Ancient and Modern Humans

Michael Schatz

Oct 2, 2014

WSBS Genomics

Agenda

1. Clustering Refresher
   1. Hierarchical Clustering
   2. PCA

2. Ancient and Modern Human Evolution
   1. Modern Diversity
   2. Ancient Hominids

3. Genetic Privacy
   1. lobSTR and Microsatellites
   2. Surname inference

Clustering Refresher

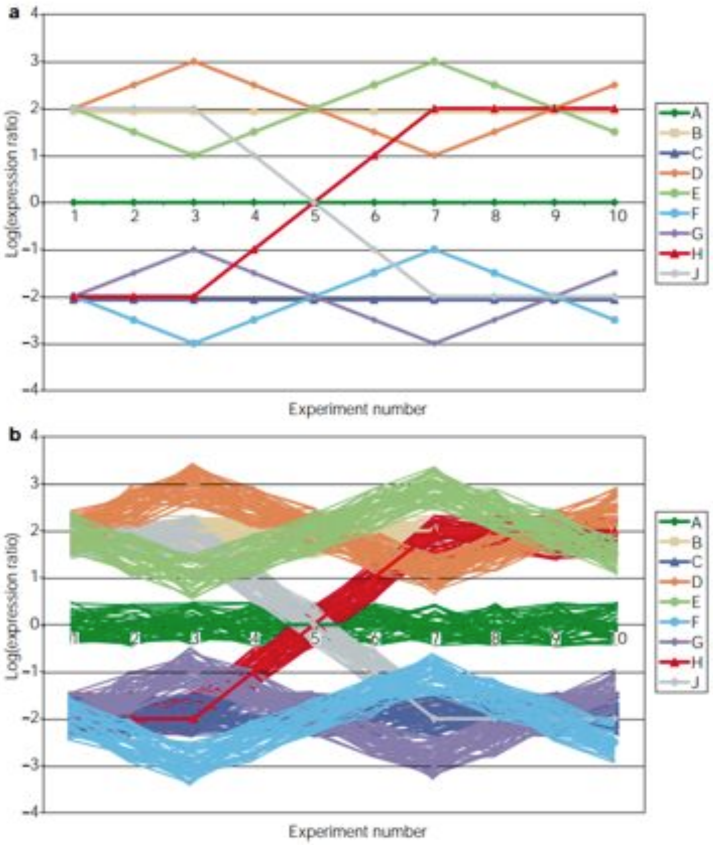
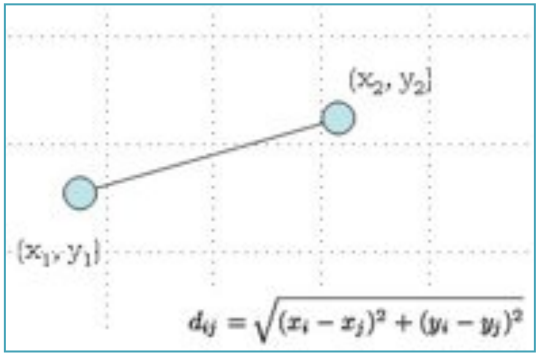
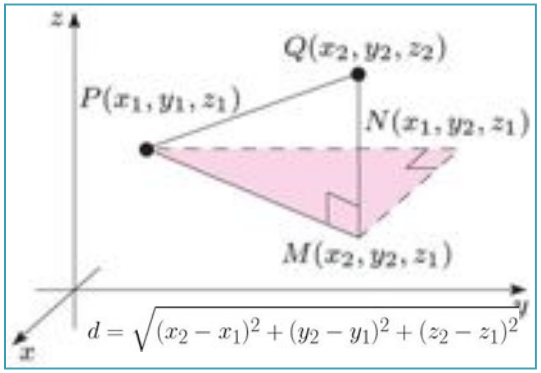


Figure 2 | **A synthetic gene-expression data set.** This data set provides an opportunity to evaluate how various clustering algorithms reveal different features of the data. **a** | Nine distinct gene-expression patterns were created with $\log_2$(ratio) expression measures defined for ten experiments. **b** | For each expression pattern, 50 additional genes were generated, representing variations on the basic patterns.

Euclidean Distance

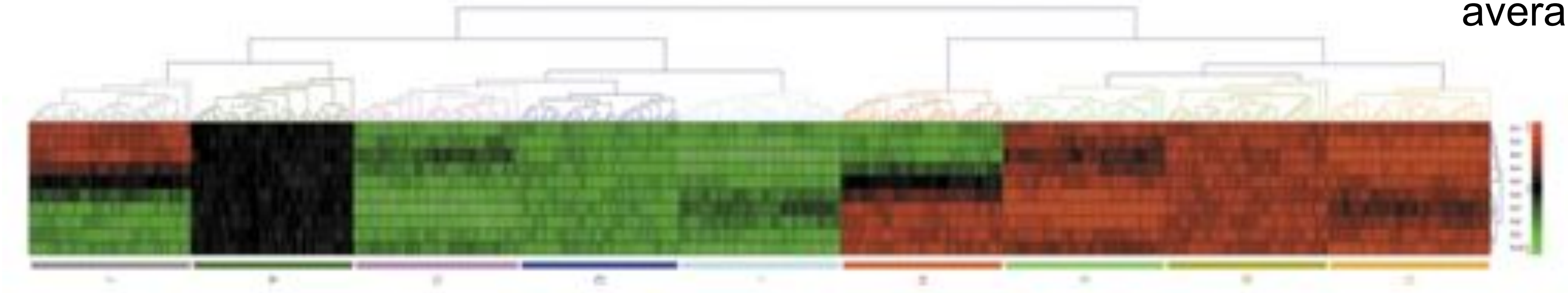




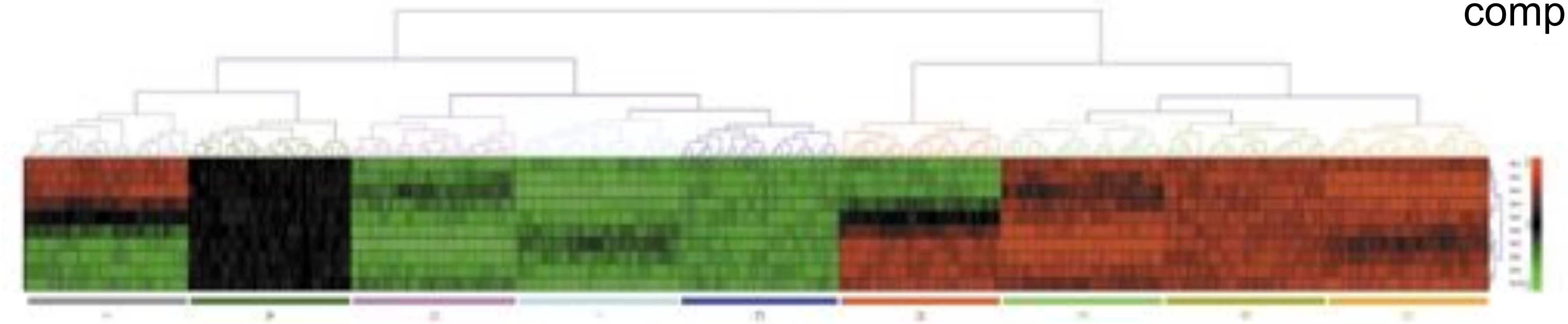
$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Computational genetics: Computational analysis of microarray data
Quackenbush (2001) *Nature Reviews Genetics*. doi:10.1038/35076576

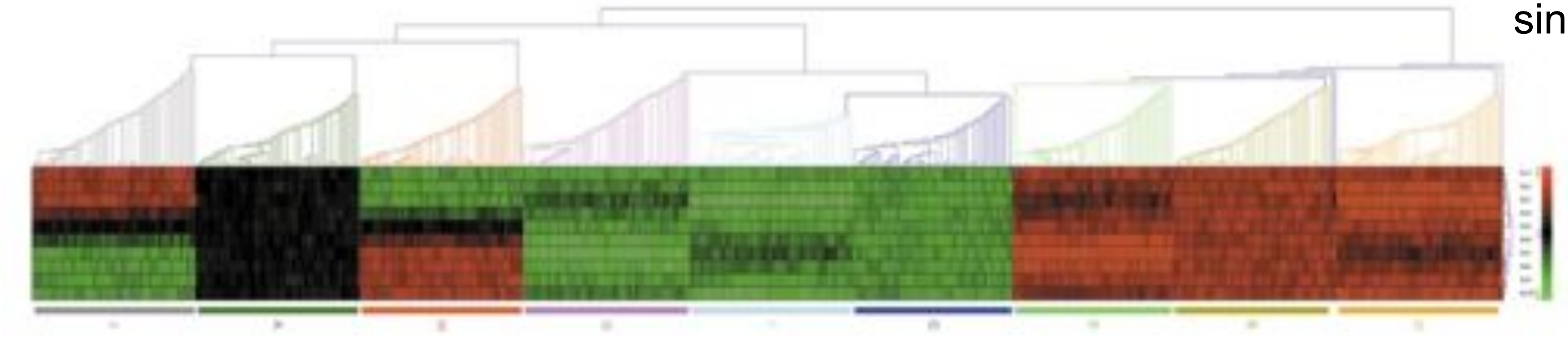
Hierarchical Clustering



average

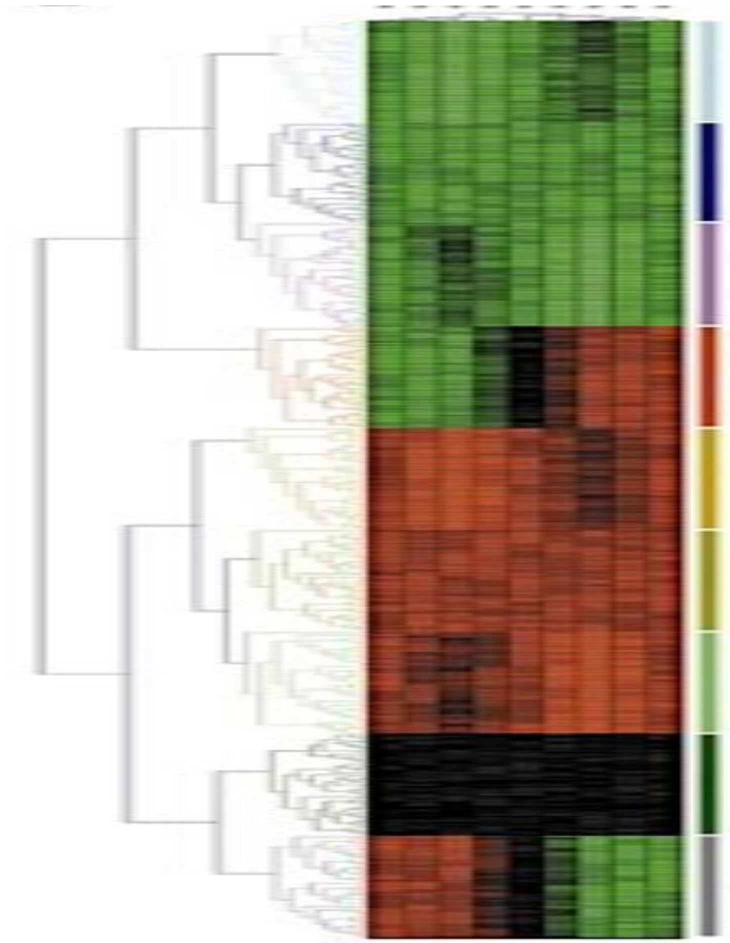


complete



single

Principle Components Analysis (PCA)



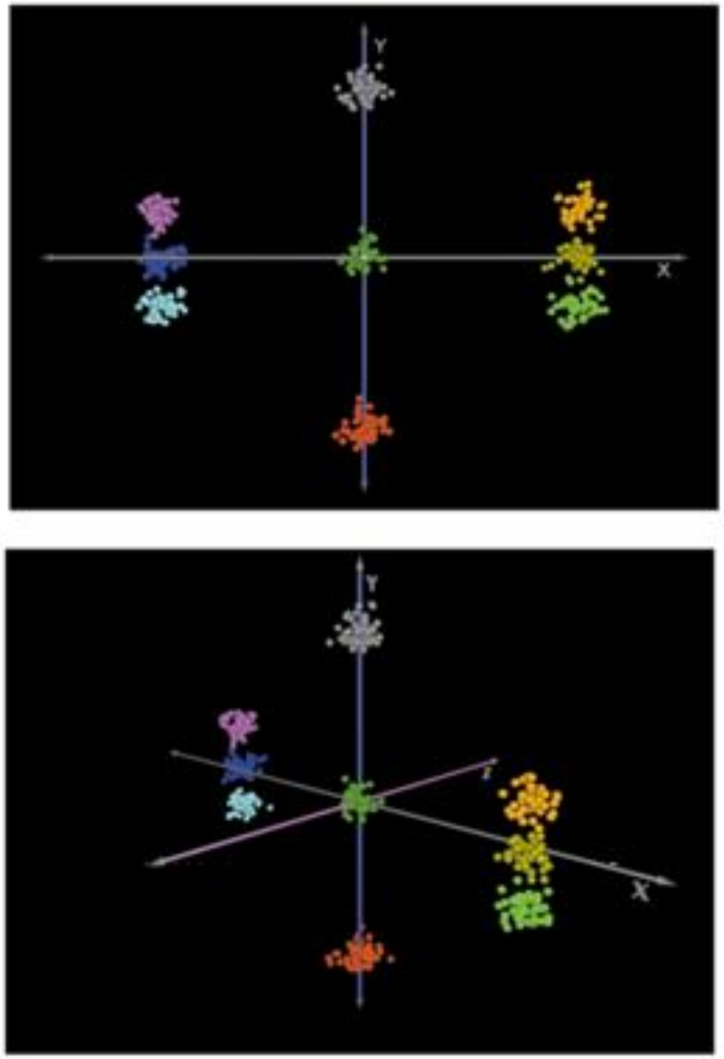


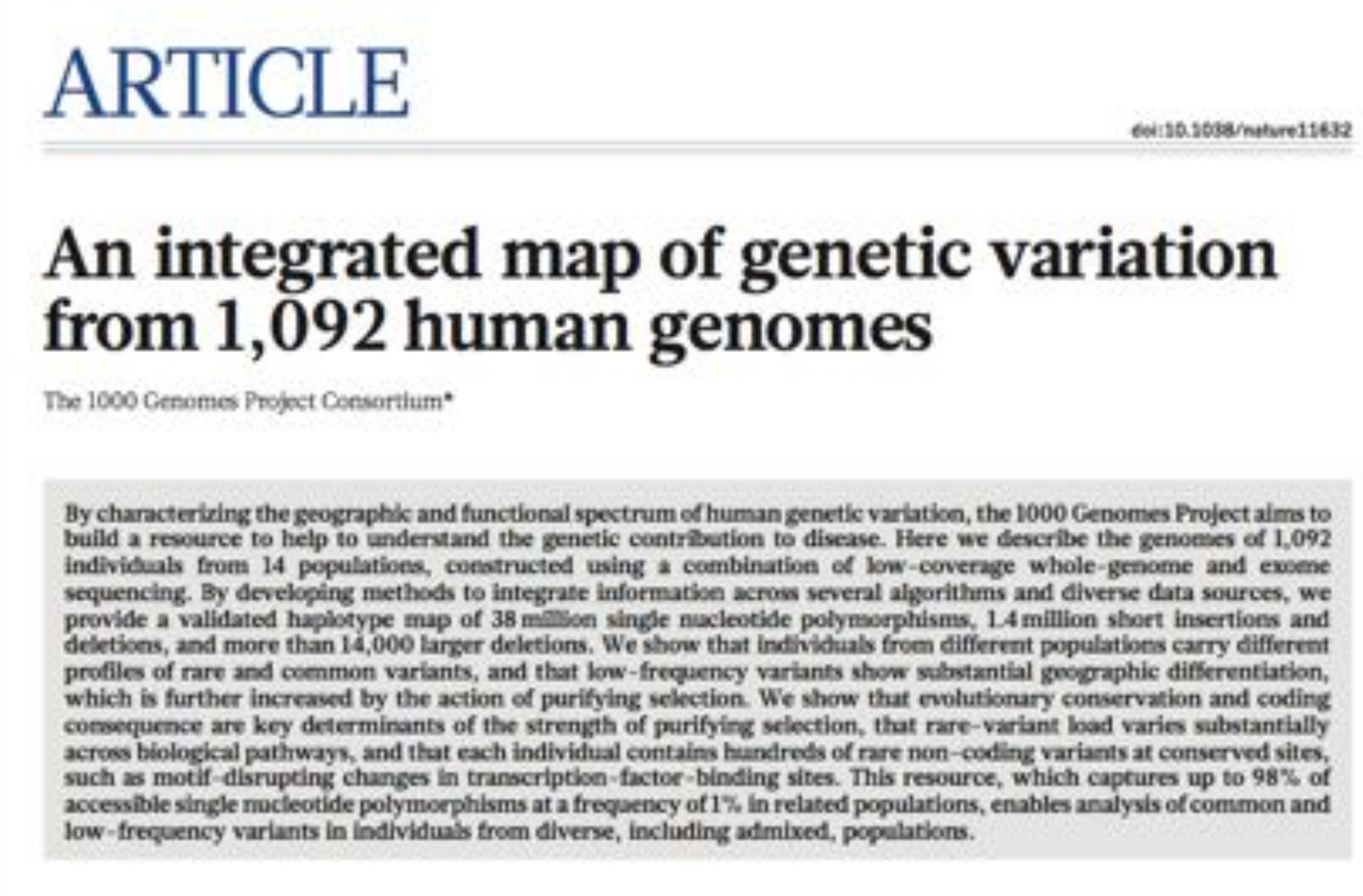
Figure 4 | **Principal component analysis.** The same demonstration data set was analysed using **a** | hierarchical (average-linkage) clustering and **b** | principal component analysis using Euclidean distance, to show how each treats the data, with genes colour coded on the basis of hierarchical clustering results for comparison.

Agenda

1. Clustering Refresher
   1. Hierarchical Clustering
   2. PCA

2. Ancient and Modern Human Evolution
   1. Modern Diversity
   2. Ancient Hominids

3. Genetic Privacy
   1. lobSTR and Microsatellites
   2. Surname inference



ARTICLE

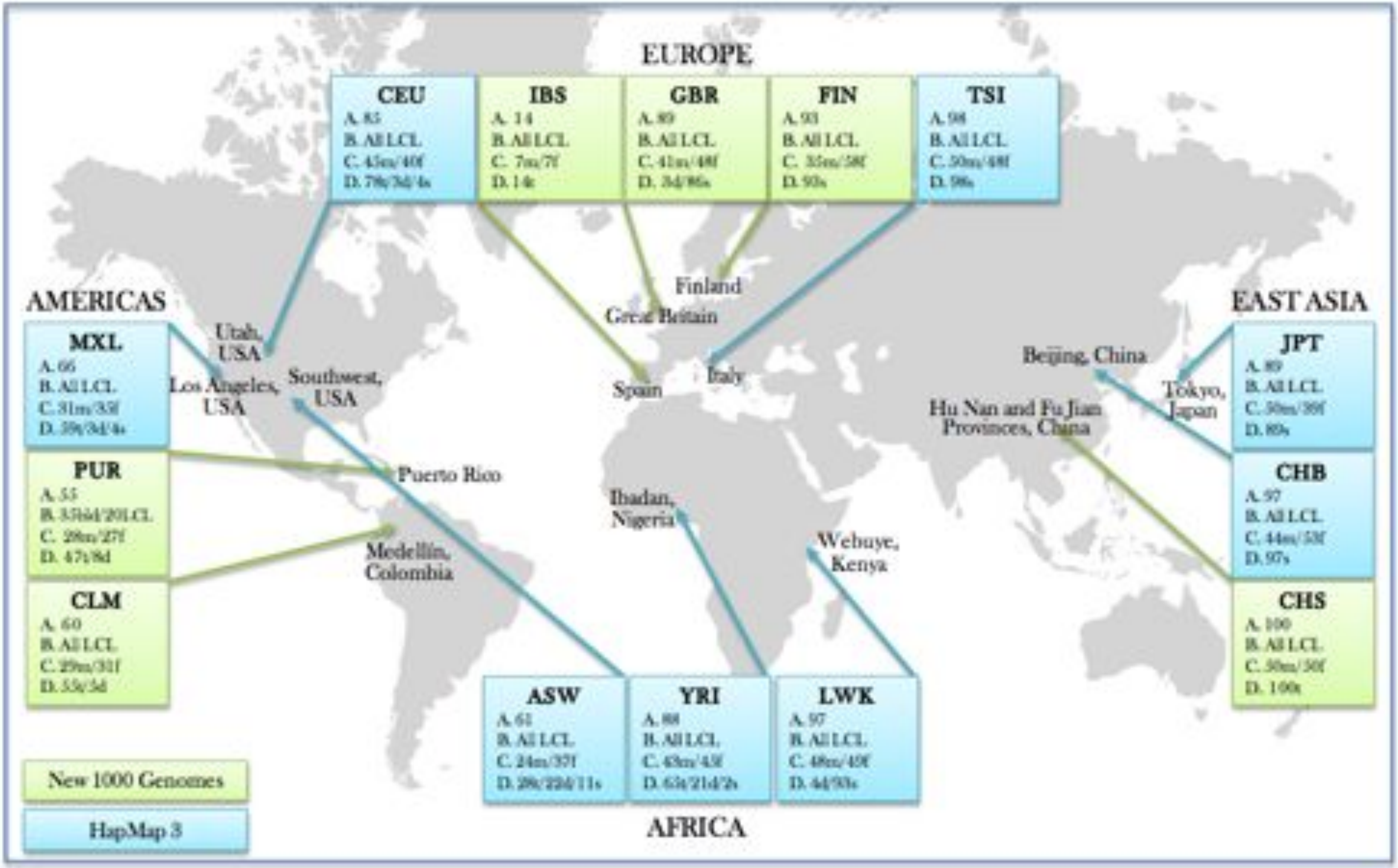
doi:10.1038/nature11632

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

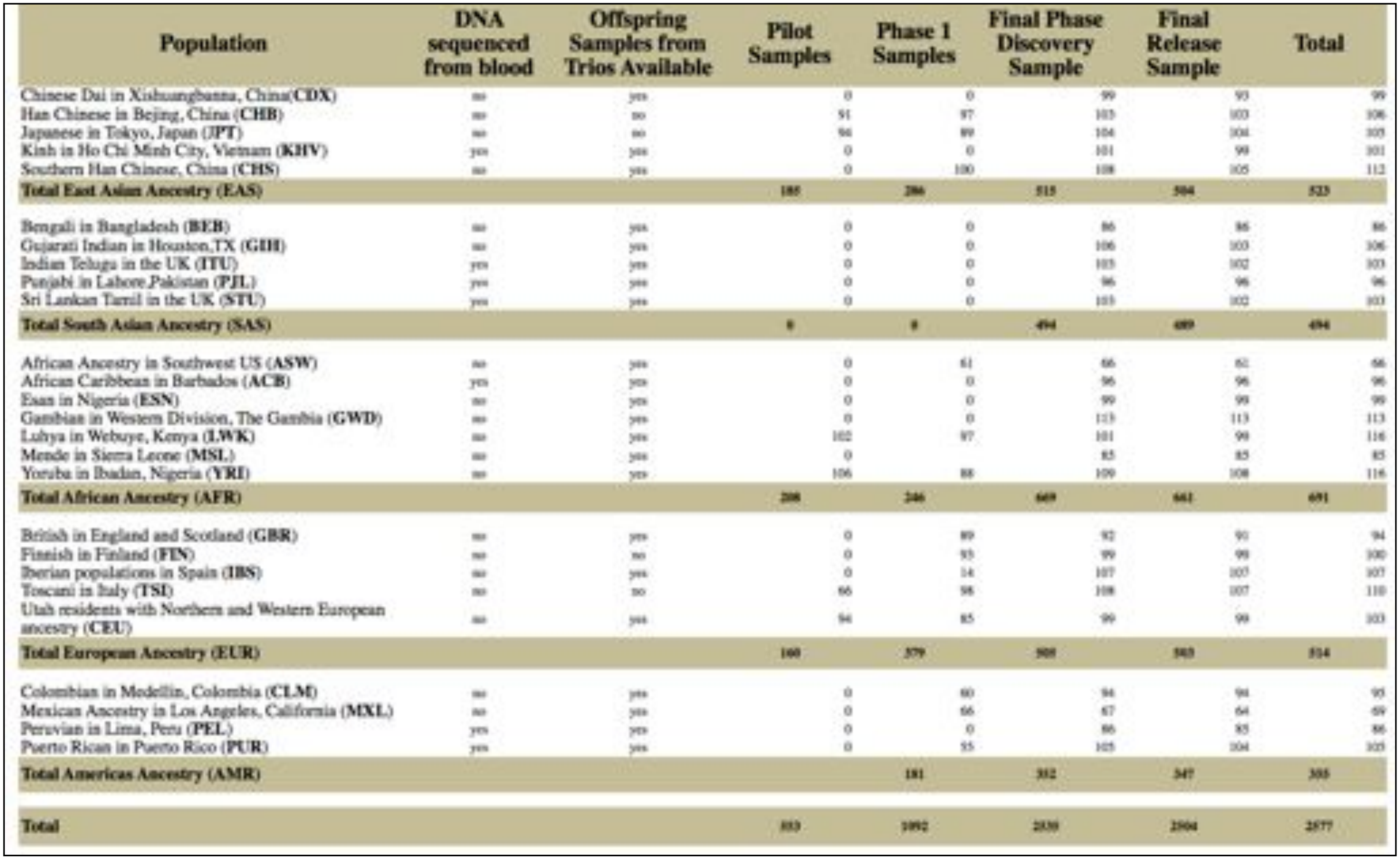
The 1000 Genomes Project Consortium*

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	99	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			**553**	**1092**	**2535**	**2504**	**2577**

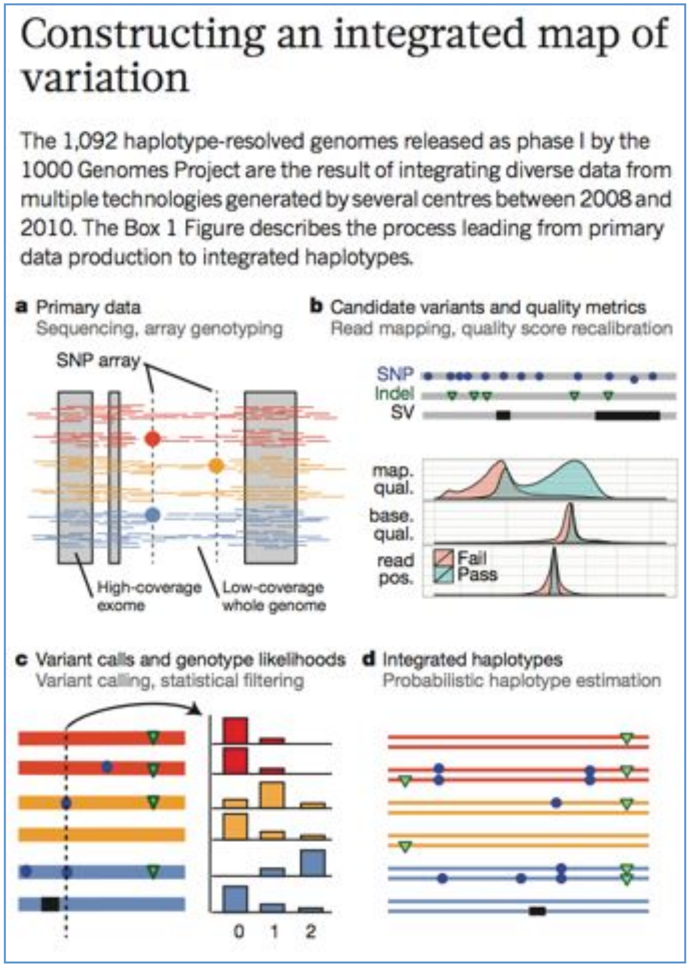
26 populations from 5 major population groups

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



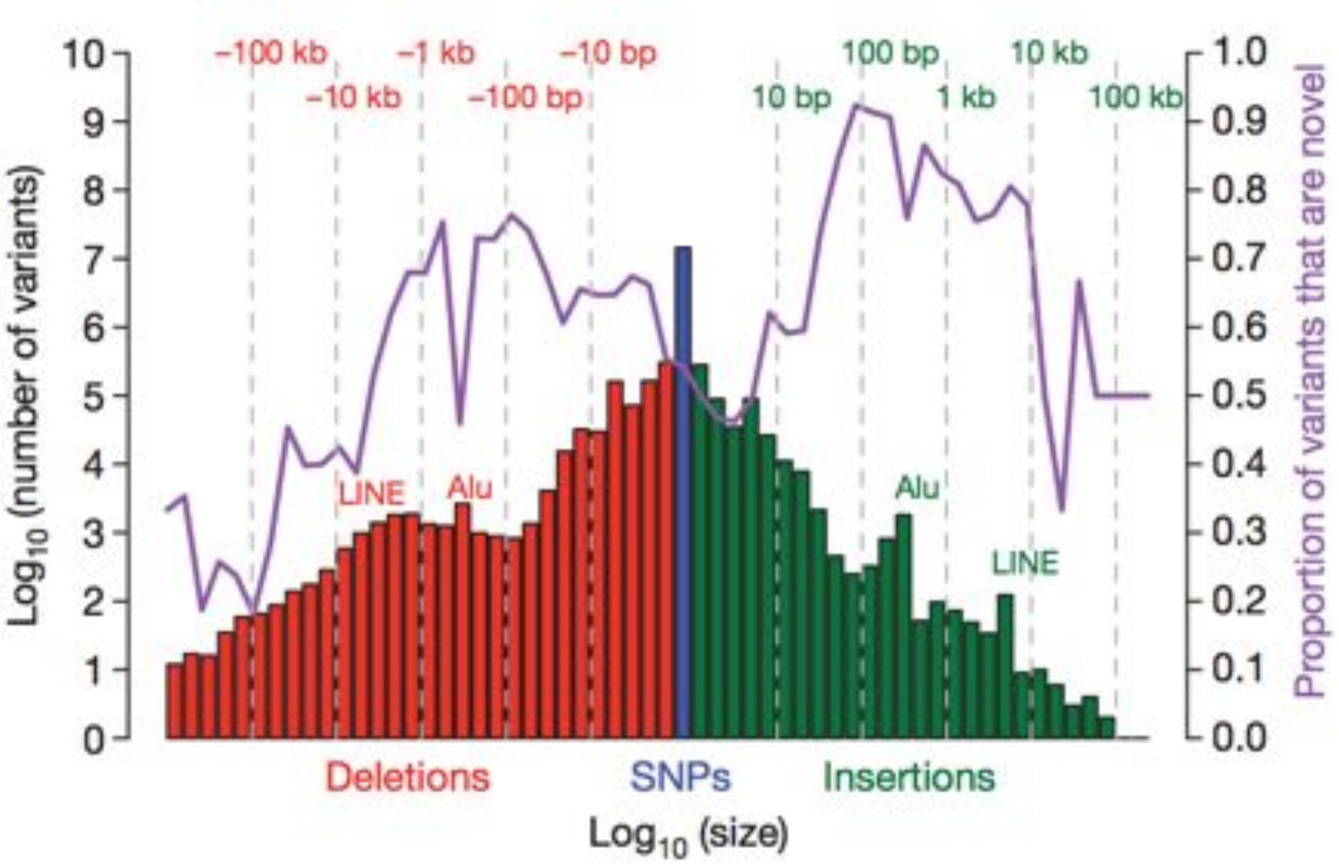
Constructing an integrated map of variation

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

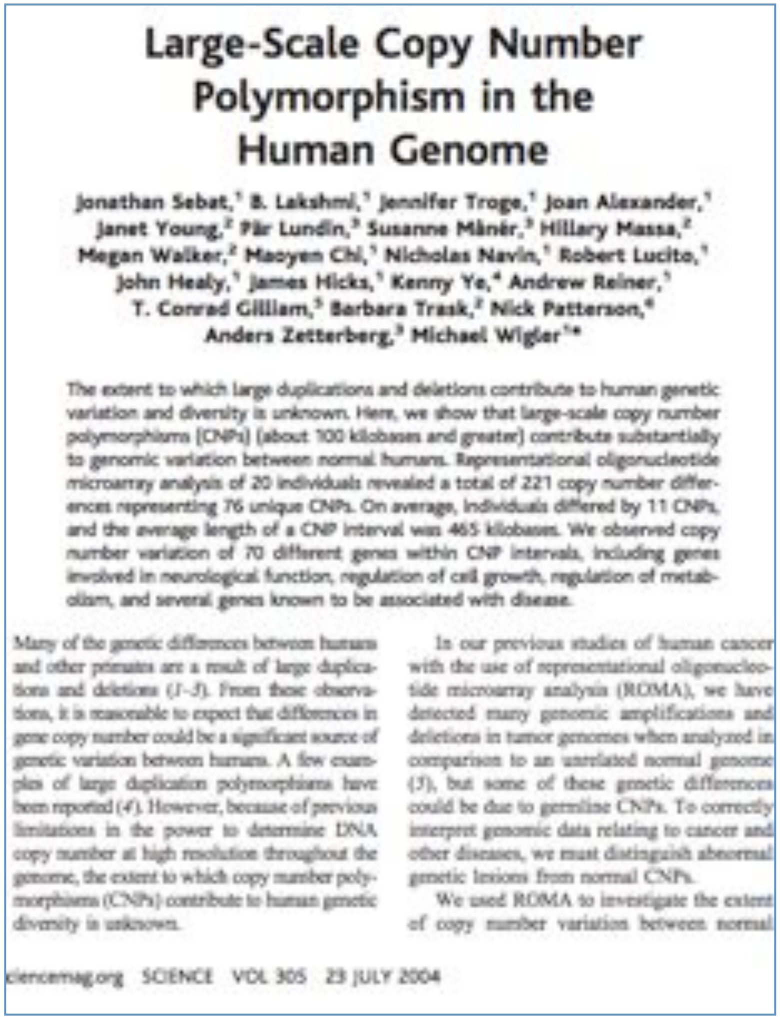
Human Mutation Types



- 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

Copy Number Variations



Large-Scale Copy Number Polymorphism in the Human Genome

Jonathan Sebat,[1] B. Lakshmi,[1] Jennifer Troge,[1] Joan Alexander,[1] Janet Young,[2] Pär Lundin,[3] Susanne Månér,[3] Hillary Massa,[2] Megan Walker,[2] Maoyen Chi,[1] Nicholas Navin,[1] Robert Lucito,[1] John Healy,[1] James Hicks,[1] Kenny Ye,[4] Andrew Reiner,[1] T. Conrad Gilliam,[5] Barbara Trask,[2] Nick Patterson,[6] Anders Zetterberg,[3] Michael Wigler[1]*

The extent to which large duplications and deletions contribute to human genetic variation and diversity is unknown. Here, we show that large-scale copy number polymorphisms (CNPs) (about 100 kilobases and greater) contribute substantially to genomic variation between normal humans. Representational oligonucleotide microarray analysis of 20 individuals revealed a total of 221 copy number differences representing 76 unique CNPs. On average, individuals differed by 11 CNPs, and the average length of a CNP interval was 465 kilobases. We observed copy number variation of 70 different genes within CNP intervals, including genes involved in neurological function, regulation of cell growth, regulation of metabolism, and several genes known to be associated with disease.

Many of the genetic differences between humans and other primates are a result of large duplications and deletions (*1–3*). From these observations, it is reasonable to expect that differences in gene copy number could be a significant source of genetic variation between humans. A few examples of large duplication polymorphisms have been reported (*4*). However, because of previous limitations in the power to determine DNA copy number at high resolution throughout the genome, the extent to which copy number polymorphisms (CNPs) contribute to human genetic diversity is unknown.

In our previous studies of human cancer with the use of representational oligonucleotide microarray analysis (ROMA), we have detected many genomic amplifications and deletions in tumor genomes when analyzed in comparison to an unrelated normal genome (*5*), but some of these genetic differences could be due to germline CNPs. To correctly interpret genomic data relating to cancer and other diseases, we must distinguish abnormal genetic lesions from normal CNPs.

We used ROMA to investigate the extent of copy number variation between normal

ciencemag.org SCIENCE VOL 305 23 JULY 2004

Strong Association of De Novo Copy Number Mutations with Autism

Jonathan Sebat,[1]* B. Lakshmi,[1] Dheeraj Malhotra,[1]* Jennifer Troge,[1]* Christa Lese-Martin,[2] Tom Walsh,[3] Boris Yamrom,[1] Seungtai Yoon,[1] Alex Krasnitz,[1] Jude Kendall,[1] Anthony Leotta,[1] Deepa Pai,[1] Ray Zhang,[1] Yoon-Ha Lee,[1] James Hicks,[1] Sarah J. Spence,[4] Annette T. Lee,[5] Kaija Puura,[6] Terho Lehtimäki,[7] David Ledbetter,[2] Peter K. Gregersen,[5] Joel Bregman,[8] James S. Sutcliffe,[9] Vaidehi Jobanputra,[10] Wendy Chung,[10] Dorothy Warburton,[10] Mary-Claire King,[3] David Skuse,[11] Daniel H. Geschwind,[12] T. Conrad Gilliam,[13] Kenny Ye,[14] Michael Wigler[1]†

We tested the hypothesis that de novo copy number variation (CNV) is associated with autism spectrum disorders (ASDs). We performed comparative genomic hybridization (CGH) on the genomic DNA of patients and unaffected subjects to detect copy number variants not present in their respective parents. Candidate genomic regions were validated by higher-resolution CGH, paternity testing, cytogenetics, fluorescence in situ hybridization, and microsatellite genotyping. Confirmed de novo CNVs were significantly associated with autism (P = 0.0005). Such CNVs were identified in 12 out of 118 (10%) of patients with sporadic autism, in 2 out of 77 (3%) of patients with an affected first-degree relative, and in 2 out of 196 (1%) of controls. Most de novo CNVs were smaller than microscopic resolution. Affected genomic regions were highly heterogeneous and included mutations of single genes. These findings establish de novo germline mutation as a more significant risk factor for ASD than previously recognized.

Autism spectrum disorders (ASDs) [Mendelian Inheritance in Man (MIM) 209850] are characterized by language impairments, social deficits, and repetitive behaviors. The onset of symptoms occurs by the age of 3 and usually requires extensive support for the lifetime of the afflicted. The prevalence of ASD is estimated to be 1 in 166 (*1*), making it a major burden to society.

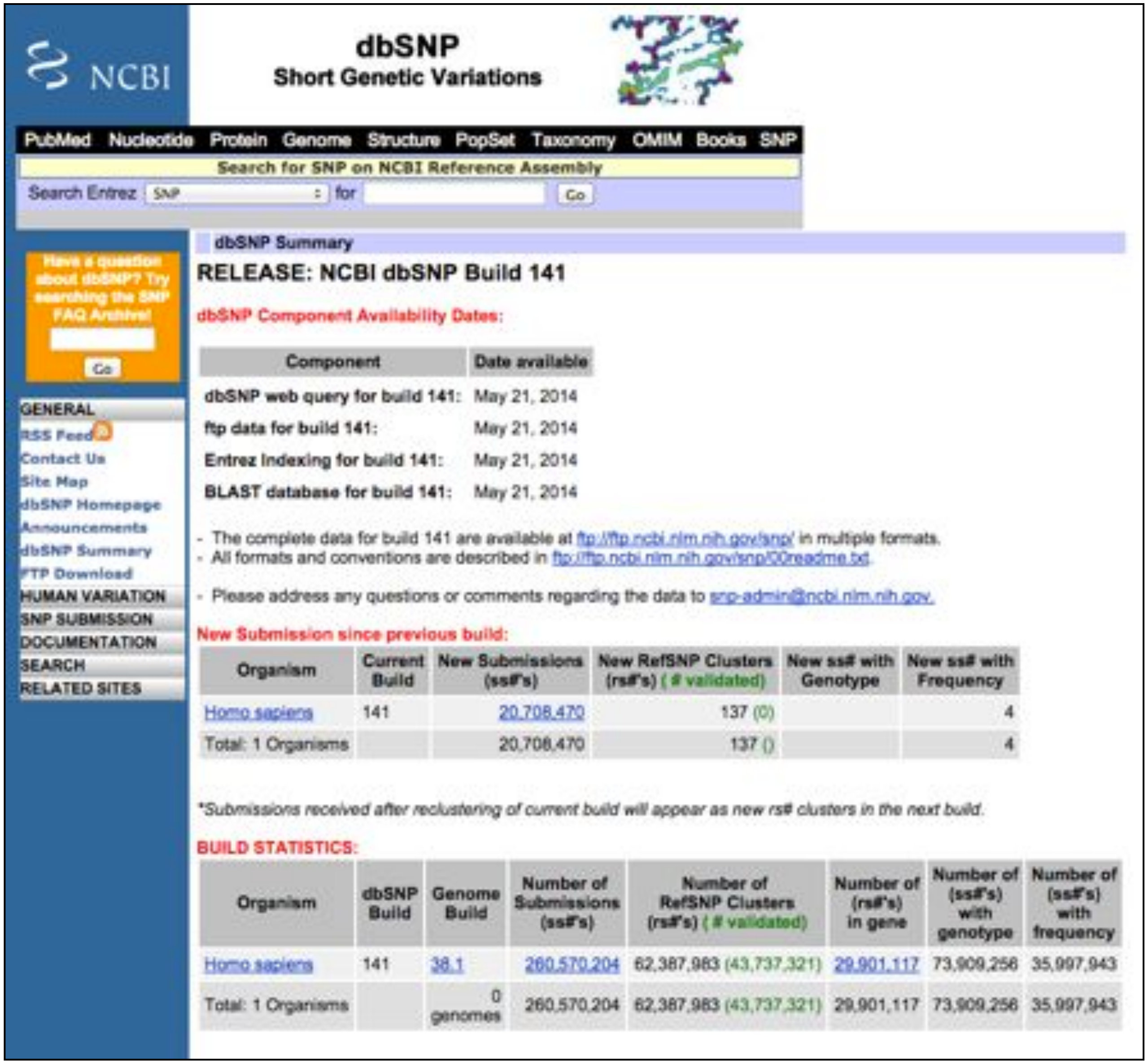
Genetics plays a major role in the etiology of autism. The concordance rates in monozygotic twins are 70% for autism and 90% for ASD, whereas the concordance rates in dizygotic twins are 5% and 10%, respectively. Previous studies suggest autism displays a high degree of genetic heterogeneity. Efforts to map disease genes using linkage analysis have found evidence for autism loci on 20 different chromosomes. Regions implicated by multiple studies include 1p, 5q, 7q, 15q, 16p, 17q, 19p, and Xq (*2*). Moreover, microscopy studies have identified cytogenetic abnormalities in >5% of affected children, involving many different loci on all chromosomes (*3*). In some rare syndromic forms of autism, such as Rett syndrome (*4*) and tuberous sclerosis (*5*), mutations in a single gene have been identified. Otherwise, neither linkage nor cytogenetics has unambiguously identified specific genes involved.

Genetic heterogeneity poses a considerable challenge to traditional approaches for gene mapping (*6*). Some of these limitations are overcome by methods that rely on the direct detection of functional variants, which in most cases are de novo events. New array-based technologies can detect differences in DNA copy number at much higher resolution than cytogenetic methods (*7*) and, hence, might reveal spontaneous mutations that were previously unidentified. These techniques have shown an abundance of copy number variants (CNVs) in humans (*8*, *9*), and the same methods have been used to find de novo chromosome aberrations below the resolution of microscopy in children with mental retardation and dysmorphic features (*10–14*), including patients with syndromic forms of autism (*15*). Yet, the association of spontaneous CNVs in idiopathic autism has not been systematically investigated. Thus, a large-scale study of genome copy number variation in

www.sciencemag.org SCIENCE VOL 316 20 APR

While fewer numbers of CNVs occur per person, the total number of bases involved can be much greater and have profound effect.

dbSNP


NCBI
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
Short Genetic Variations

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

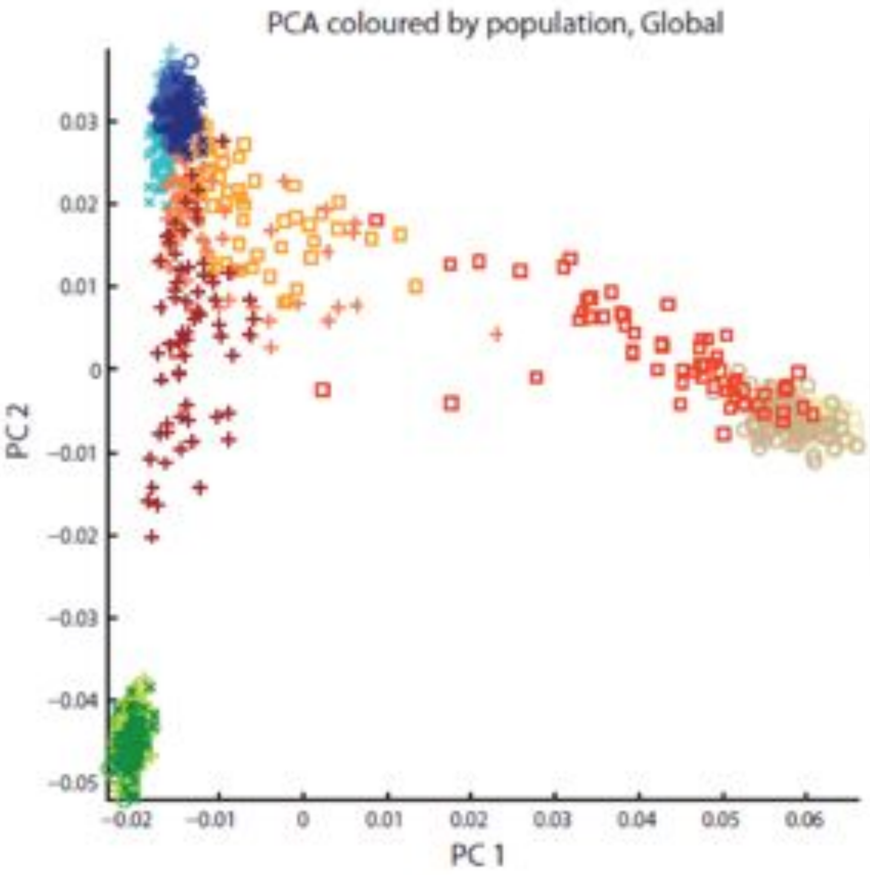
*Submissions received after reclustering of current build will appear as new rs# clusters in the next build.

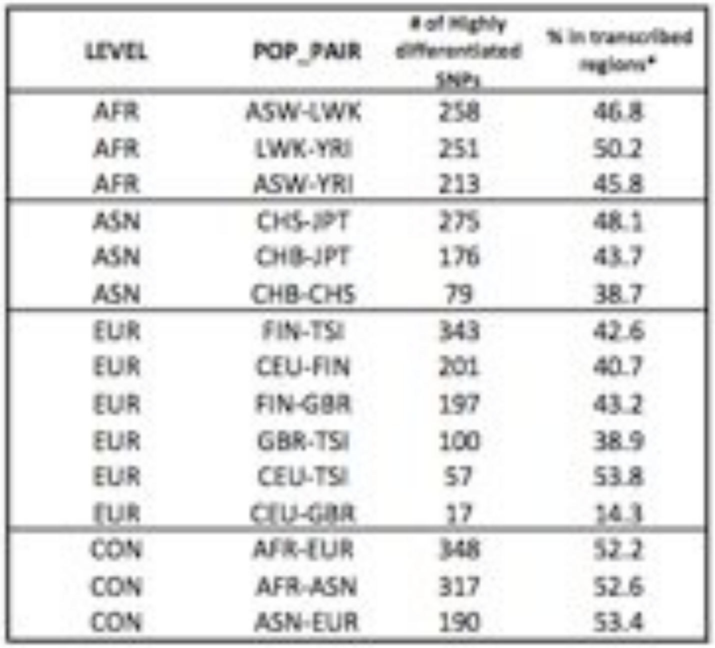
BUILD STATISTICS:

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.

Variation across populations



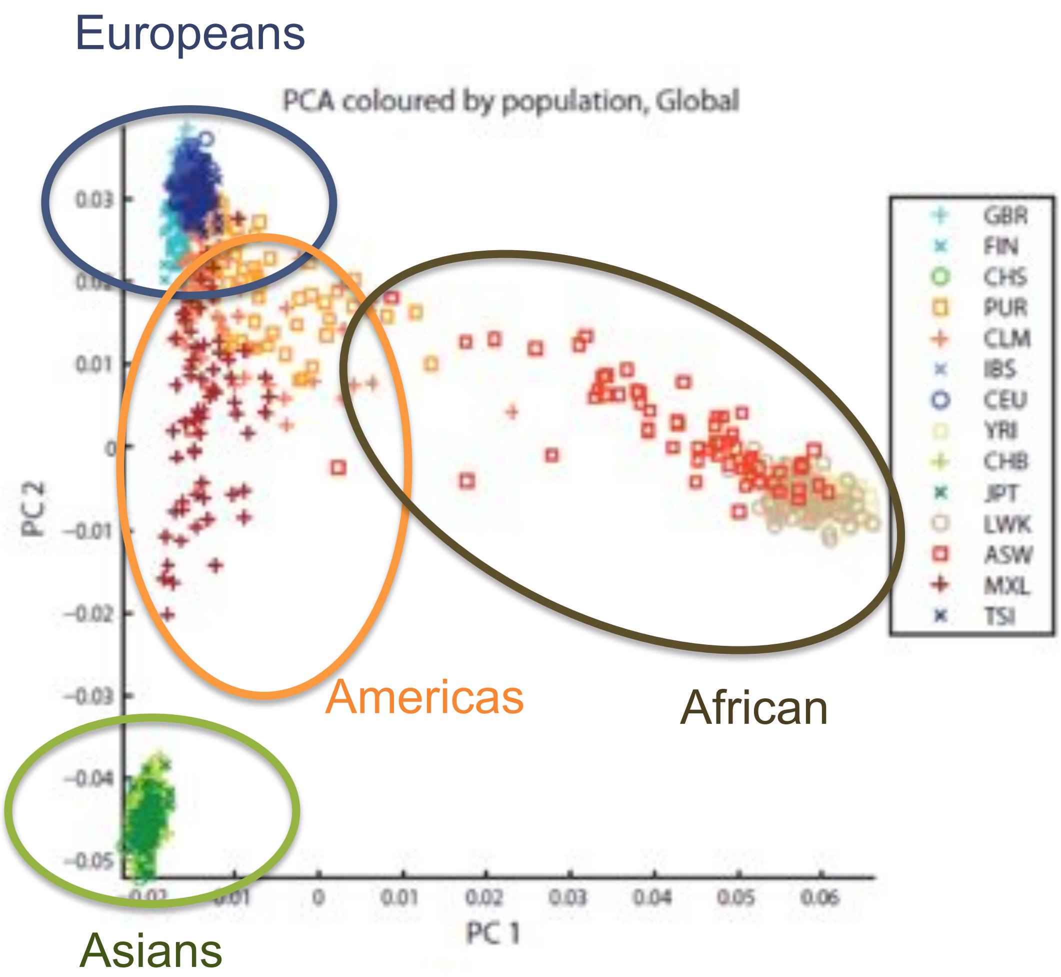


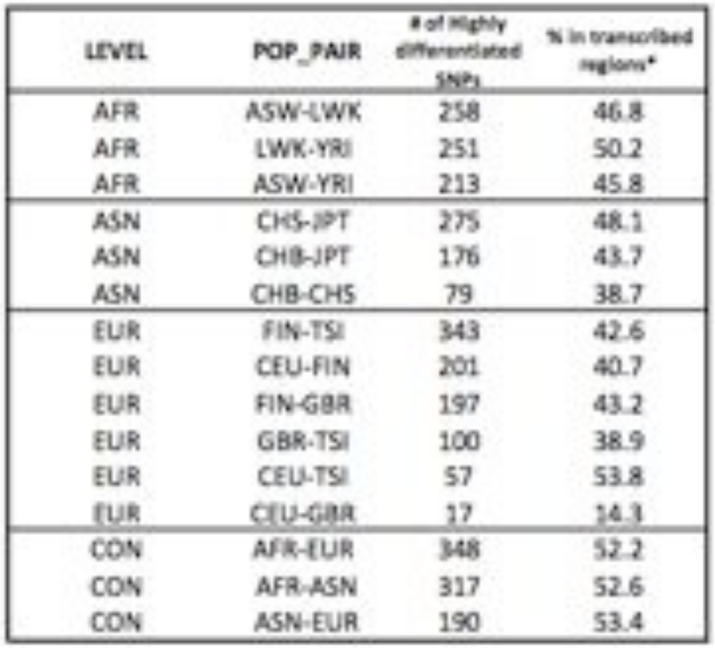
LEVEL	POP_PAIR	# of Highly differentiated SNPs	% in transcribed regions*
AFR	ASW-LWK	258	46.8
AFR	LWK-YRI	251	50.2
AFR	ASW-YRI	213	45.8
ASN	CHS-JPT	275	48.1
ASN	CHB-JPT	176	43.7
ASN	CHB-CHS	79	38.7
EUR	FIN-TSI	343	42.6
EUR	CEU-FIN	201	40.7
EUR	FIN-GBR	197	43.2
EUR	GBR-TSI	100	38.9
EUR	CEU-TSI	57	53.8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



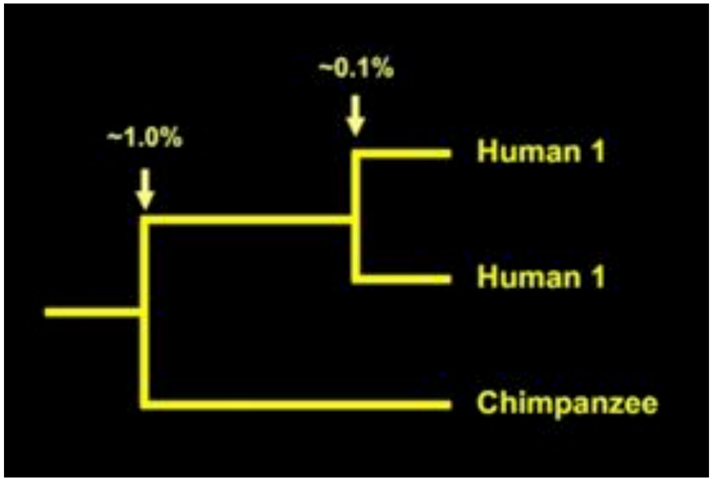


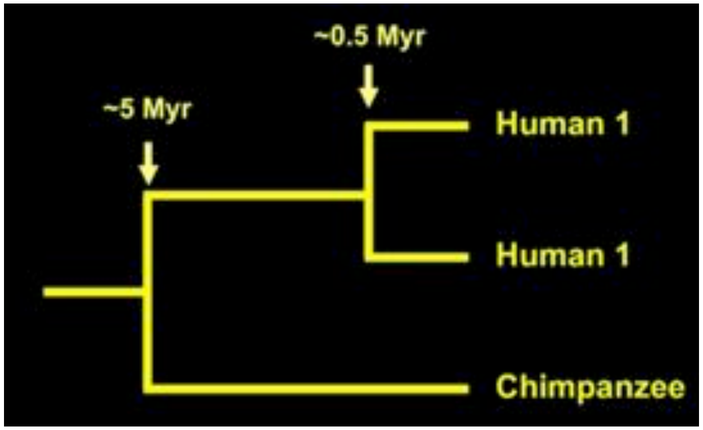
LEVEL	POP_PAIR	# of Highly differentiated SNPs	% in transcribed regions*
AFR	ASW-LWK	258	46.8
AFR	LWK-YRI	251	50.2
AFR	ASW-YRI	213	45.8
ASN	CHS-JPT	275	48.1
ASN	CHB-JPT	176	43.7
ASN	CHB-CHS	79	38.7
EUR	FIN-TSI	343	42.6
EUR	CEU-FIN	201	40.7
EUR	FIN-GBR	197	43.2
EUR	GBR-TSI	100	38.9
EUR	CEU-TSI	57	53.8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

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

- 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)!

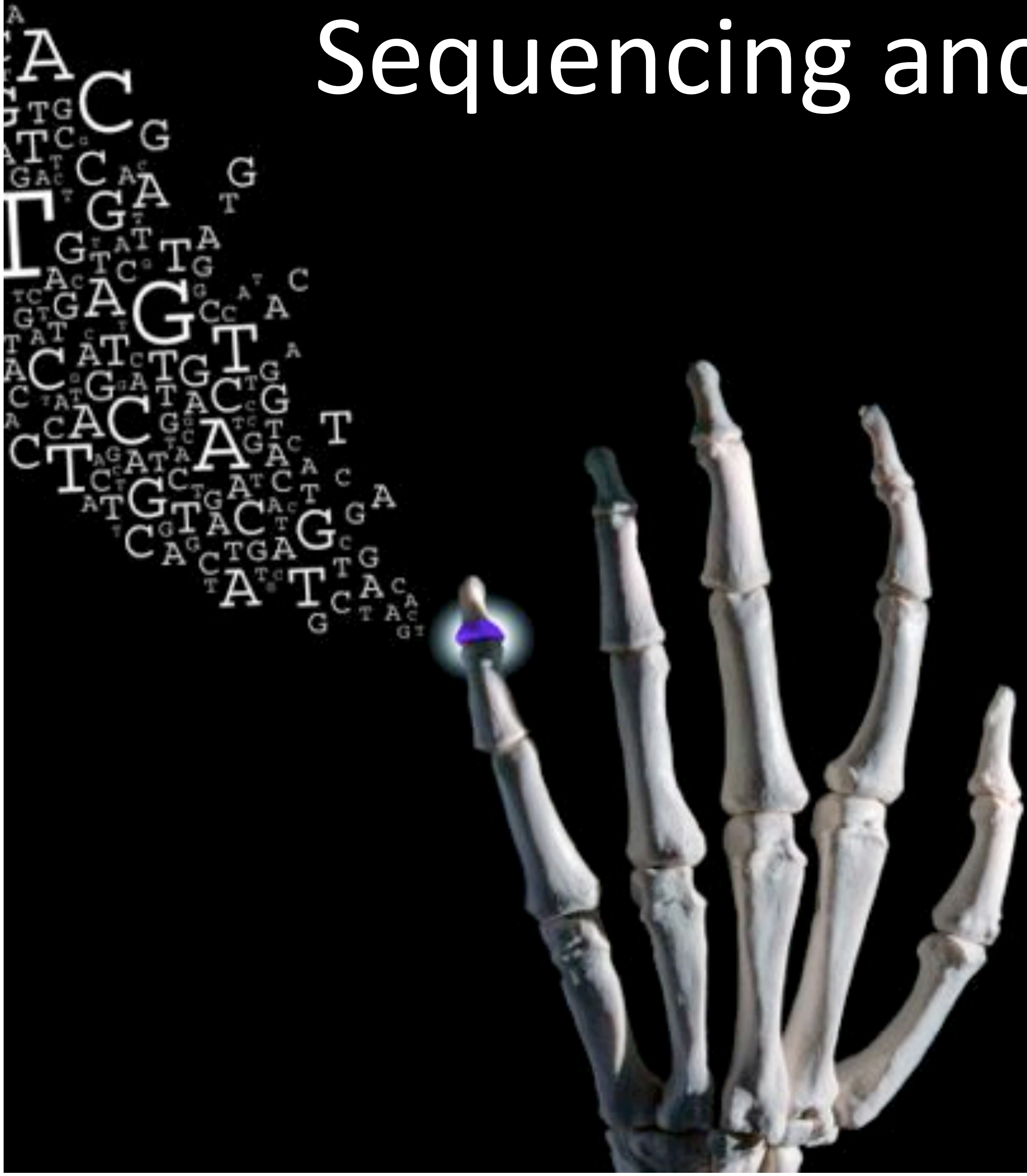
Open question if/how early modern humans interacted with earlier hominid

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

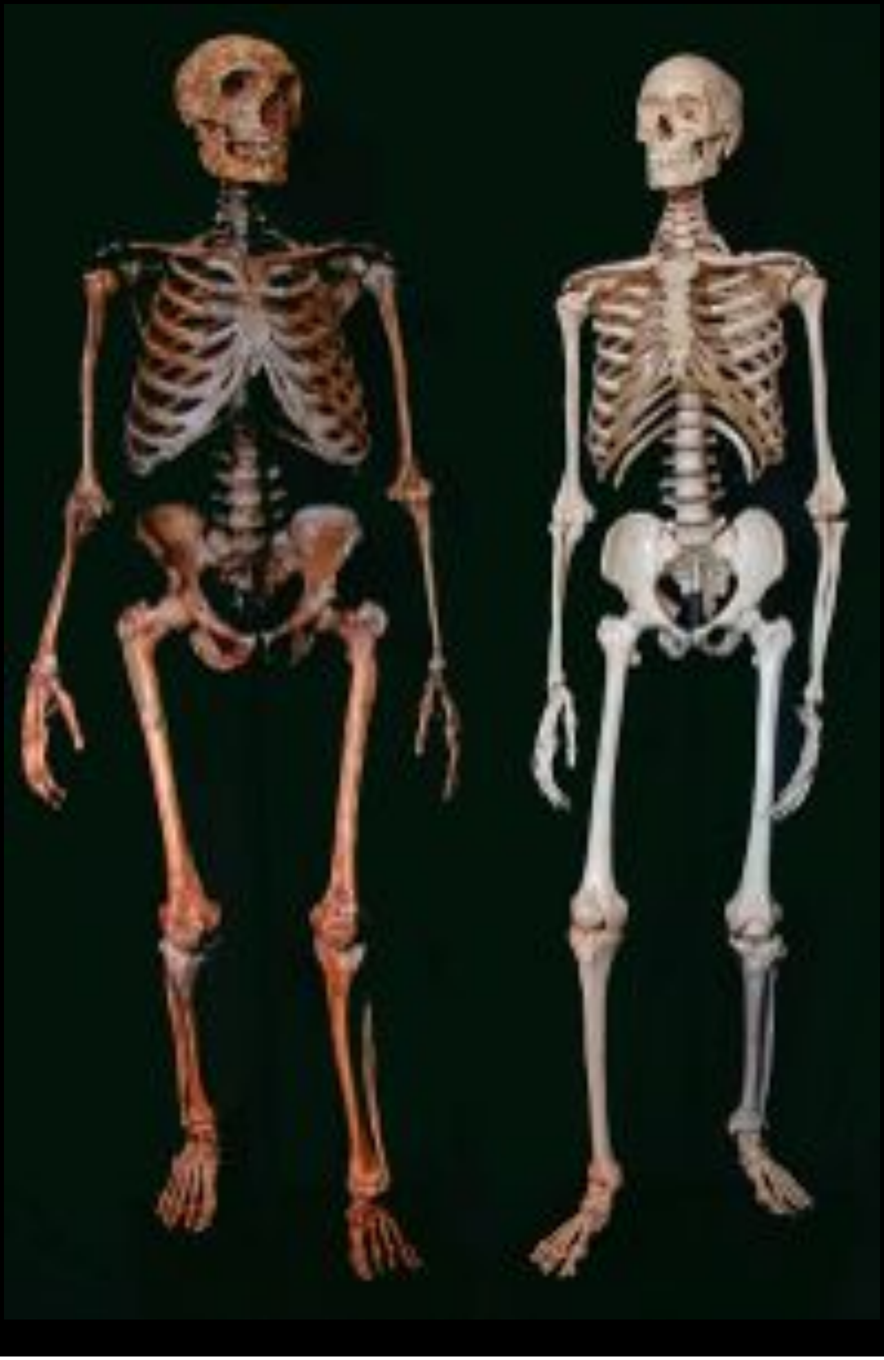
Janet Kelso

Max-Planck Institute



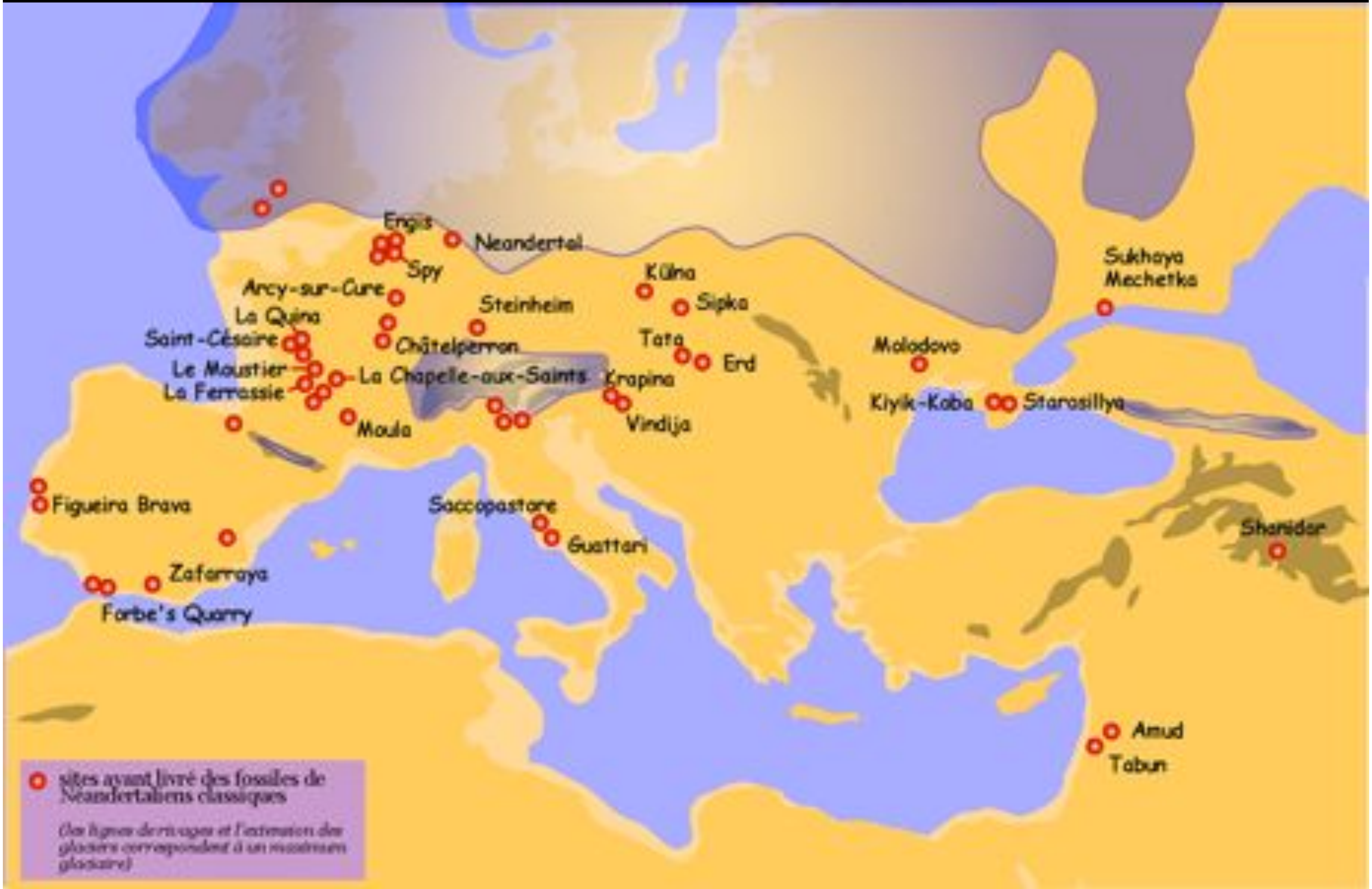
Homo neanderthalensis

- 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
Steinheim
Saint-Césaire
Châtelperron
Le Moustier
La Chapelle-aux-Saints
La Ferrassie
Moula
Külna
Sipka
Tata
Erd
Krapina
Vindija
Sukhaya
Mechetka
Molodova
Kiyik-Koba
Starosillya
Figueira Brava
Saccopastore
Guattari
Zafarraya
Forbe's Quarry
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)

A Draft Sequence of the Neandertal Genome

Richard E. Green, *et al.*

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

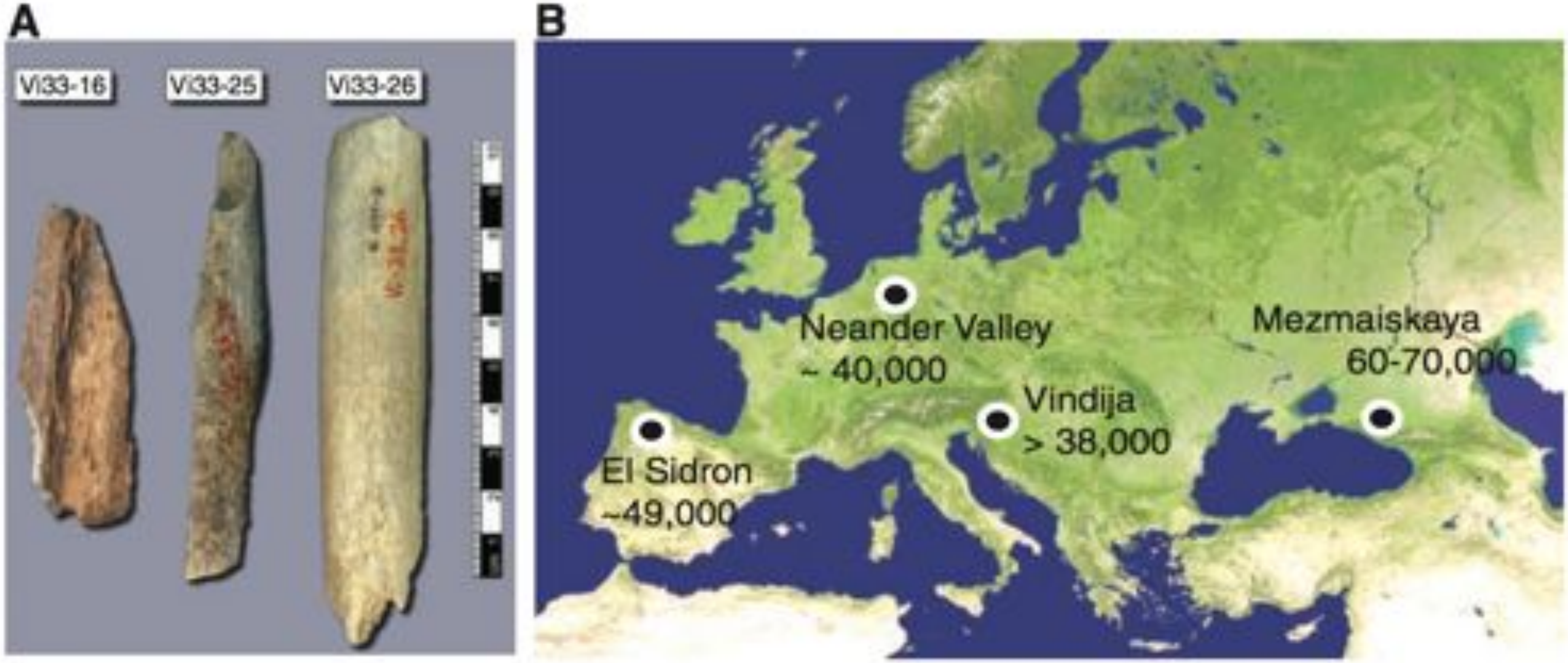
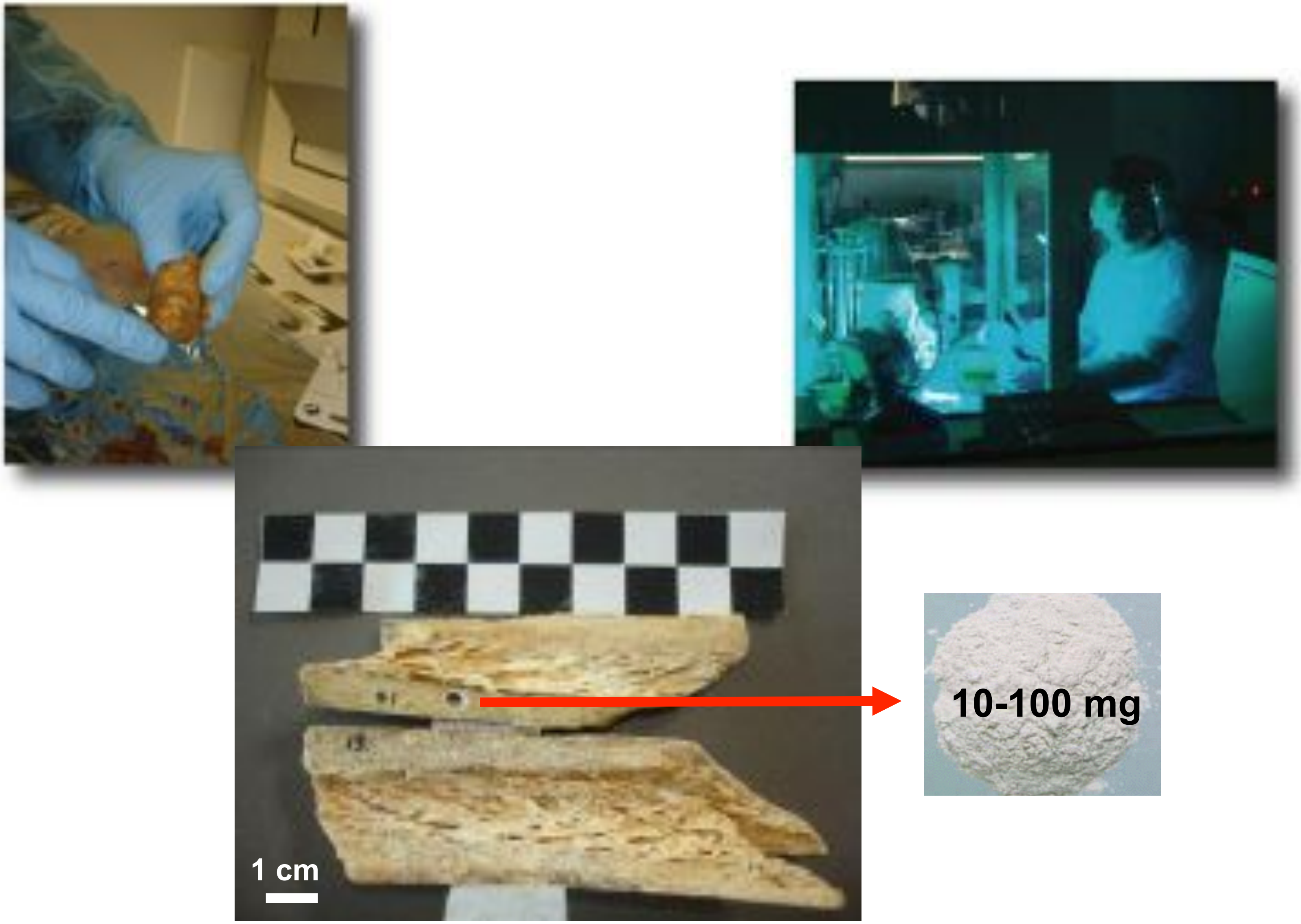
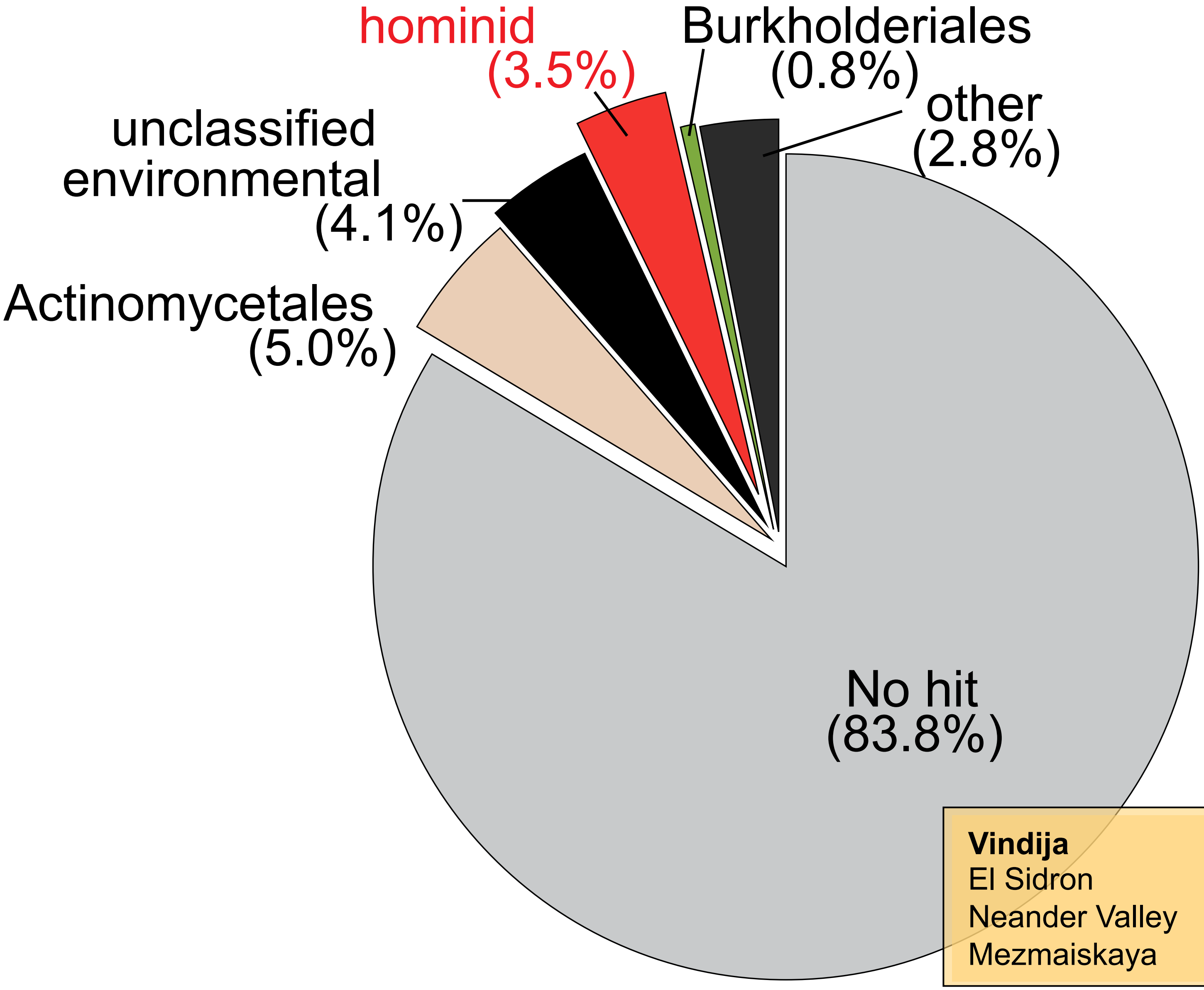


Fig. 1. Samples and sites from which DNA was retrieved. (**A**) The three bones from Vindija from which Neandertal DNA was sequenced. (**B**) Map showing the four archaeological sites from which bones were used and their approximate dates (years B.P.).

Extracting Ancient DNA

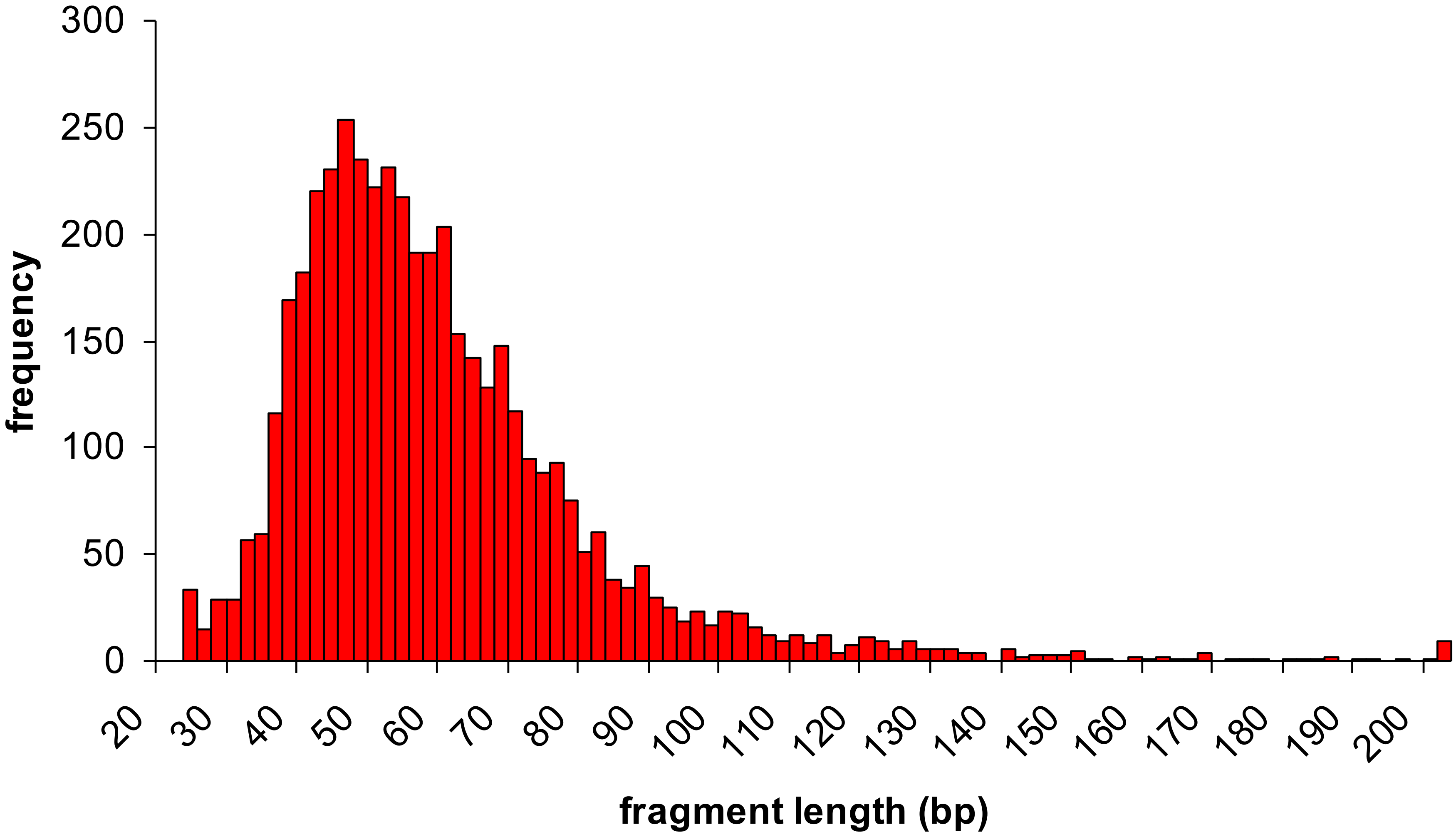


DNA is from mixed sources

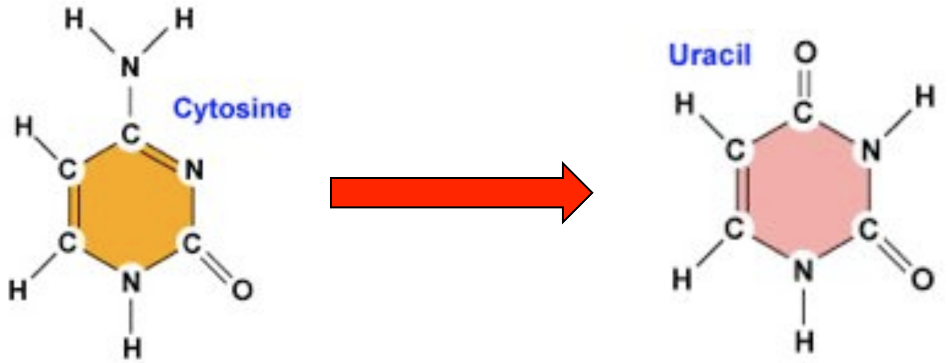


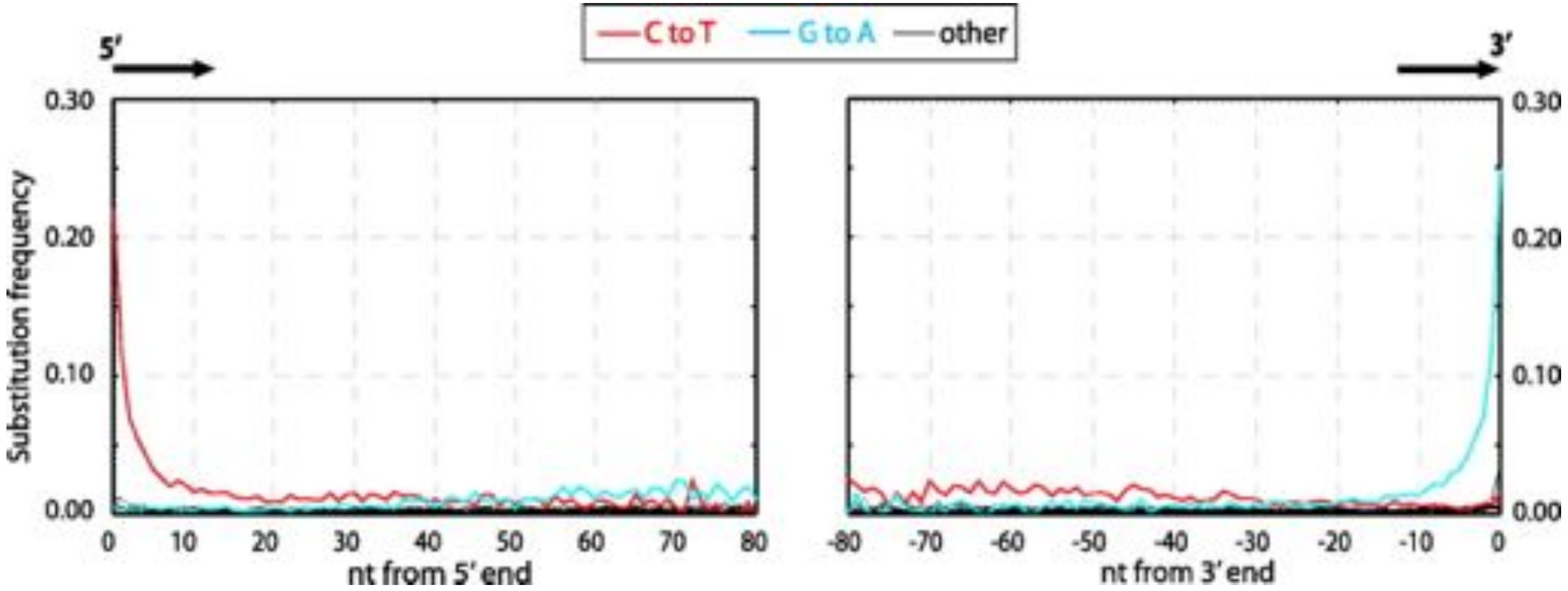
Vindija	**0.2 – 3.5%**
El Sidron	0.1 - 0.4%
Neander Valley	0.2 - 0.5%
Mezmaiskaya	0.8 - 1.5%

DNA is degraded



DNA is chemically damaged





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

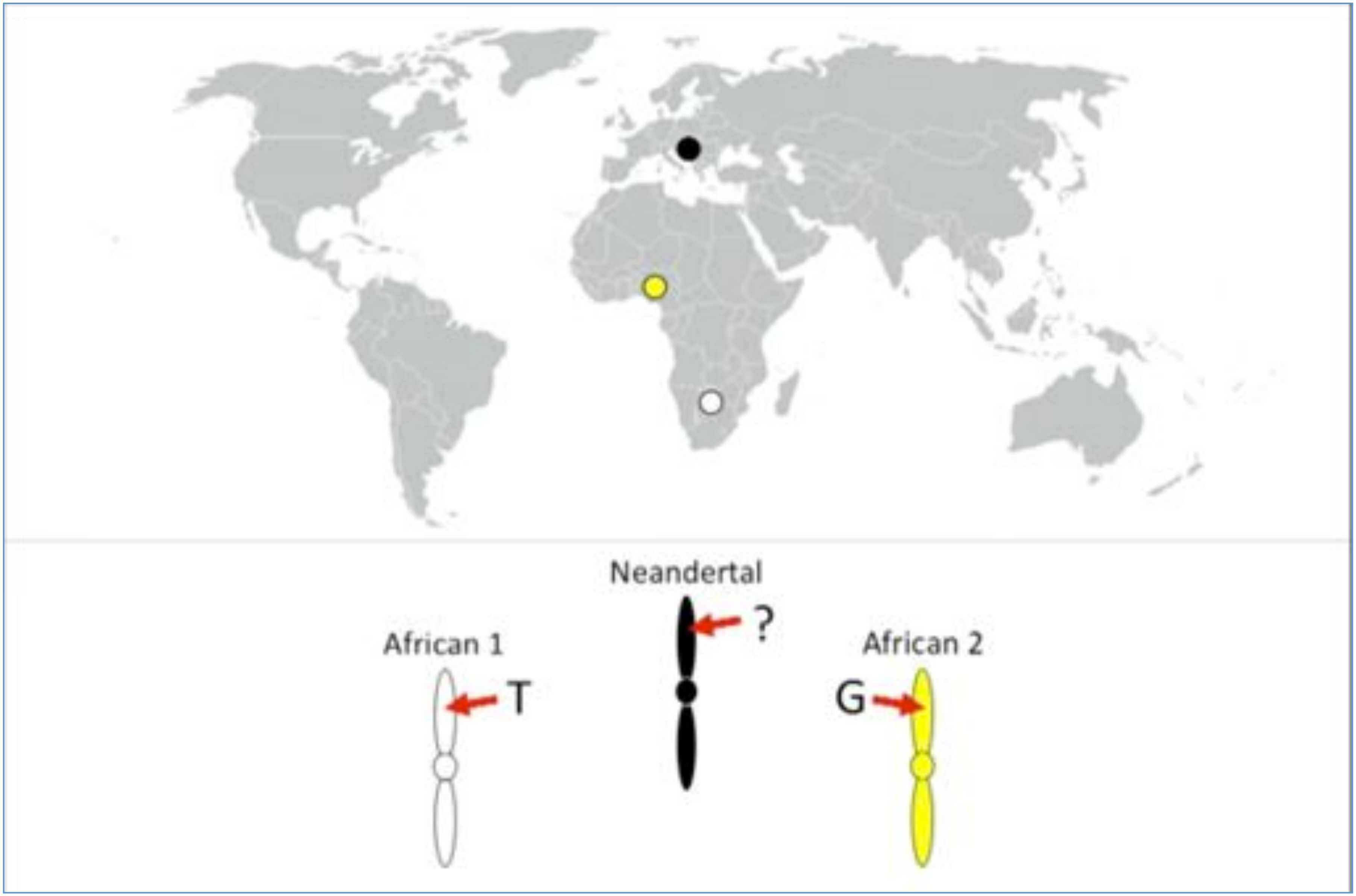
Green et al. 2010

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

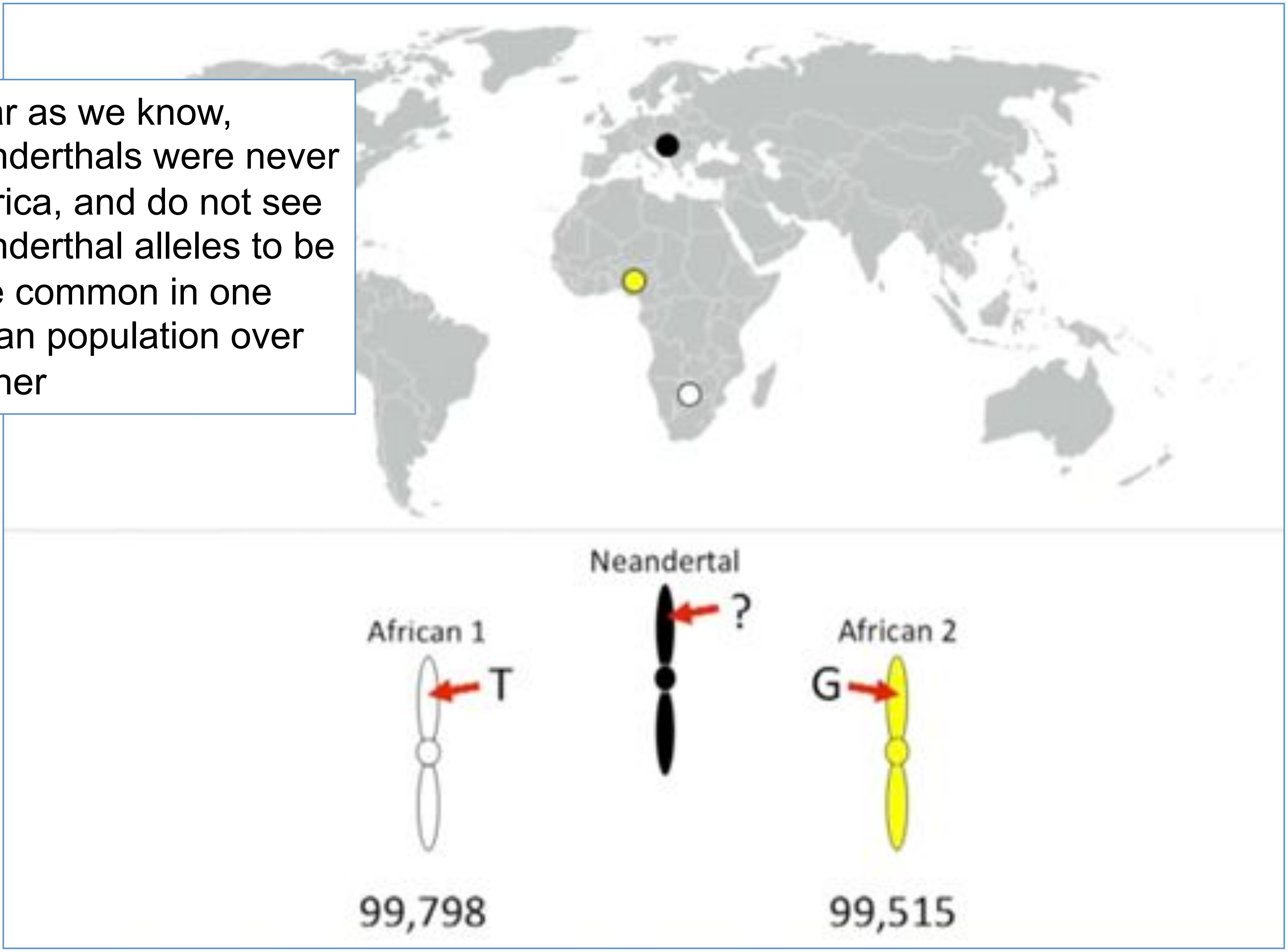
Genome coverage ~1.3 X

Did we mix?

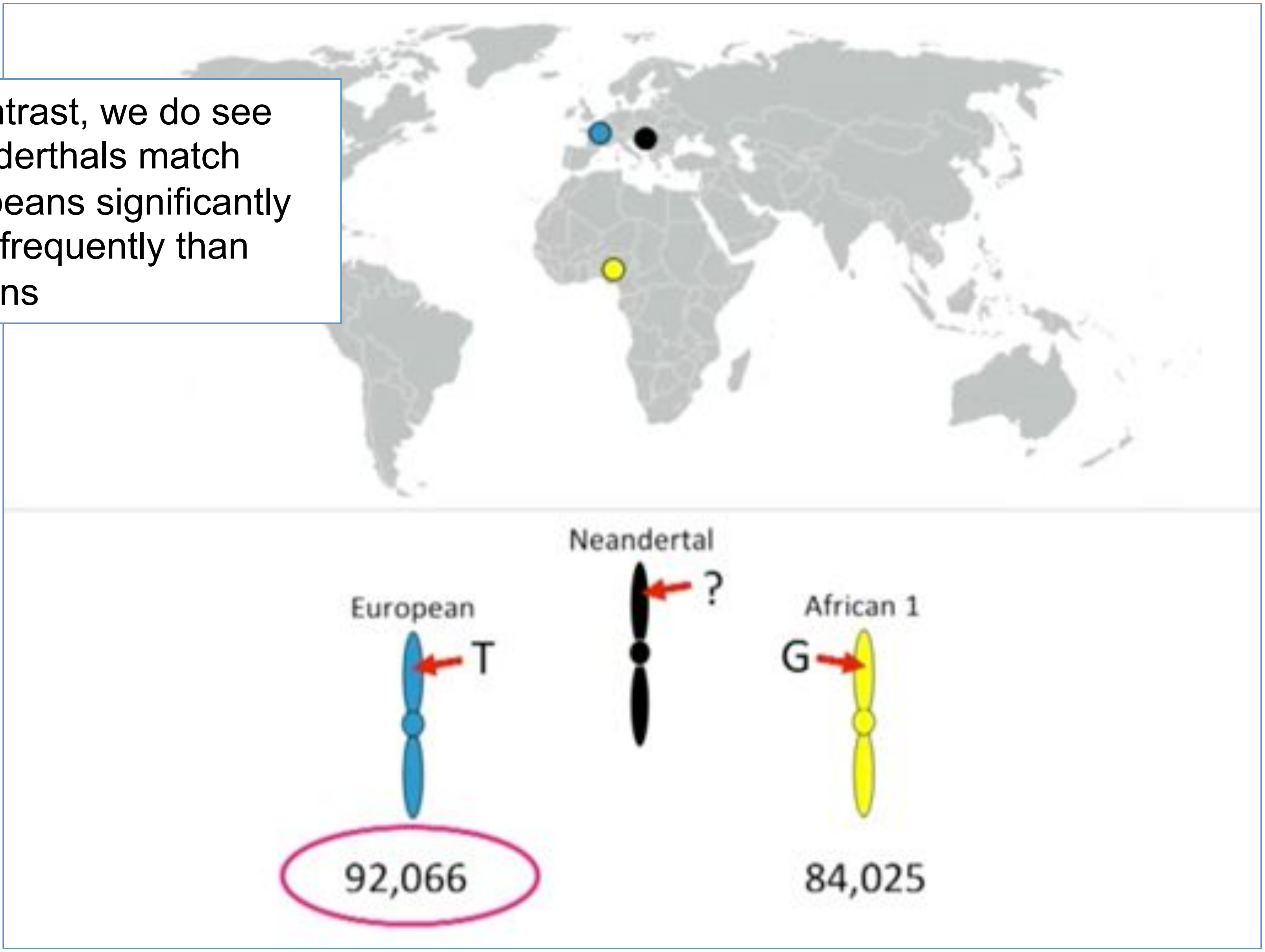


Did we mix?

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?

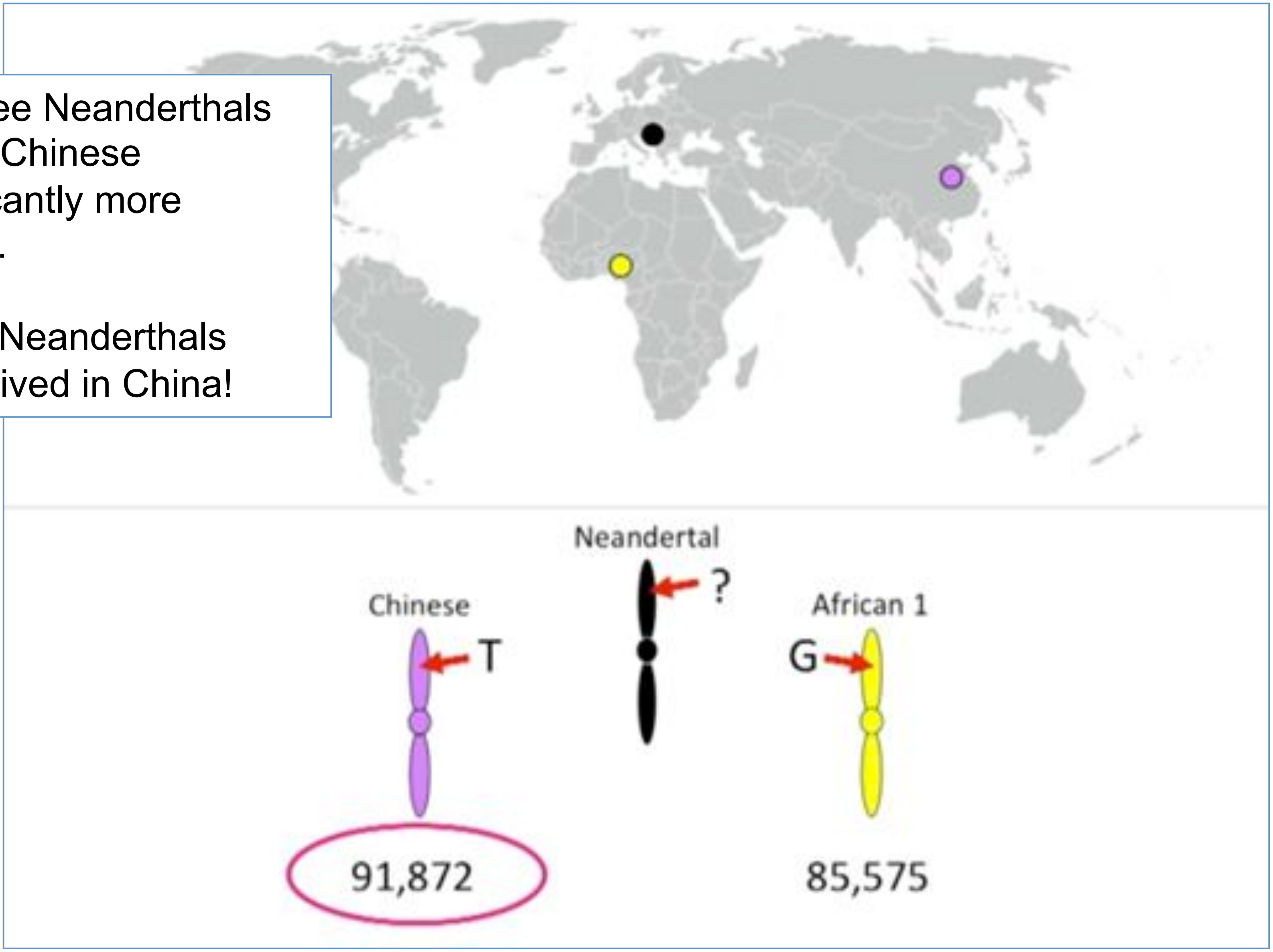


In contrast, we do see Neanderthals match Europeans significantly more frequently than Africans

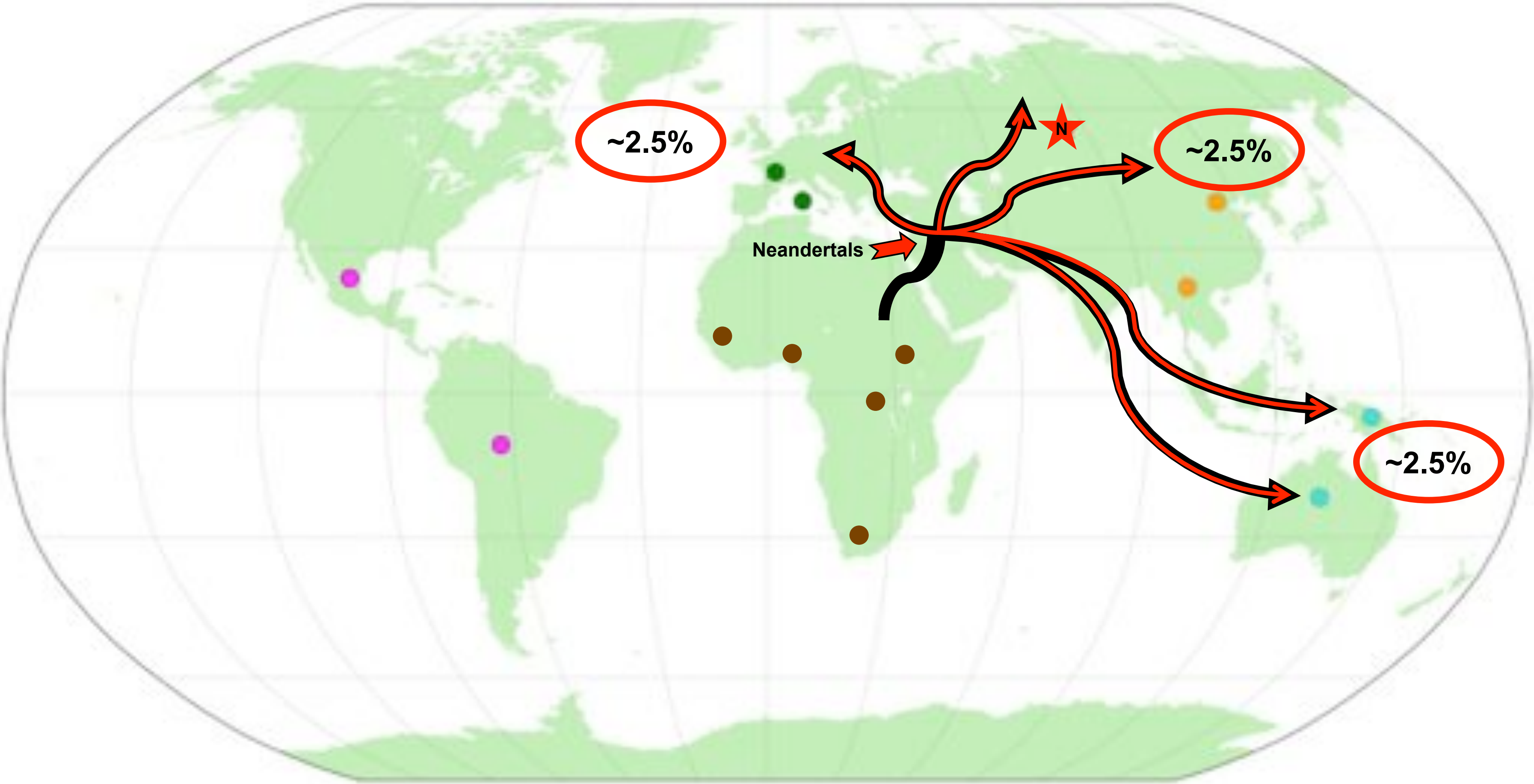
Did we mix?

Also see Neanderthals match Chinese significantly more often…

… but Neanderthals never lived in China!

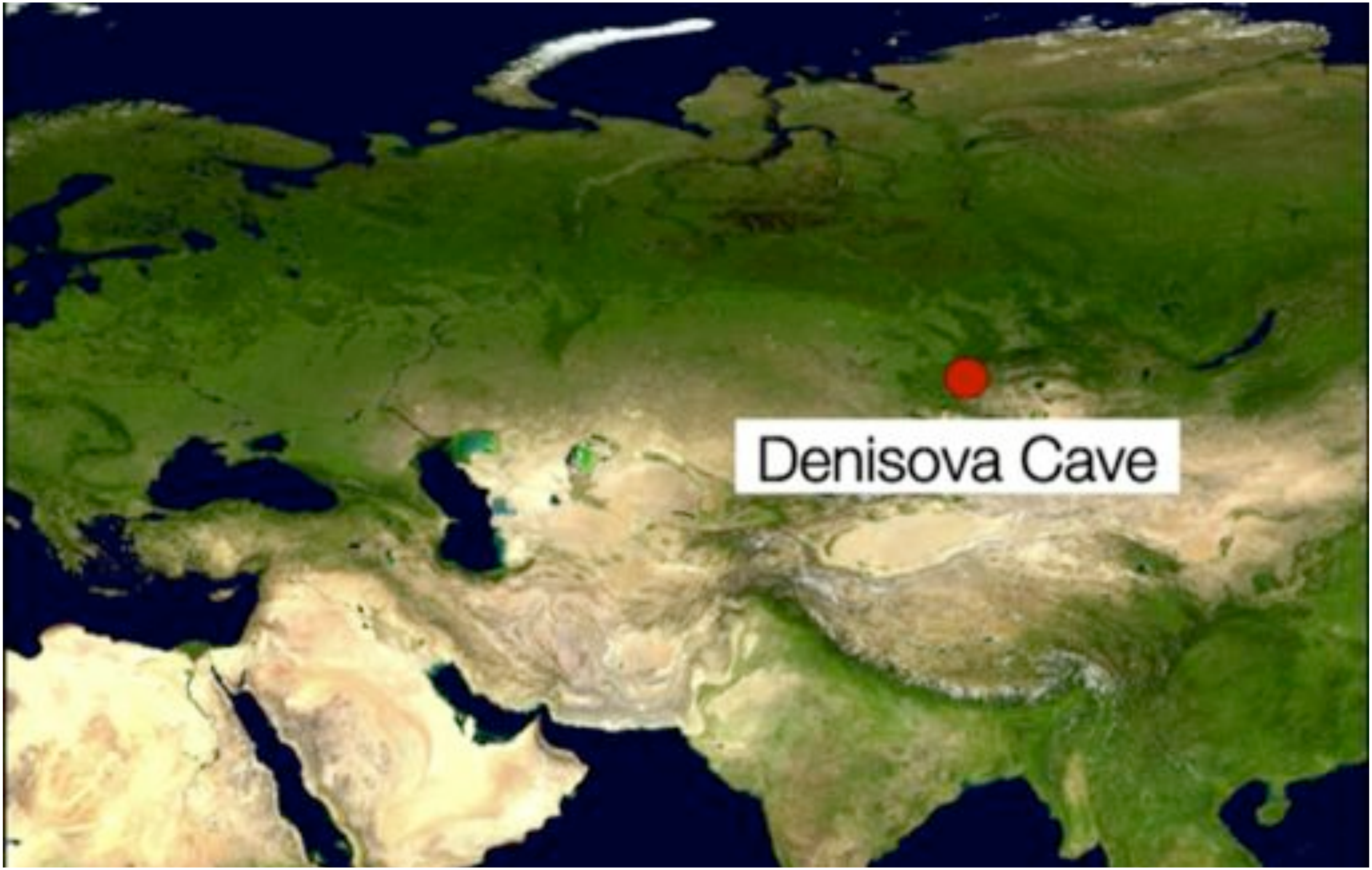


Neanderthal Interbreeding



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?

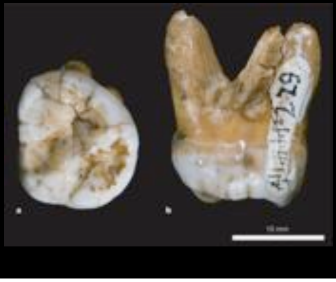


Denisova cave

Altai mountains

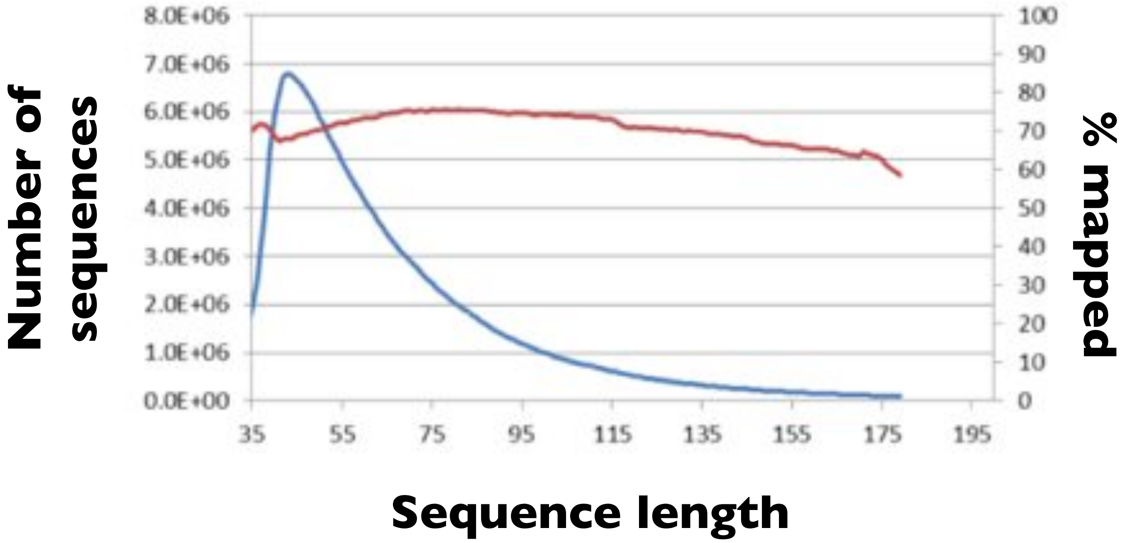
Russia

Academician A.P. Derevianko



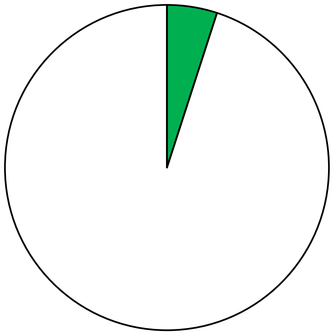
UNITED STATES OF AMERICA
E·PLURIBUS UNUM
ONE CENT

Extraordinary preservation

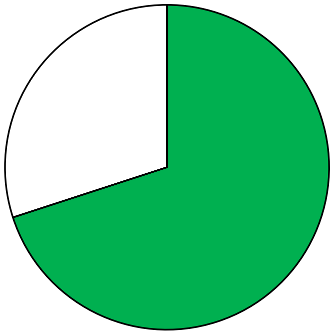


5% endogenous DNA

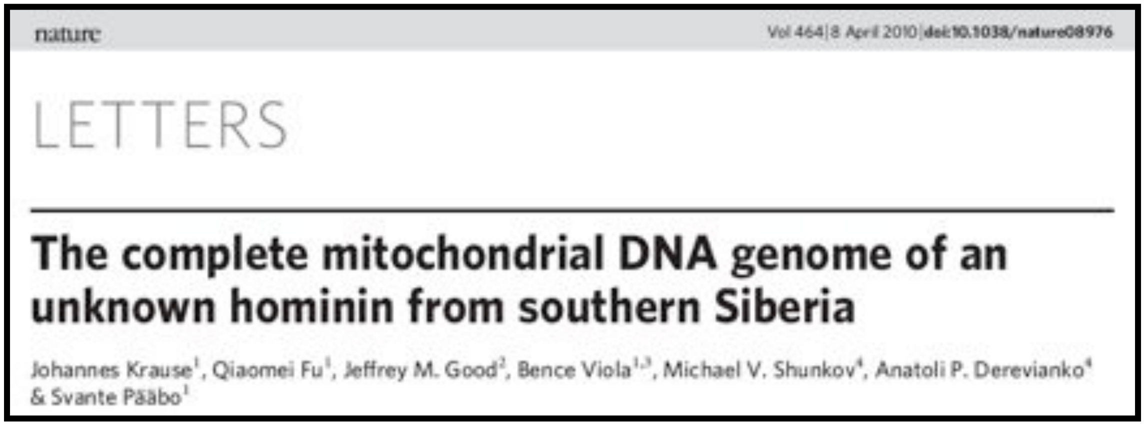
Best Neandertal bone



>70% endogenous DNA



Denisova bone

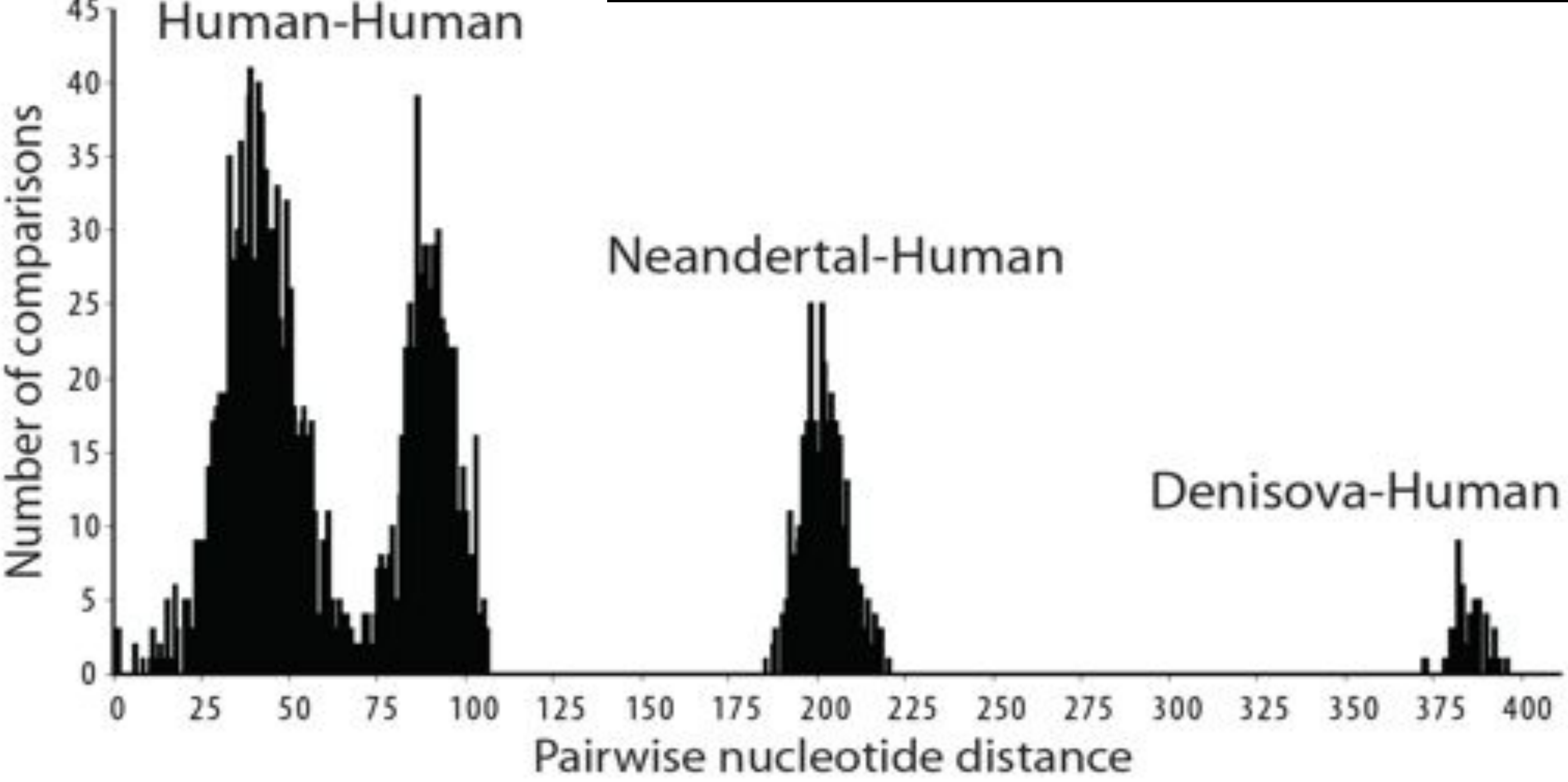


nature Vol 464|8 April 2010|doi:10.1038/nature08976

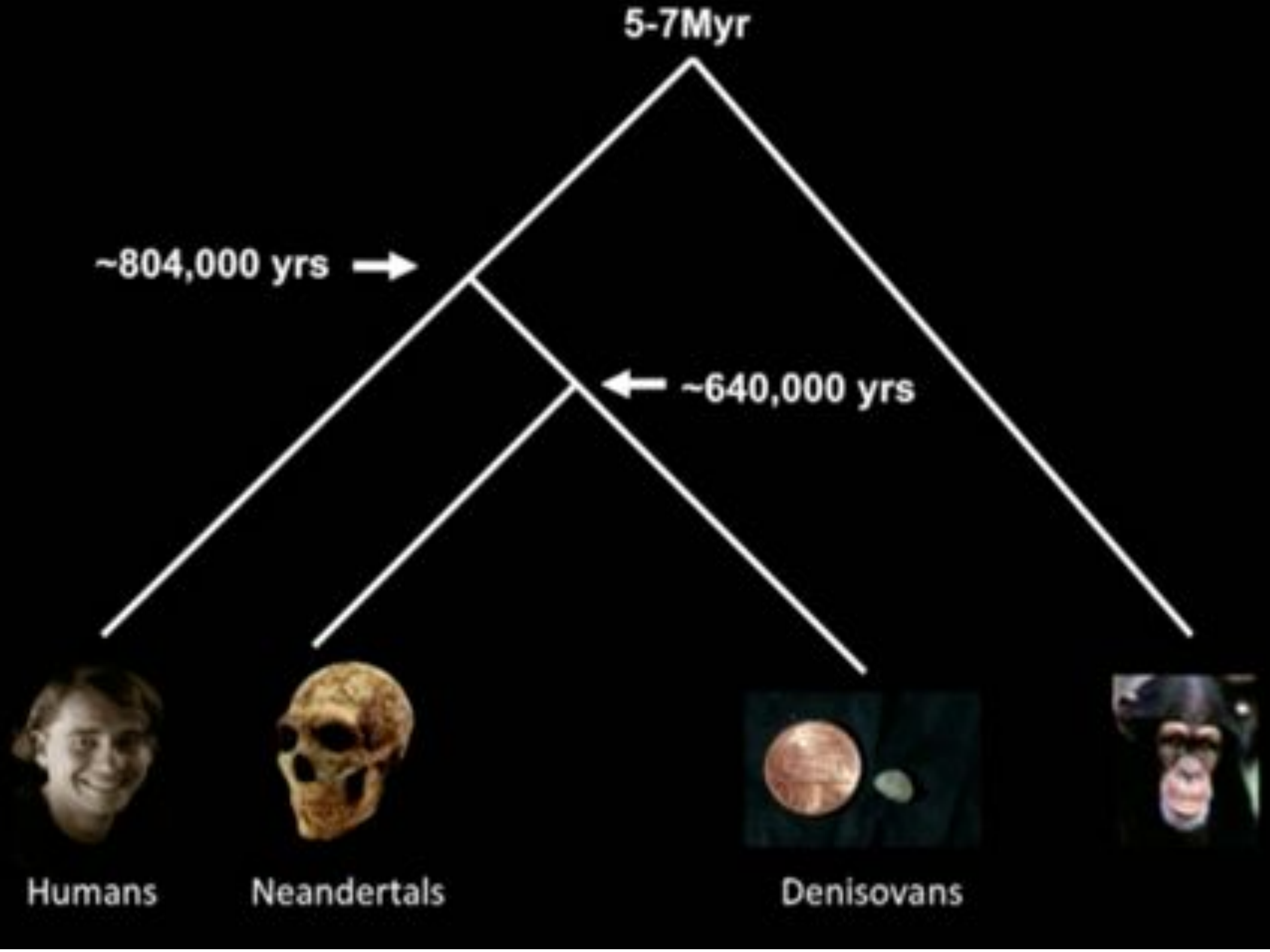
LETTERS

The complete mitochondrial DNA genome of an unknown hominin from southern Siberia

Johannes Krause[1], Qiaomei Fu[1], Jeffrey M. Good[2], Bence Viola[1,3], Michael V. Shunkov[4], Anatoli P. Derevianko[4] & Svante Pääbo[1]



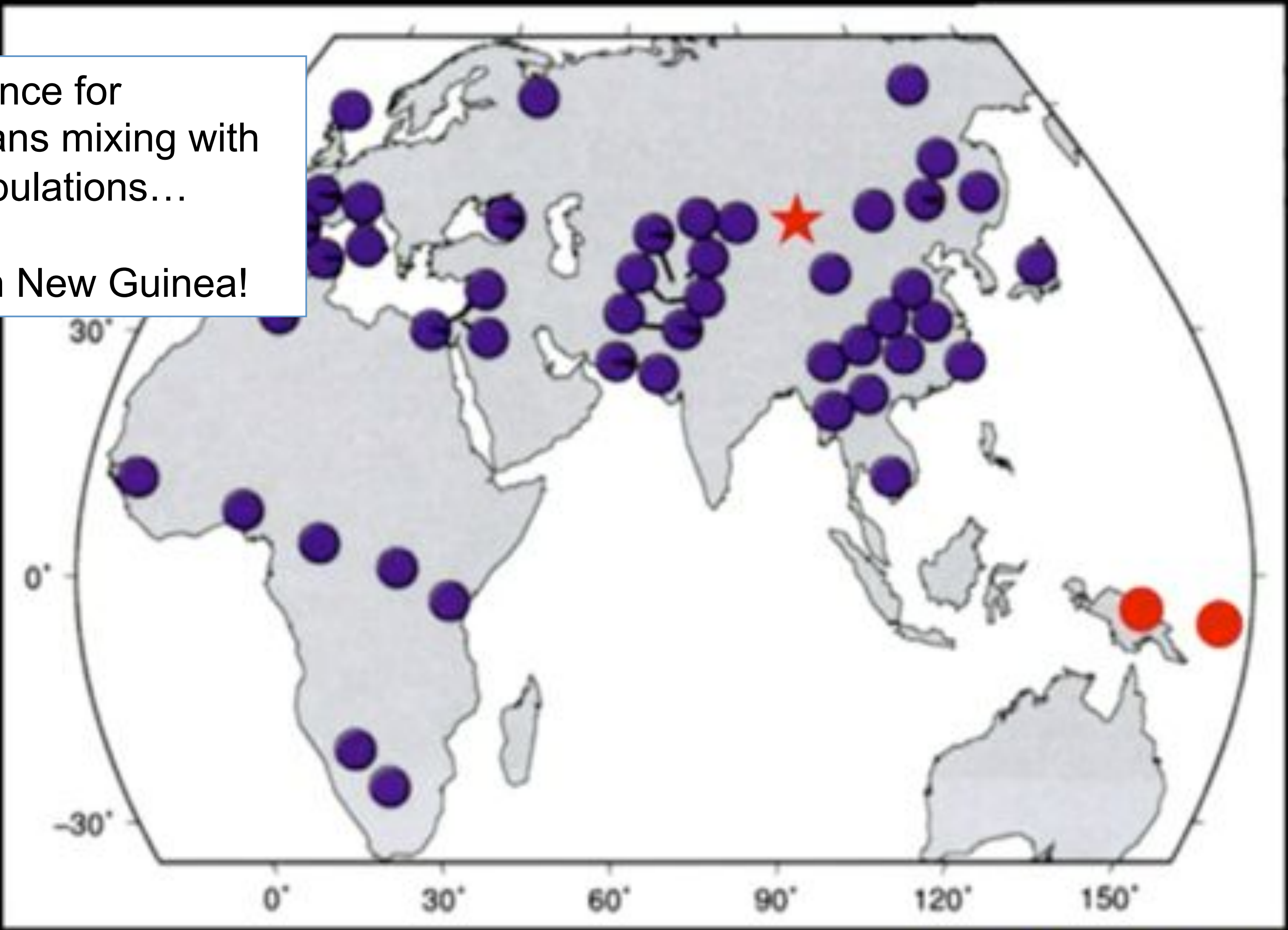
Denisovans & Neandertals



Did we mix?

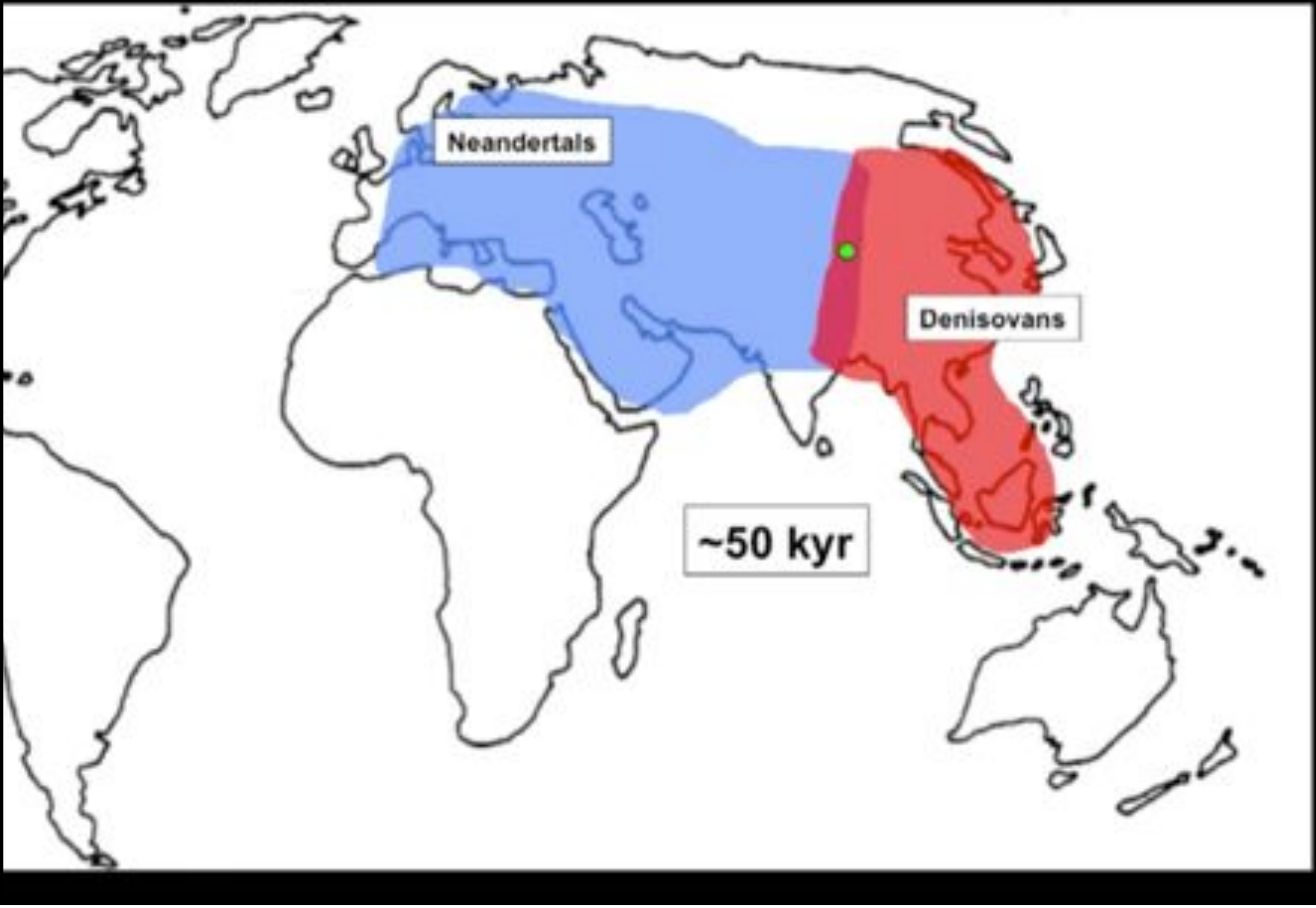
No evidence for Denisovans mixing with other populations…

Except in New Guinea!

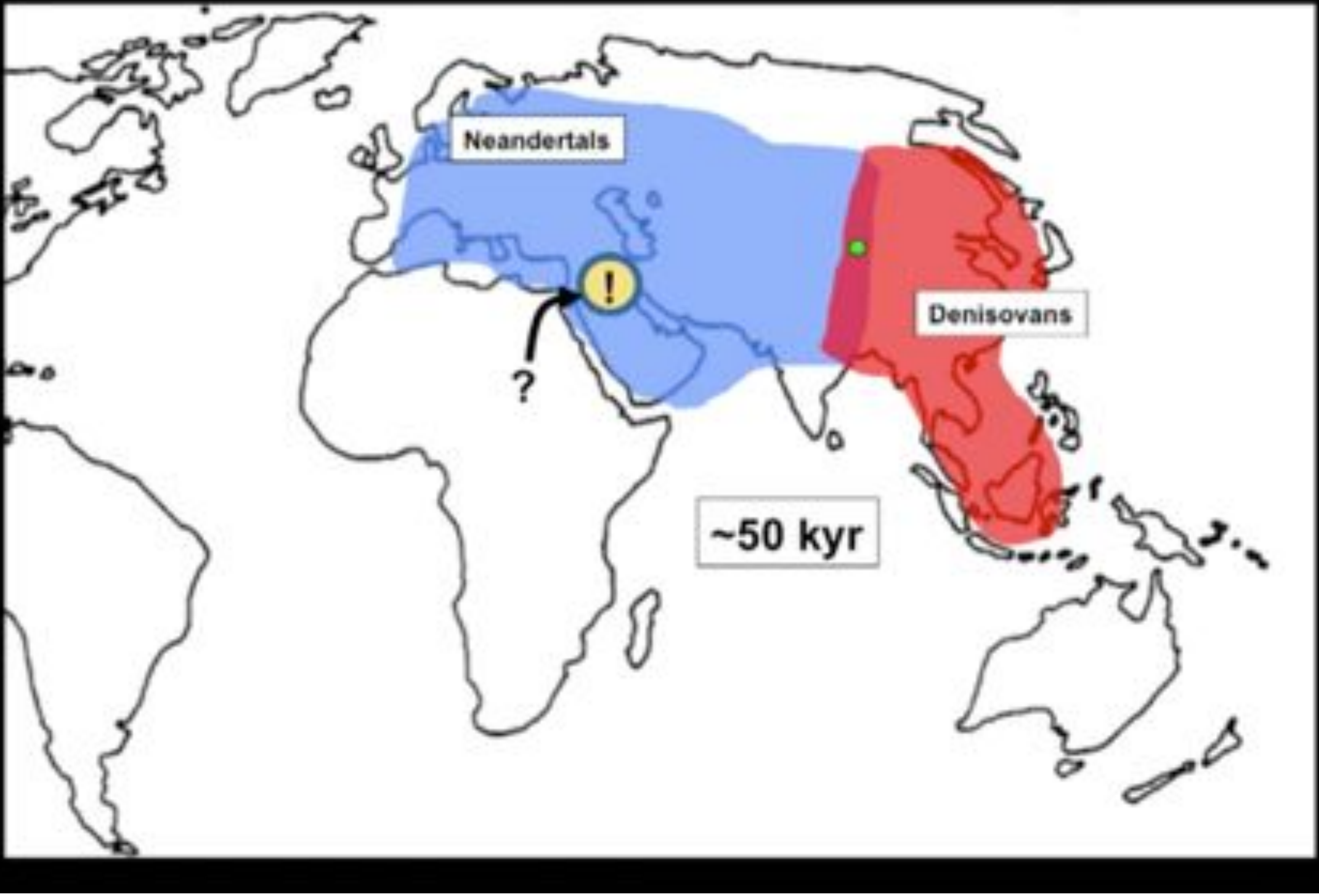


Map after Pickrell et al., 2009

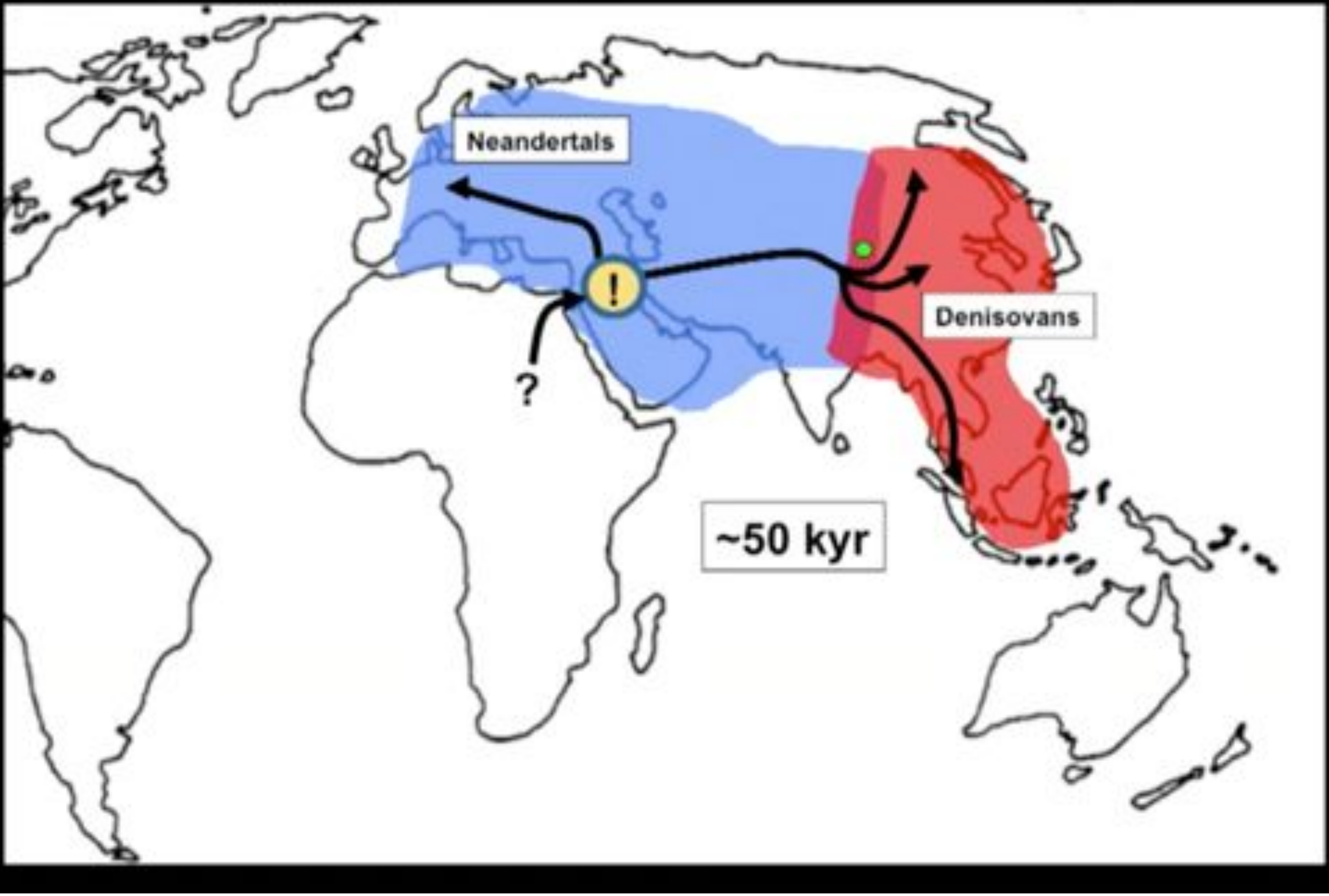
Timeline of ancient hominids



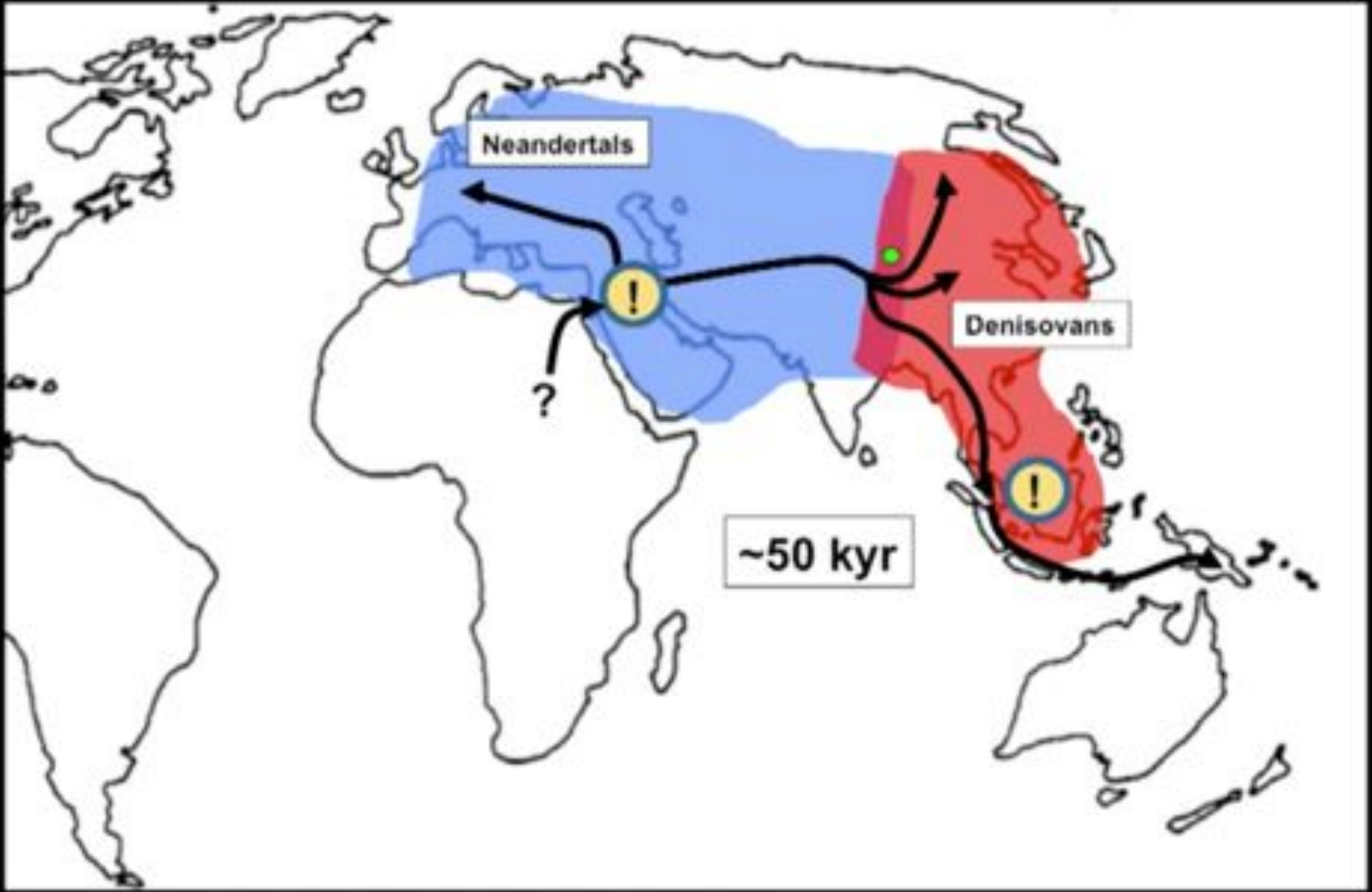
Timeline of ancient hominids



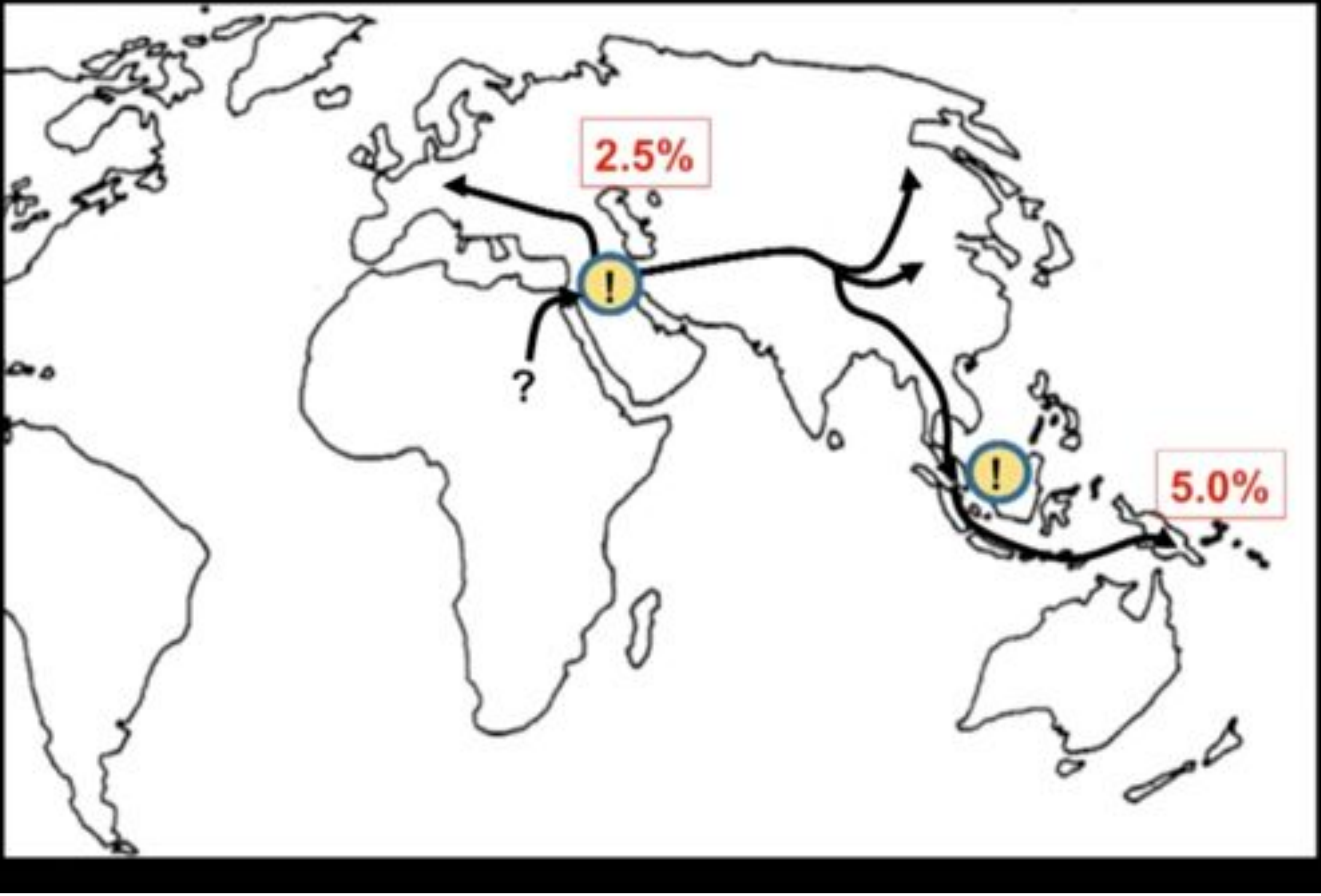
Timeline of ancient hominids



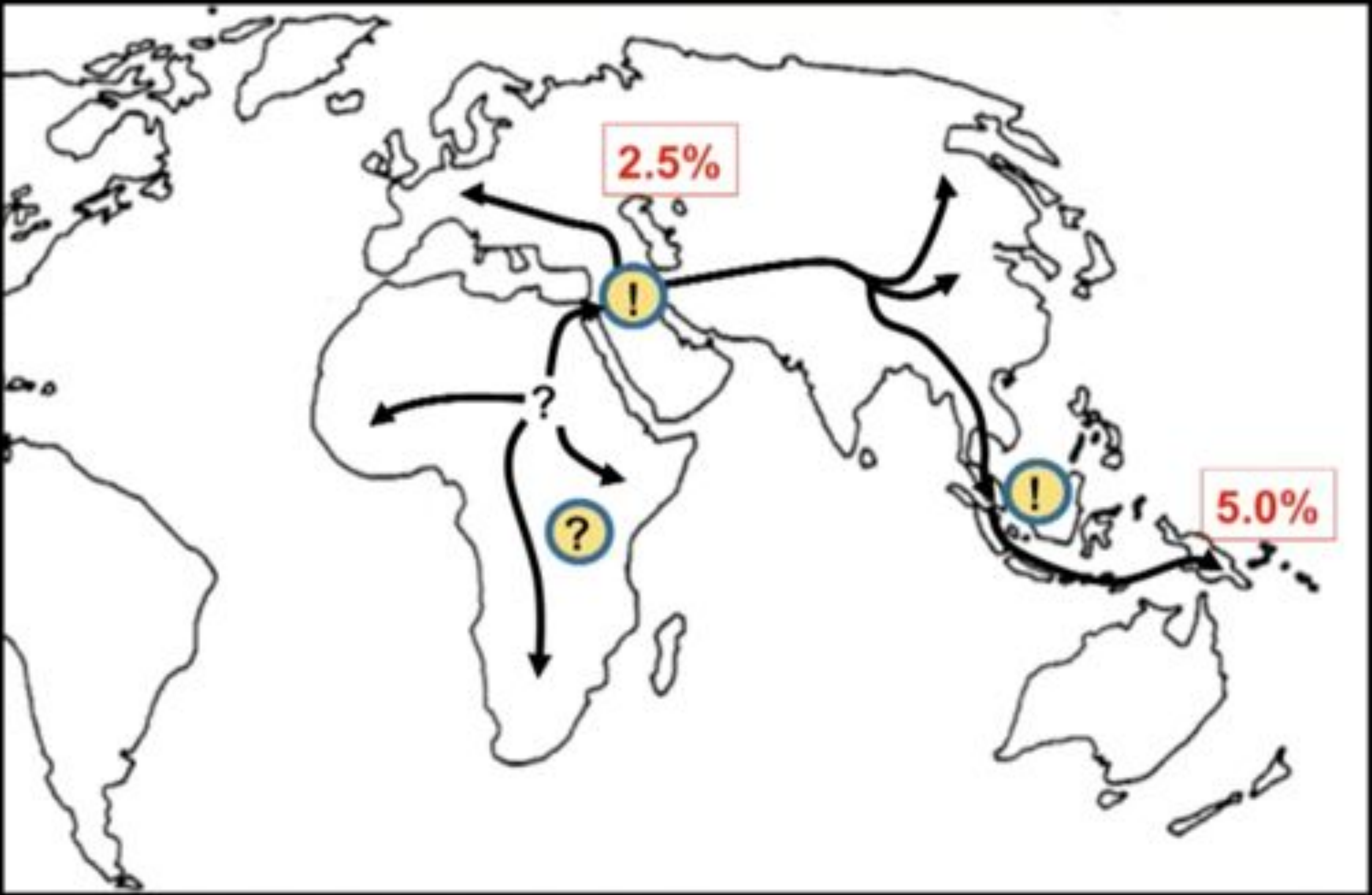
Timeline of ancient hominids



Timeline of ancient hominids

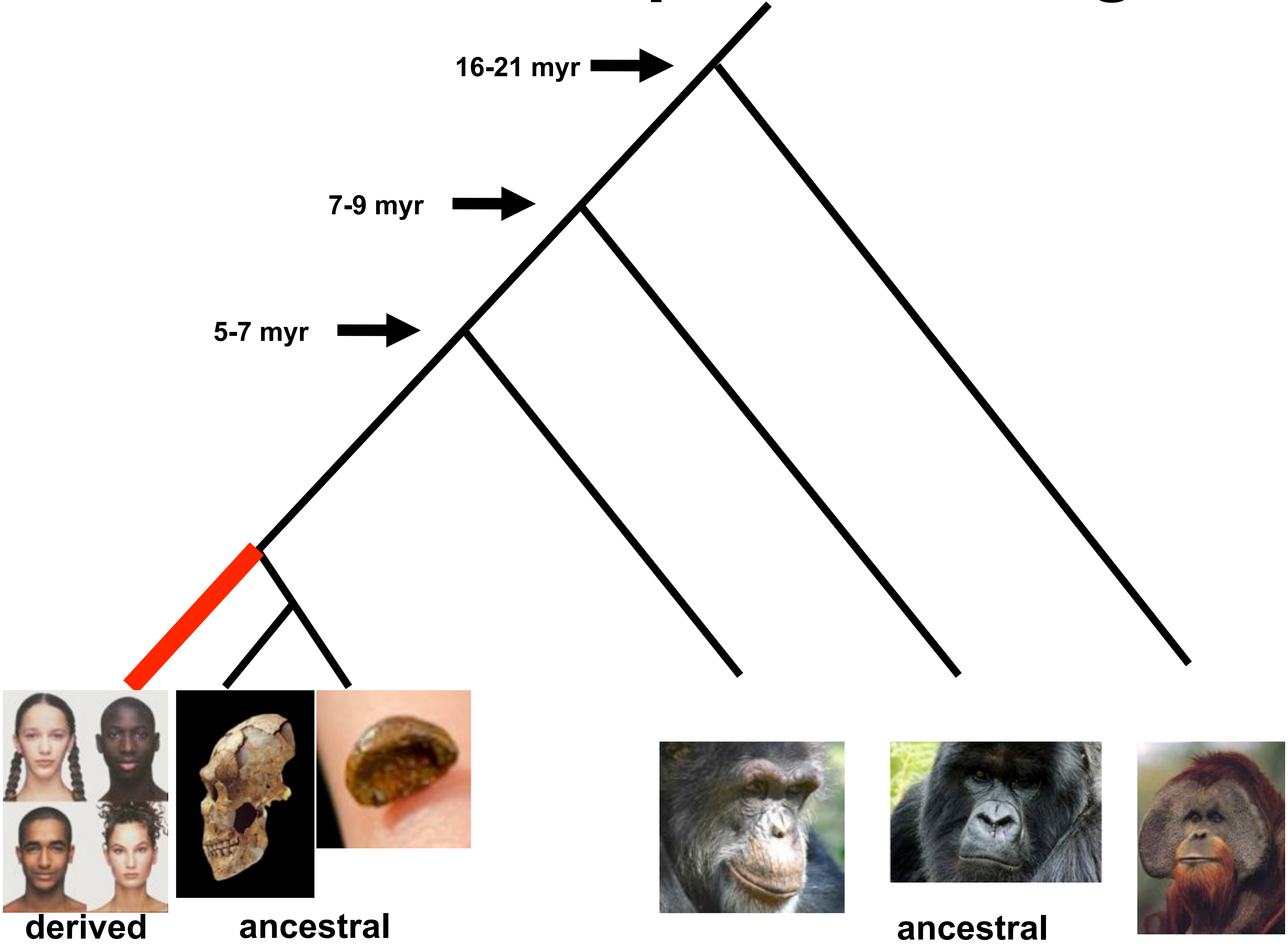


Timeline of ancient hominids



We have always mixed!

Modern human-specific changes



16-21 myr

7-9 myr

5-7 myr

derived

ancestral

ancestral

Recipe for a modern human

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

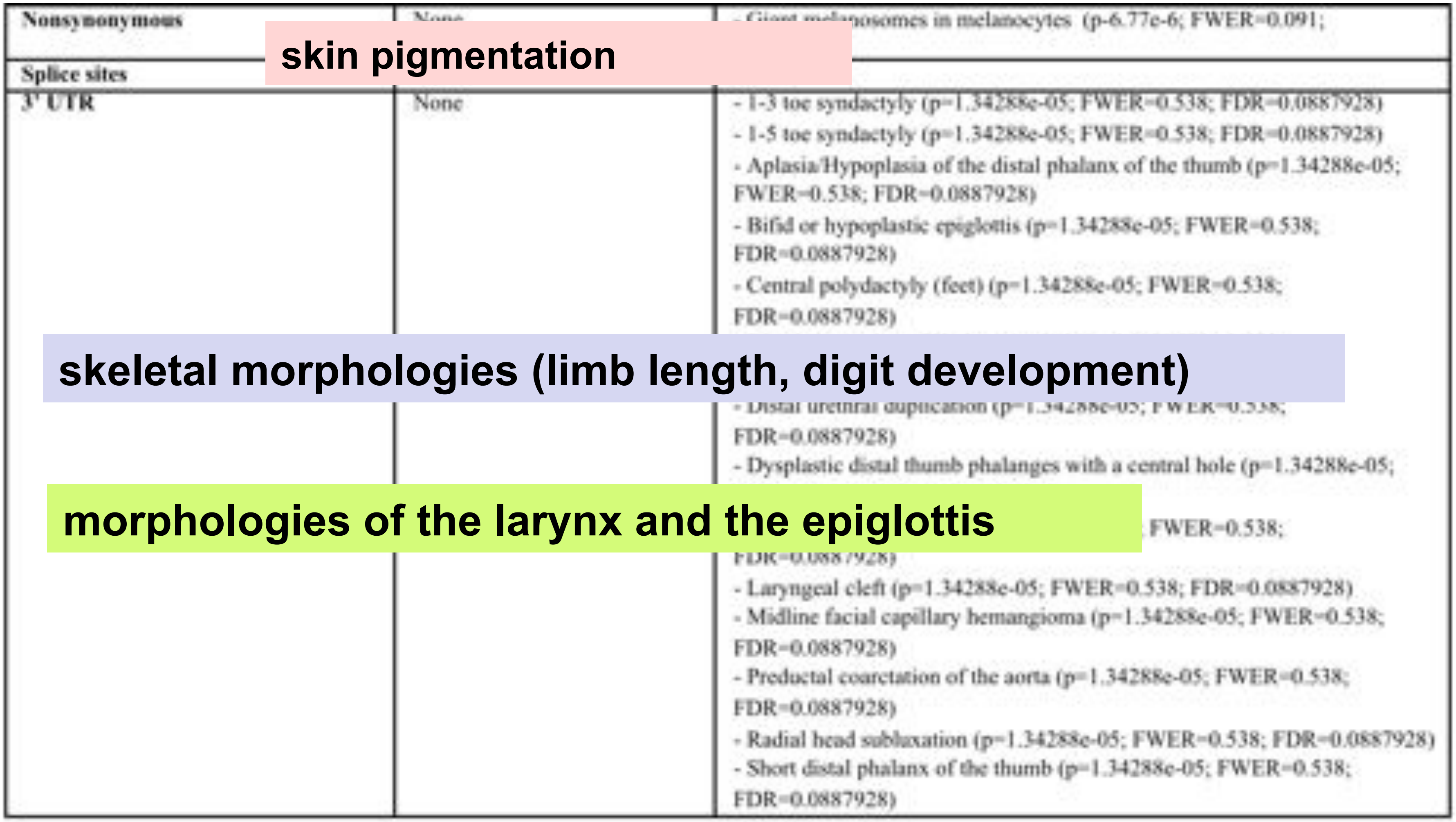
Changes in protein coding genes

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



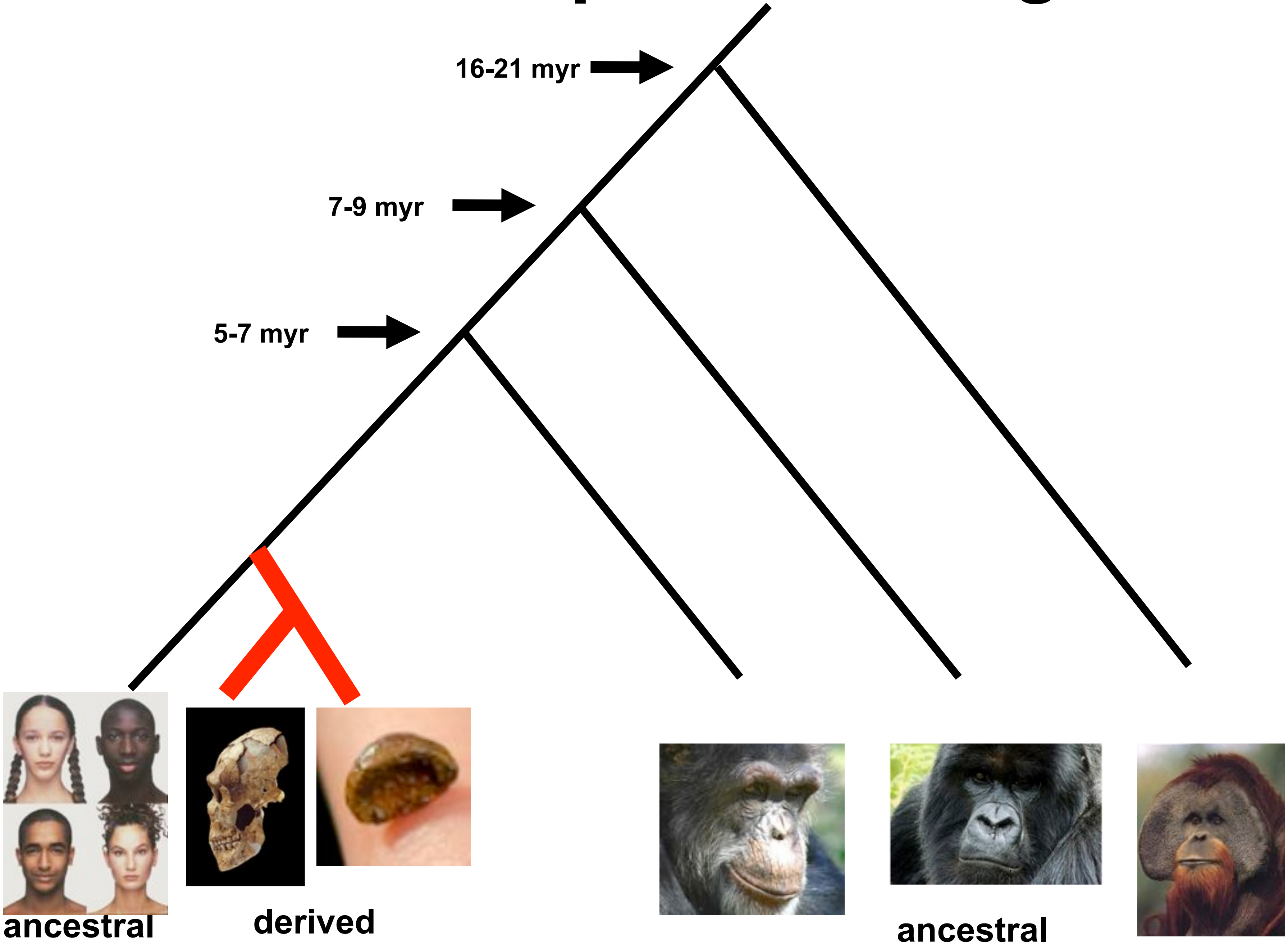
Nonsynonymous	None	- Giant melanosomes in melanocytes (p-6.77e-6; FWER=0.091;
Splice sites		
3' UTR	None	- 1-3 toe syndactyly (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - 1-5 toe syndactyly (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Aplasia/Hypoplasia of the distal phalanx of the thumb (p=1.34288e-05;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

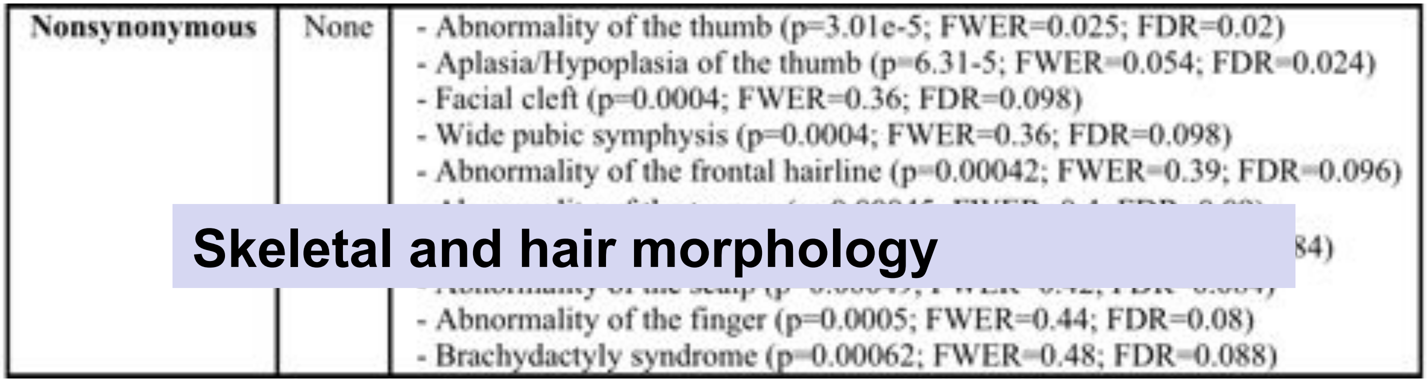
skeletal morphologies (limb length, digit development)

morphologies of the larynx and the epiglottis

Neandertal-specific changes

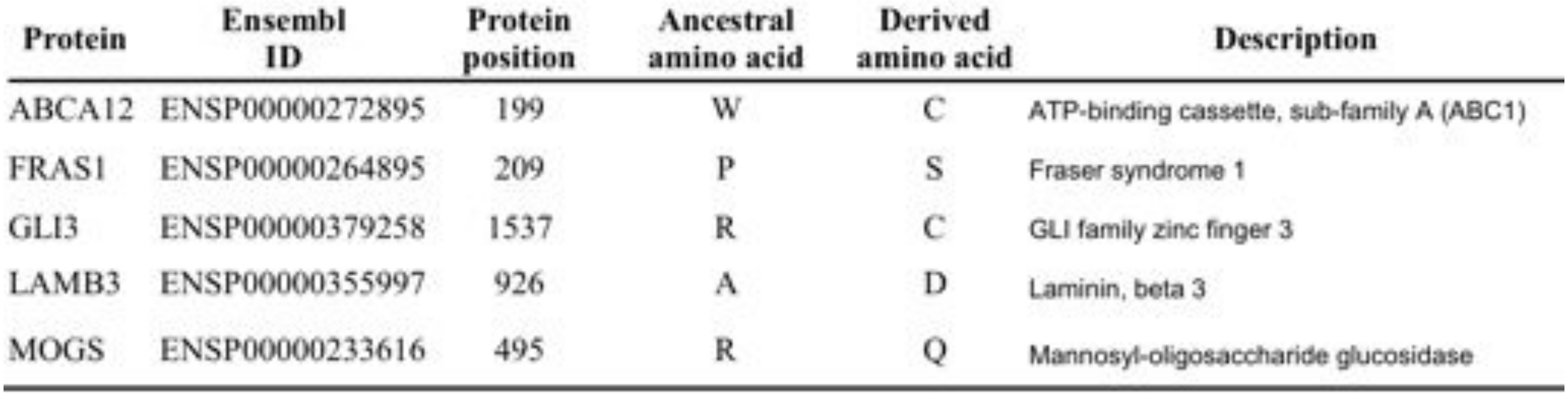


Enrichment analysis



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

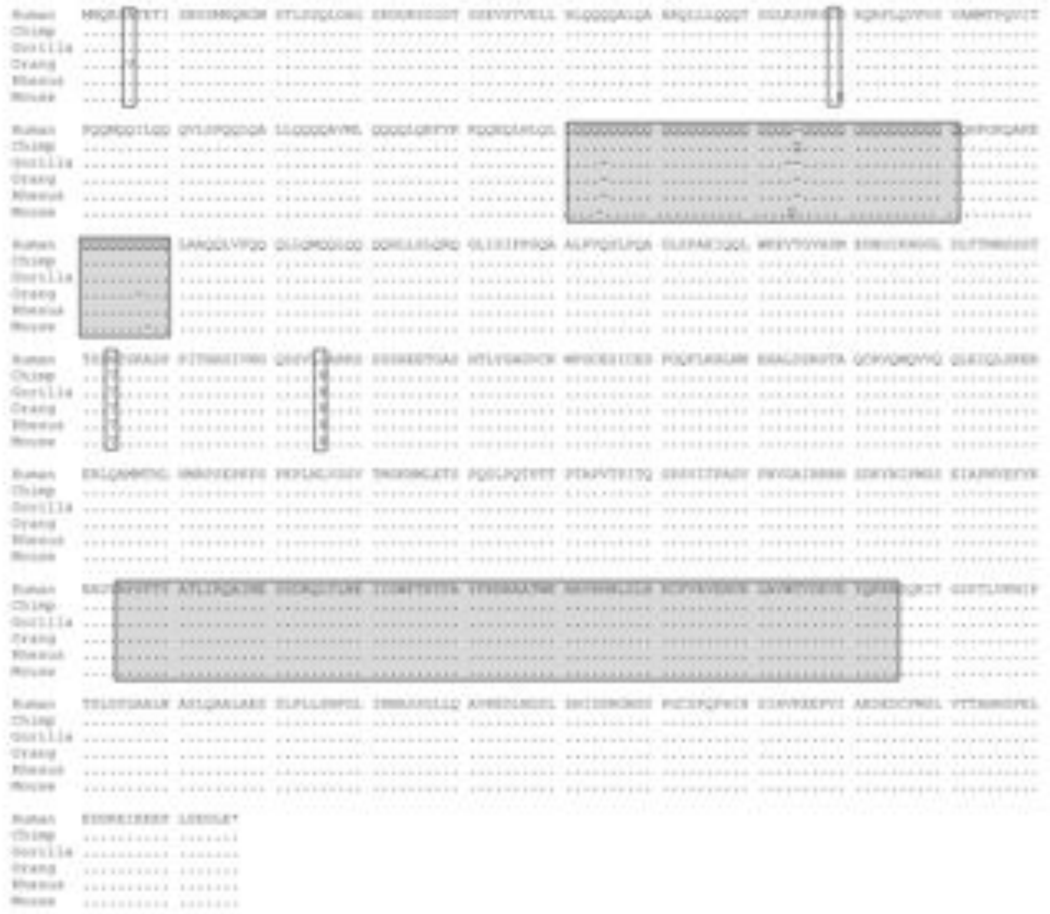

Human
Chimp
Gorilla
Orang
Rhesus
Mouse

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 f
human sequence are boxed.

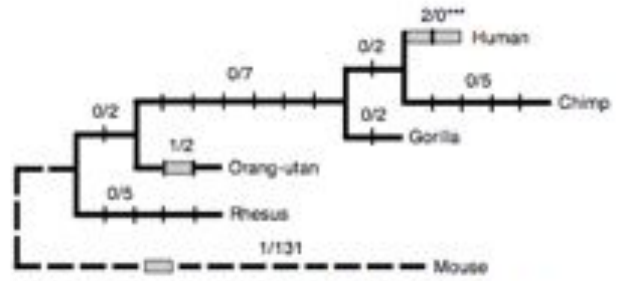


Figure 2 Silent and replacement nucleotide substitutions mapped on a phylogeny of primates. Bars represent nucleotide changes. Grey bars indicate amino-acid changes.

- 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.

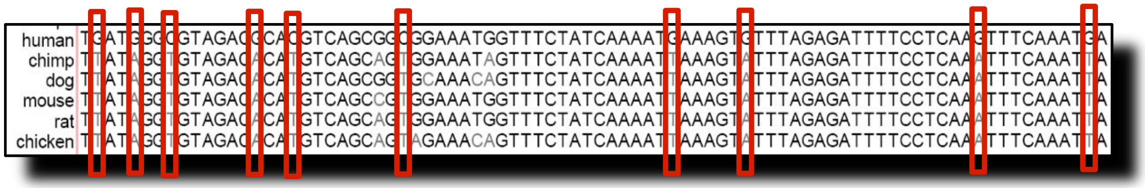
Molecular evolution of FOXP2, a gene involved in speech and language

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

What makes us human?

“*Human Accelerated Regions*”

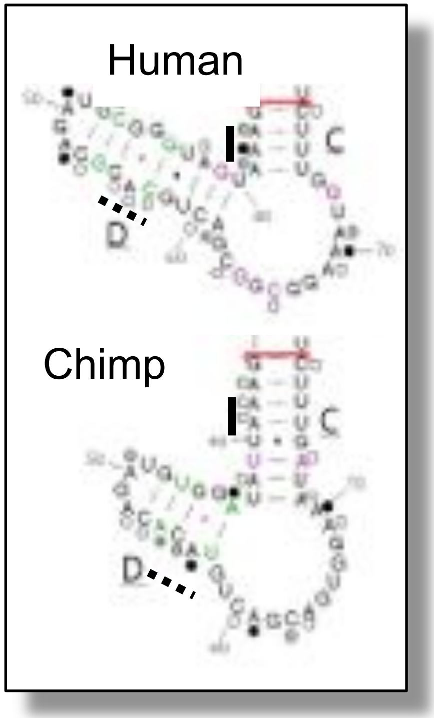
ANDREW SULLIVAN ON BLIND FAITH ■ LEO & JACK & MATT & MARTY
TIME
HOW WE BECAME HUMAN
Chimps and humans share almost 99% of their DNA. Now discoveries reveal how we can be so alike— and yet so different



```
human   TGATGGCGTAGACGCACGTCAGCGGCGGAAATGGTTTCTATCAAAATGAAAGTGTTTAGAGATTTTCCTCAAGTTTCAAATGA
chimp   TTATAGGTGTAGACACATGTCAGCAGTGGAAATAGTTTCTATCAAAATTAAAGTATTTAGAGATTTTCCTCAAATTTCAAATTA
dog     TTATAGGTGTAGACACATGTCAGCGGTGCAAACAGTTTCTATCAAAATTAAAGTATTTAGAGATTTTCCTCAAATTTCAAATTA
mouse   TTATAGGTGTAGACACATGTCAGCCGTGGAAATGGTTTCTATCAAAATTAAAGTATTTAGAGATTTTCCTCAAATTTCAAATTA
rat     TTATAGGTGTAGACACATGTCAGCAGTGGAAATGGTTTCTATCAAAATTAAAGTATTTAGAGATTTTCCTCAAATTTCAAATTA
chicken TTATAGGTGTAGACACATGTCAGCAGTAGAAACAGTTTCTATCAAAATTAAAGTATTTAGAGATTTTCCTCAAATTTCAAATTA
```

Systematic scan of recent human evolution identified the gene *HAR1F* as the most dramatic “human accelerated region”.

Follow up analysis found it was specifically expressed in Cajal-Retzius neurons in the human brain from 6 to 19 gestational weeks.



(Pollard et al., *Nature*, 2006)

Agenda

1. Clustering Refresher
   1. Hierarchical Clustering
   2. PCA

2. Ancient and Modern Human Evolution
   1. Modern Diversity
   2. Ancient Hominids

3. Genetic Privacy
   1. lobSTR and Microsatellites
   2. Surname inference

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 ➔ $(CT)_{13}$ tCAACAACAACAACAACAACAAa ➔ $(CAA)_{7}$

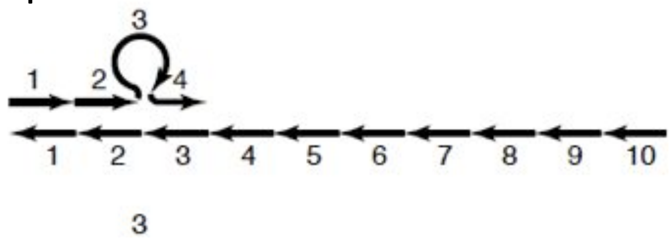
tTTGTCTTGTCTTGTCTTGTCTTGTCTTGTCc ➔ $(TTGTC)_{6}$ cCATTCATTCATTCATTa ➔ $(CATT)_{4}$

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

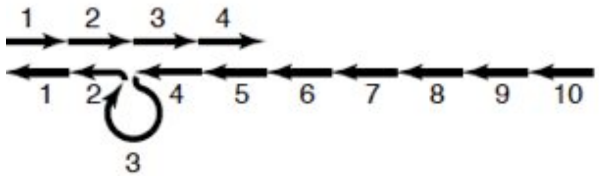
Replication slippage

- **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:



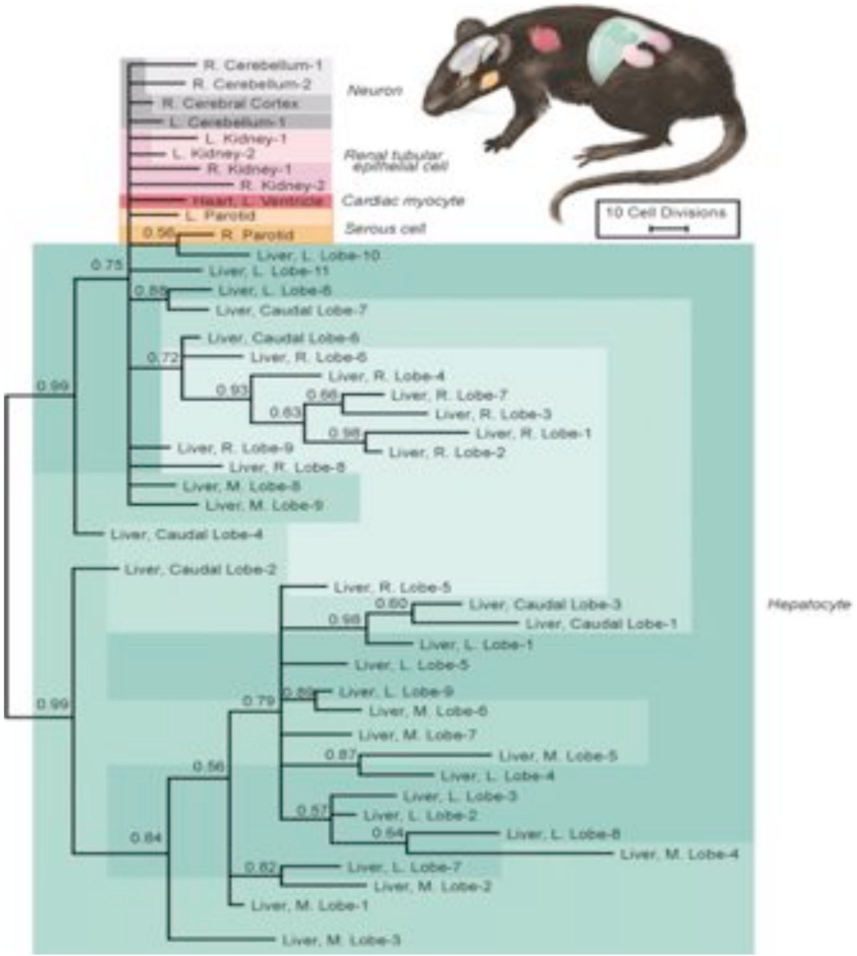
Contraction:



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

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

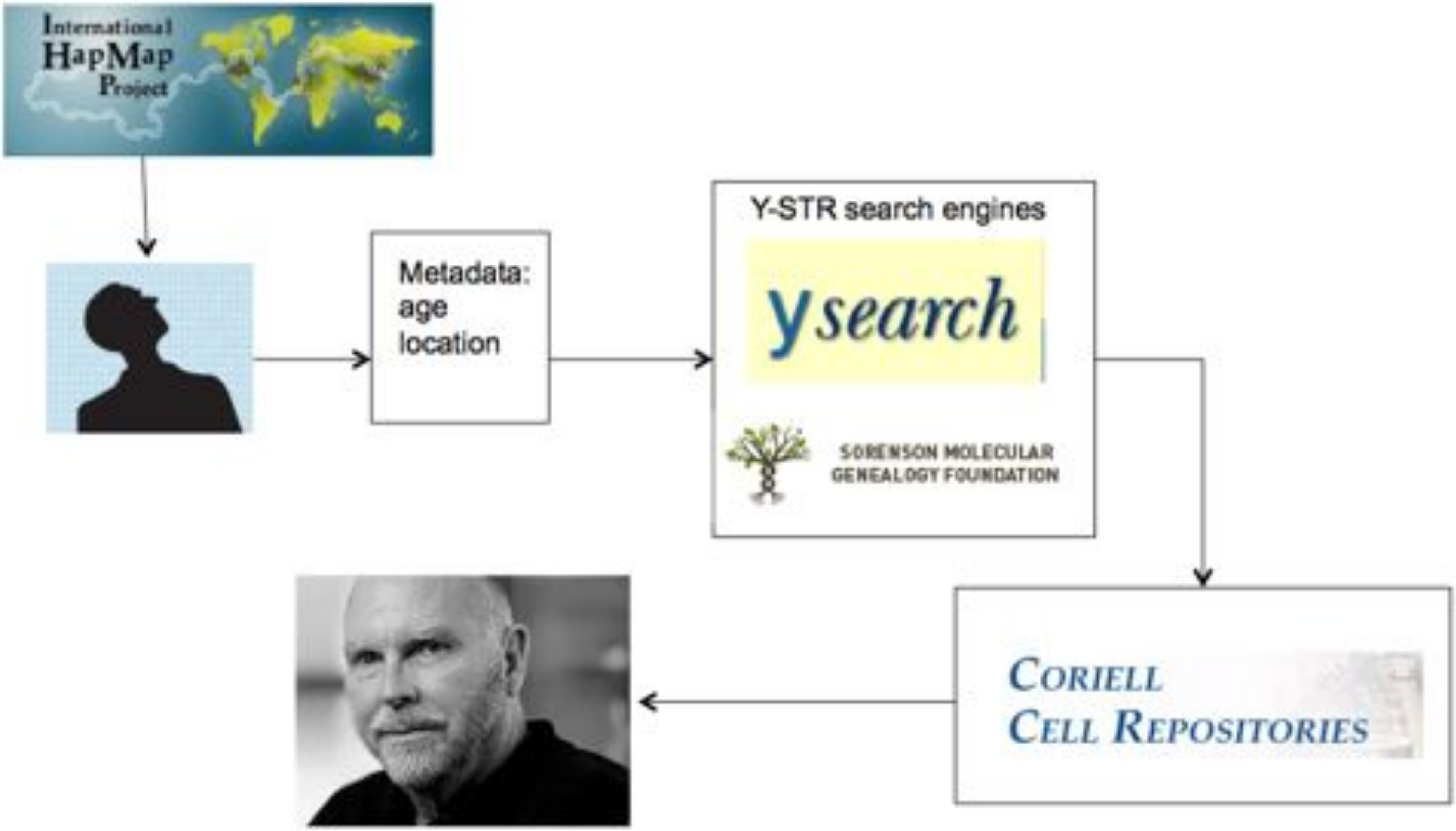
Genealogy Databases

DNA fingerprint
ysearch
SORENSON MOLECULAR GENEALOGY FOUNDATION
International HapMap Project
CORIELL CELL REPOSITORIES

GENETICS

Genealogy Databases Enable Naming Of Anonymous DNA Donors

Surname Inference Overview

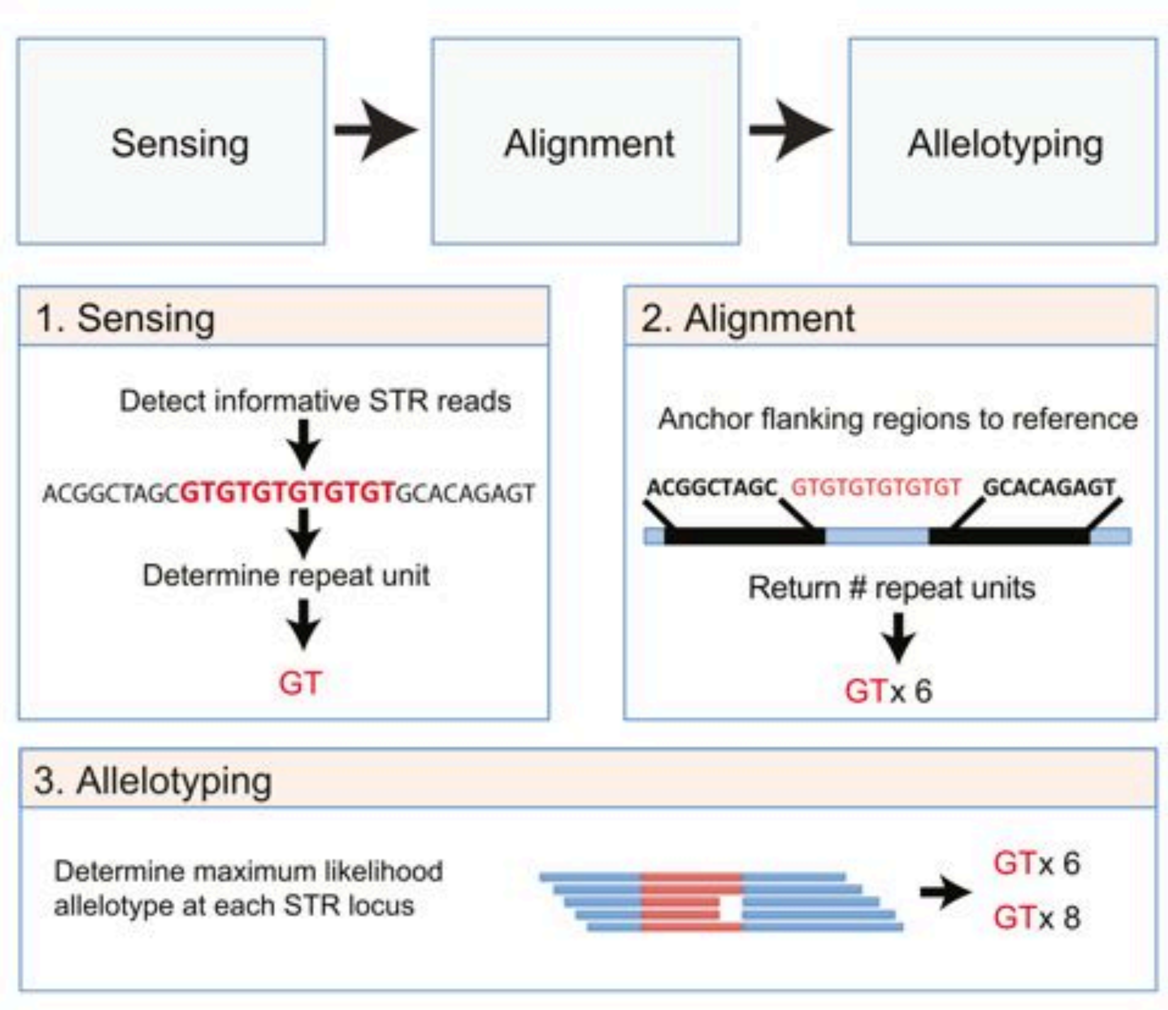


Metadata:
age
location

Y-STR search engines

International HapMap Project
ysearch
SORENSON MOLECULAR GENEALOGY FOUNDATION
CORIELL CELL REPOSITORIES

lobSTR Algorithm Overview



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

lobSTR Accuracy

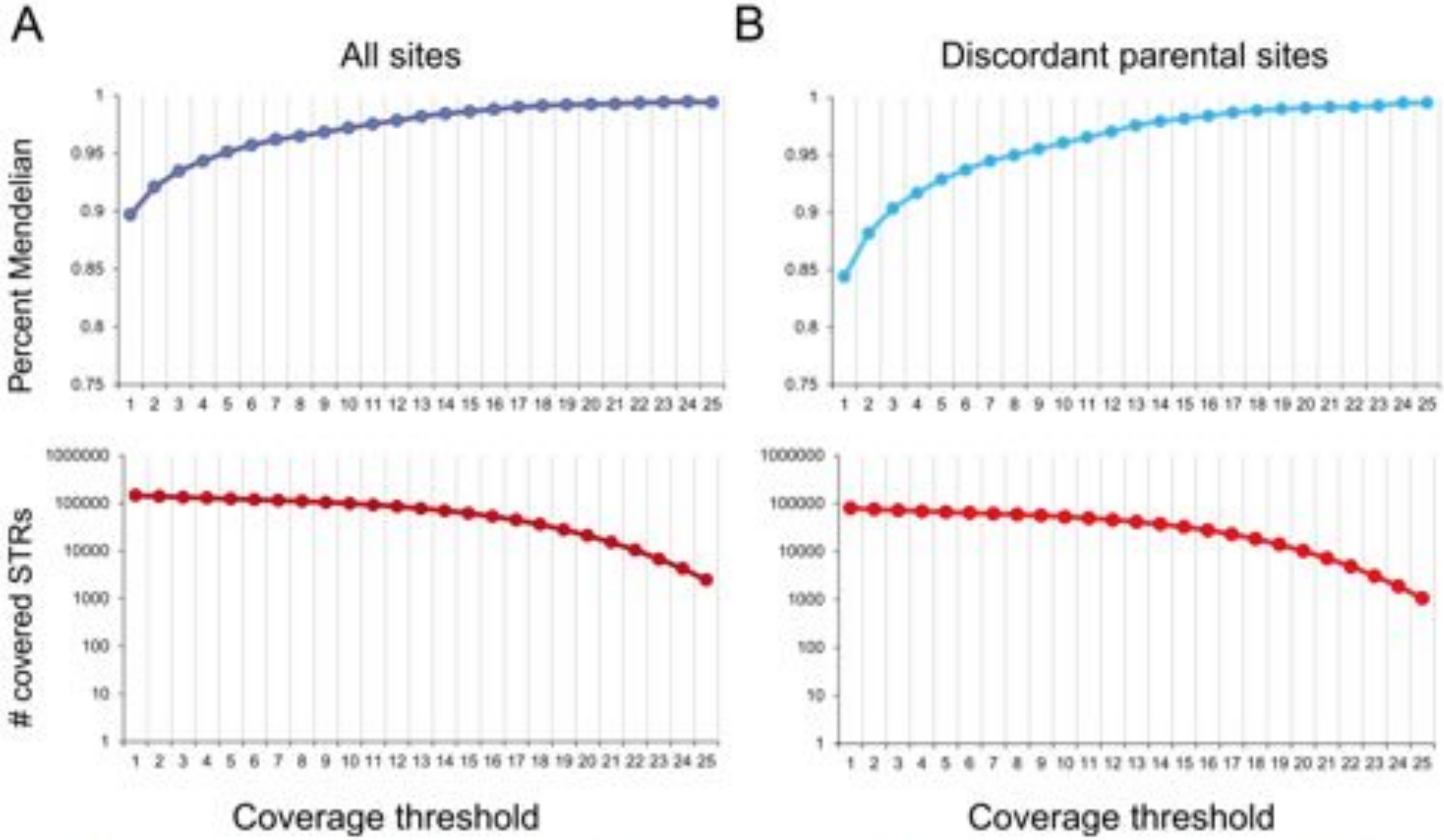
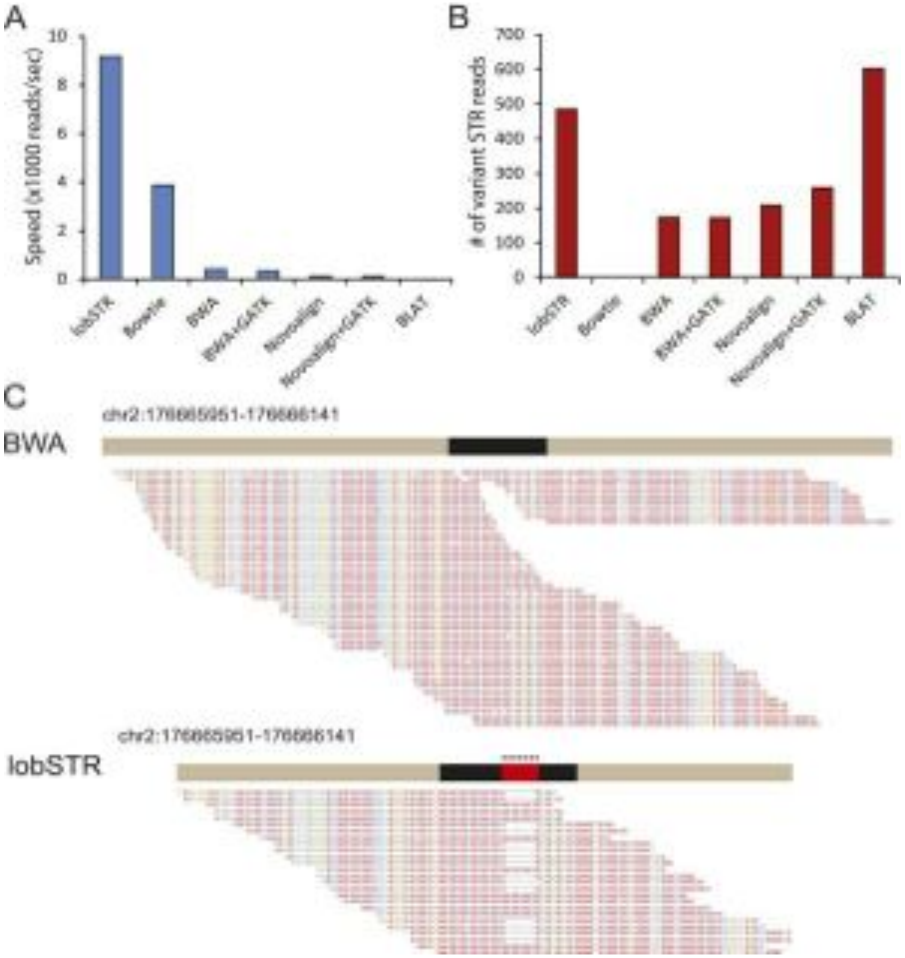


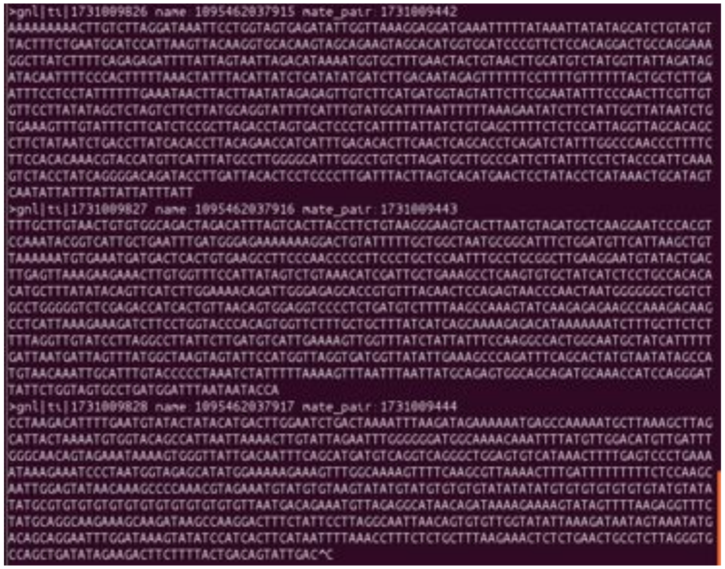
Figure 4. Validating lobSTR by Mendelian inheritance in a HapMap trio. Mendelian inheritance (blue and cyan) rose to 99% above 17× coverage. (Dark and light red) The number of covered loci at each coverage threshold. (*A*) Mendelian inheritance of all covered loci. (*B*) Mendelian inheritance of loci with discordant parental allelotypes.

lobSTR Performance



- LobSTR processes reads between 2.5 and 1000 times faster than mainstream aligners.
- Only BLAT detected more STR variations than lobSTR.
- LobSTR accurately detects pathogenic trinucleotide expansions that are normally discarded by mainstream aligners.
  - BWA only reports normal allele.
  - LobSTR identifies both alleles present at the simulated loci.

Surname Inference

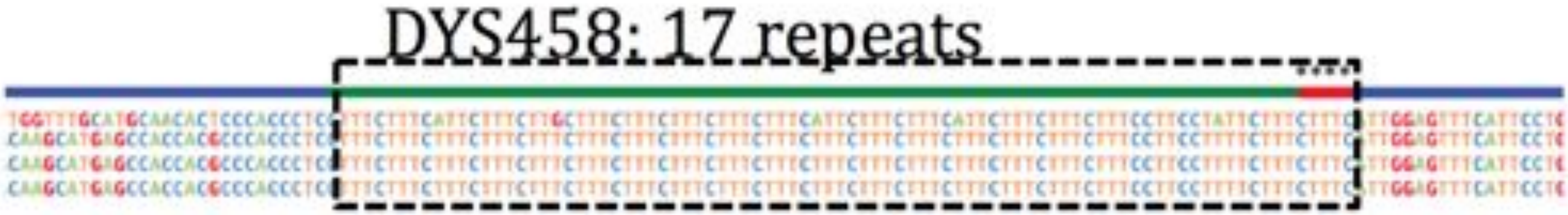


```
>gnl|ti|1731009826 name: 1095462037915 mate_pair: 1731009442
AAAAAAAAACTTGTCTTAGGATAAATTCCTGGTAGTGAGATATTGGTTAAAGGAGGATGAAATTTTTATAAATTATATAGCATCTGTATGT
TACTTTCTGAATGCATCCATTAAGTTACAAGGTGCACAAGTAGCAGAAGTAGCACATGGTGCATCCCGTTCTCCACAGGACTGCCAGGAAA
GGCTTATCTTTTCAGAGAGATTTTATTAGTAATTAGACATAAAATGGTGCTTTGAACTACTGTAACTTGCATGTCTATGGTTATTAGATAG
ATACAATTTTCCCACTTTTTAAACTATTTACATTATCTCATATATGATCTTGACAATAGAGTTTTTTCCTTTTGTTTTTTACTGCTCTTGA
ATTTCCTCCTATTTTTTGAAATAACTTACTTAATATAGAGAGTTGTCTTCATGATGGTAGTATTCTTCGCAATATTTCCCAACTTCGTTGT
GTTCCTTATATAGCTCTAGTCTTCTTATGCAGGTATTTTCATTTGTATGCATTTAATTTTTTAAAGAATATCTTCTATTGCTTATAATCTG
TGAAAGTTTGTATTTCTTCATCTCCGCTTAGACCTAGTGACTCCCTCATTTTATTATCTGTGAGCTTTTCTCTCCATTAGGTTAGCACAGC
CTTCTATAATCTGACCTTATCACACCTTACAGAACCATCATTTGACACACTTCAACTCAGCACCTCAGATCTATTTGGCCCAACCCTTTTC
TTCCACACAAACGTACCATGTTCATTTATGCCTTGGGGCATTTGGCCTGTCTTAGATGCTTGCCCATTCTTATTTCCTCTACCCATTCAAA
GTCTACCTATCAGGGGACAGATACCTTGATTACACTCCTCCCCTTGATTTACTTAGTCACATGAACTCCTATACCTCATAAACTGCATAGT
CAATATTATTTATTATTATTTATT
>gnl|ti|1731009827 name: 1095462037916 mate_pair: 1731009443
TTTGCTTGTAACTGTGTGGCAGACTAGACATTTAGTCACTTACCTTCTGTAAGGGAAGTCACTTAATGTAGATGCTCAAGGAATCCCACGT
CCAAATACGGTCATTGCTGAATTTGATGGGAGAAAAAAAGGACTGTATTTTTGCTGGCTAATGCGGCATTTCTGGATGTTCATTAAGCTGT
TAAAAAATGTGAAATGATGACTCACTGTGAAGCCTTCCCAACCCCCTTCCCTGCTCCAATTTGCCTGCGGCTTGAAGGAATGTATACTGAC
TTGAGTTAAAGAAGAAACTTGTGGTTTCCATTATAGTCTGTAAACATCGATTGCTGAAAGCCTCAAGTGTGCTATCATCTCCTGCCACACA
CATGCTTTATATACAGTTCATCTTGGAAAACAGATTGGGAGAGCACCGTGTTTACAACTCCAGAGTAACCCAACTAATGGGGGGCTGGTCT
GCCTGGGGGTCTCGAGACCATCACTGTTAACAGTGGAGGTCCCCTCTGATGTCTTTTAAGCCAAAGTATCAAGAGAGAAGCCAAAGACAAG
CCTCATTAAAGAAAGATCTTCCTGGTACCCACAGTGGTTCTTTGCTGCTTTATCATCAGCAAAAGAGACATAAAAAAATCTTTGCTTCTCT
TTTAGGTTGTATCCTTAGGCCTTATTCTTGATGTCATTGAAAAGTTGGTTTATCTATTATTTCCAAGGCCACTGGCAATGCTATCATTTTT
GATTAATGATTAGTTTATGGCTAAGTAGTATTCCATGGTTAGGTGATGGTTATATTGAAAGCCCAGATTTCAGCACTATGTAATATAGCCA
TGTAACAAATTGCATTTGTACCCCCTAAATCTATTTTTAAAAGTTTAATTTAATTATGCAGAGTGGCAGCAGATGCAAACCATCCAGGGAT
TATTCTGGTAGTGCCTGATGGATTTAATAATACCA
>gnl|ti|1731009828 name: 1095462037917 mate_pair: 1731009444
CCTAAGACATTTTGAATGTATACTATACATGACTTGGAATCTGACTAAAATTTAAGATAGAAAAAATGAGCCAAAAATGCTTAAAGCTTAG
CATTACTAAAATGTGGTACAGCCATTAATTAAAACTTGTATTAGAATTTGGGGGGATGGCAAAACAAATTTTATGTTGGACATGTTGATTT
GGGCAACAGTAGAAATAAAAGTGGGTTATTGACAATTTCAGCATGATGTCAGGTCAGGGCTGGAGTGTCATAAACTTTTGAGTCCCTGAAA
ATAAAGAAATCCCTAATGGTAGAGCATATGGAAAAAGAAAGTTTGGCAAAAGTTTTCAAGCGTTAAAACTTTGATTTTTTTTTCTCCAAGC
AATTGGAGTATAACAAAGCCCCAAACGTAGAAATGTATGTGTAAGTATATGTATGTGTGTGTATATATATGTGTGTGTGTGTATGTATA
TATGCGTGTGTGTGTGTGTGTGTGTGTGTGTTAATGACAGAAATGTTAGAGGCATAACAGATAAAAGAAAAGTATAGTTTTAAGAGGTTTC
TATGCAGGCAAGAAAGCAAGATAAGCCAAGGACTTTCTATTCCTTAGGCAATTAACAGTGTGTTGGTATATTAAAGATAATAGTAAATATG
ACAGCAGGAATTTGGATAAAGTATATCCATCACTTCATAATTTTAAACCTTTCTCTGCTTTAAGAACTCTCTGAACTGCCTCTTAGGGTG
CCAGCTGATATAGAAGACTTCTTTTACTGACAGTATTGAC^C
```

Whose sequence reads are these?

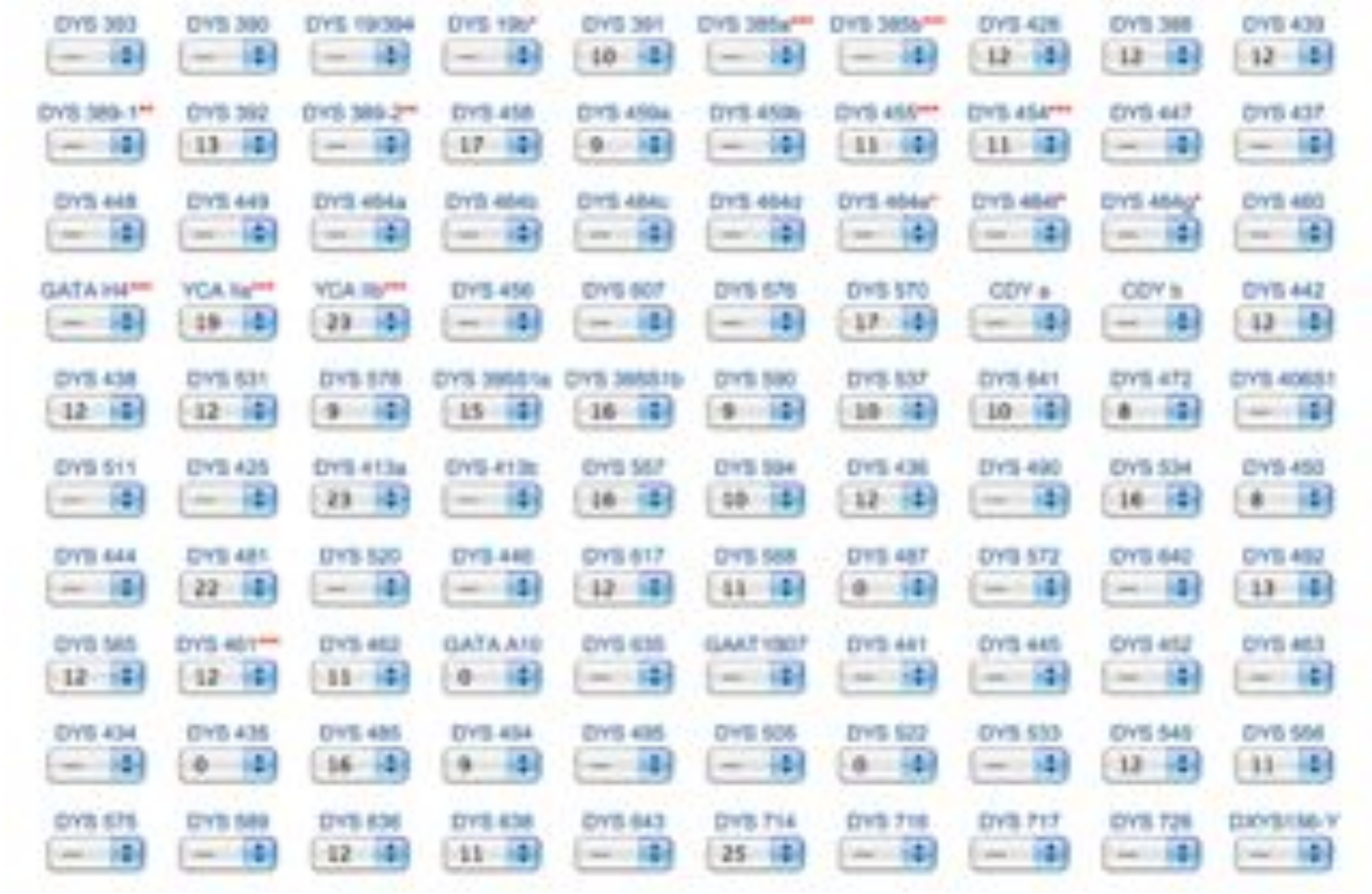
Identifying Personal Genomes by Surname Inference
Gymrek et al (2013) *Science*. doi: 10.1126/science.1229566

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.

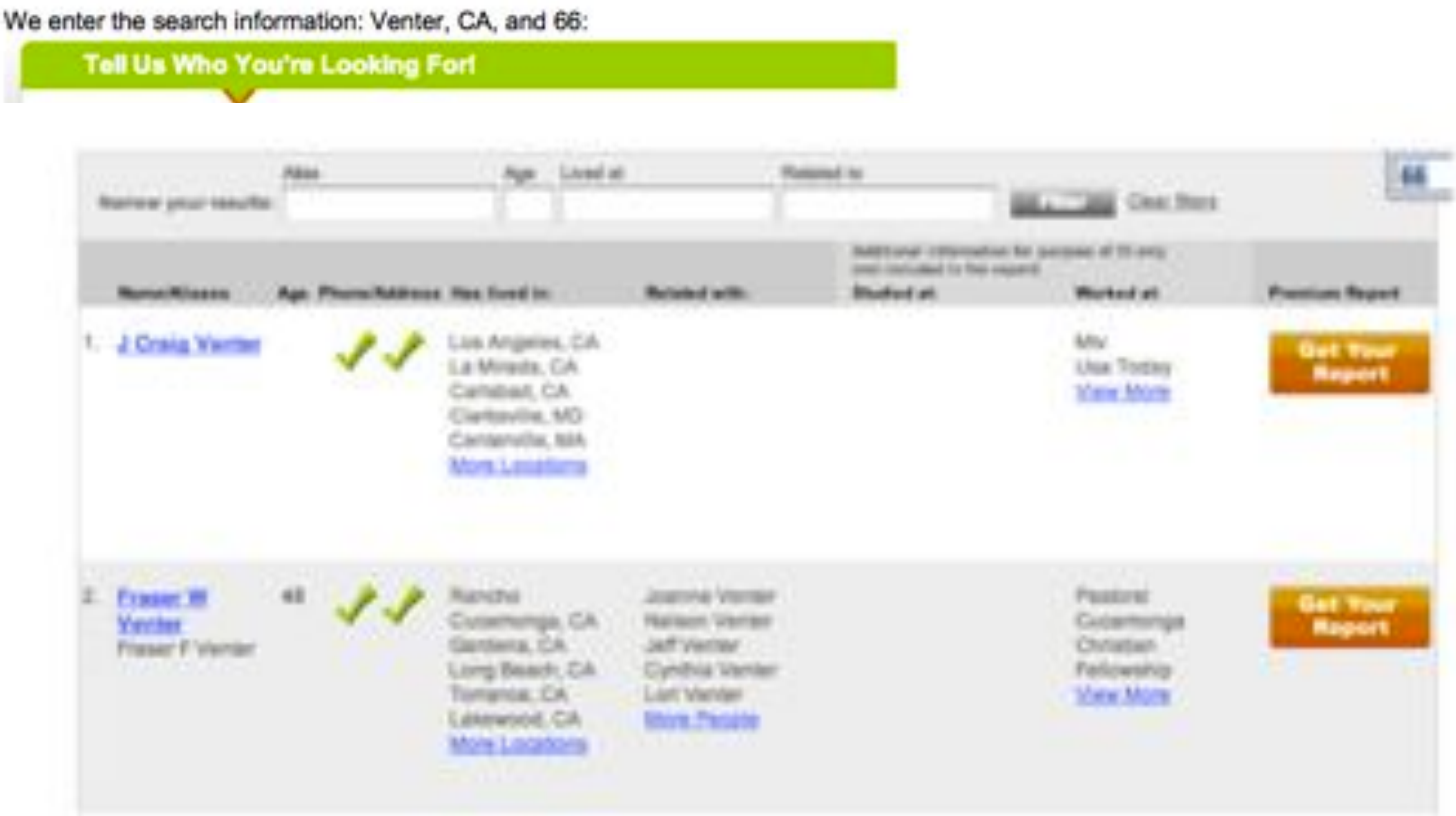
Step 2. Search for a surname hit in online genetic genealogy databases.



DYS 393	DYS 390	DYS 19/394	DYS 19b*	DYS 391	DYS 385a***	DYS 385b***	DYS 426	DYS 388	DYS 439
—	—	—	—	10	—	—	12	12	12
DYS 389-1**	DYS 392	DYS 389-2**	DYS 458	DYS 459a	DYS 459b	DYS 455***	DYS 454***	DYS 447	DYS 437
—	13	—	17	9	—	11	11	—	—
DYS 448	DYS 449	DYS 464a	DYS 464b	DYS 464c	DYS 464d	DYS 464e*	DYS 464f*	DYS 464g*	DYS 460
—	—	—	—	—	—	—	—	—	—
GATA H4***	YCA IIa***	YCA IIb***	DYS 456	DYS 607	DYS 576	DYS 570	CDY a	CDY b	DYS 442
—	19	23	—	—	—	17	—	—	12
DYS 438	DYS 531	DYS 578	DYS 395S1a	DYS 395S1b	DYS 590	DYS 537	DYS 641	DYS 472	DYS 406S1
12	12	9	15	16	9	10	10	8	—
DYS 511	DYS 425	DYS 413a	DYS 413b	DYS 557	DYS 594	DYS 436	DYS 490	DYS 534	DYS 450
—	—	23	—	16	10	12	—	16	8
DYS 444	DYS 481	DYS 520	DYS 446	DYS 617	DYS 568	DYS 487	DYS 572	DYS 640	DYS 492
—	22	—	—	12	11	0	—	—	13
DYS 565	DYS 461***	DYS 462	GATA A10	DYS 635	GAAT1B07	DYS 441	DYS 445	DYS 452	DYS 463
12	12	11	0	—	—	—	—	—	—
DYS 434	DYS 435	DYS 485	DYS 494	DYS 495	DYS 505	DYS 522	DYS 533	DYS 549	DYS 556
—	0	16	9	—	—	0	—	12	11
DYS 575	DYS 589	DYS 636	DYS 638	DYS 643	DYS 714	DYS 716	DYS 717	DYS 726	DXYS156-Y
—	—	12	11	—	25	—	—	—	—

http://www.ysearch.org

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



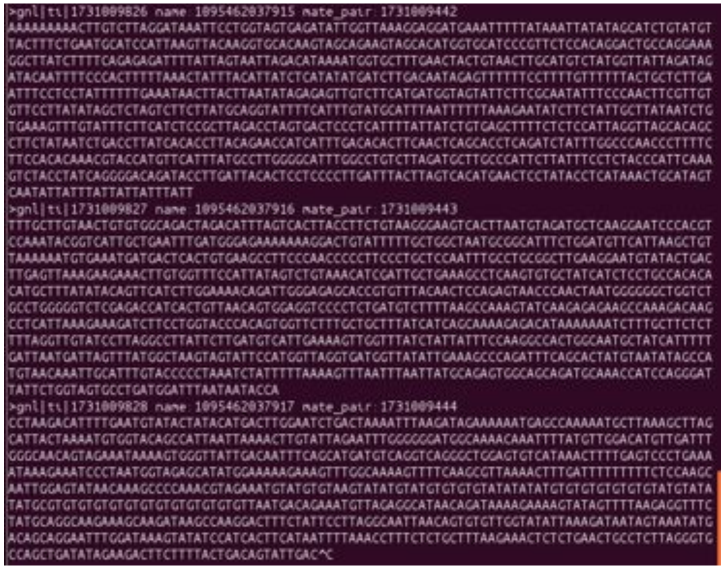
We enter the search information: Venter, CA, and 66:

Narrow your results: Alias ____ Age ____ Lived at ____ Related to ____ Filter Clear filters

					Additional information for purpose of ID only; not included in the report		
Name/Aliases	Age	Phone/Address	Has lived in:	Related with:	Studied at	Worked at	Premium Report
1. J Craig Venter		✓ ✓	Los Angeles, CA La Mirada, CA Carlsbad, CA Clarksville, MD Centerville, MA More Locations			Mtv Usa Today View More	Get Your Report
2. Fraser W Venter Fraser F Venter	48	✓ ✓	Rancho Cucamonga, CA Gardena, CA Long Beach, CA Torrance, CA Lakewood, CA More Locations	Joanna Venter Nelson Venter Jeff Venter Cynthia Venter Lori Venter More People		Pastoral Cucamonga Christian Fellowship View More	Get Your Report

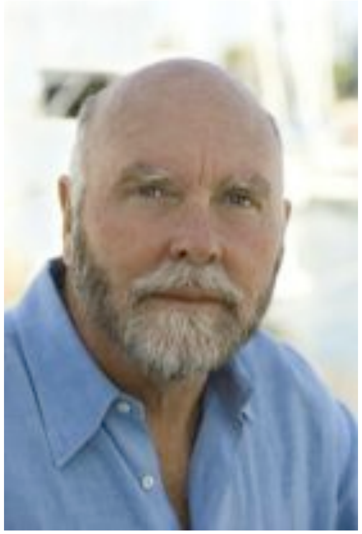
http://www.ussearch.com

Surname Inference



```
>gnl|ti|1731009826 name: 1095462037915 mate_pair: 1731009442
AAAAAAAAACTTGTCTTAGGATAAATTCCTGGTAGTGAGATATTGGTTAAAGGAGGATGAAATTTTTATAAATTATATAGCATCTGTATGT
TACTTTCTGAATGCATCCATTAAGTTACAAGGTGCACAAGTAGCAGAAGTAGCACATGGTGCATCCCGTTCTCCACAGGACTGCCAGGAAA
GGCTTATCTTTTCAGAGAGATTTTATTAGTAATTAGACATAAAATGGTGCTTTGAACTACTGTAACTTGCATGTCTATGGTTATTAGATAG
ATACAATTTTCCCACTTTTTAAACTATTTACATTATCTCATATATGATCTTGACAATAGAGTTTTTTCCTTTTGTTTTTTACTGCTCTTGA
ATTTCCTCCTATTTTTTGAAATAACTTACTTAATATAGAGAGTTGTCTTCATGATGGTAGTATTCTTCGCAATATTTCCCAACTTCGTTGT
GTTCCTTATATAGCTCTAGTCTTCTTATGCAGGTATTTTCATTTGTATGCATTTAATTTTTTAAAGAATATCTTCTATTGCTTATAATCTG
TGAAAGTTTGTATTTCTTCATCTCCGCTTAGACCTAGTGACTCCCTCATTTTATTATCTGTGAGCTTTTCTCTCCATTAGGTTAGCACAGC
CTTCTATAATCTGACCTTATCACACCTTACAGAACCATCATTTGACACACTTCAACTCAGCACCTCAGATCTATTTGGCCCAACCCTTTTC
TTCCACACAAACGTACCATGTTCATTTATGCCTTGGGGCATTTGGCCTGTCTTAGATGCTTGCCCATTCTTATTTCCTCTACCCATTCAAA
GTCTACCTATCAGGGGACAGATACCTTGATTACACTCCTCCCCTTGATTTACTTAGTCACATGAACTCCTATACCTCATAAACTGCATAGT
CAATATTATTTATTATTATTTATT
>gnl|ti|1731009827 name: 1095462037916 mate_pair: 1731009443
TTTGCTTGTAACTGTGTGGCAGACTAGACATTTAGTCACTTACCTTCTGTAAGGGAAGTCACTTAATGTAGATGCTCAAGGAATCCCACGT
CCAAATACGGTCATTGCTGAATTTGATGGGAGAAAAAAAGGACTGTATTTTTGCTGGCTAATGCGGCATTTCTGGATGTTCATTAAGCTGT
TAAAAAATGTGAAATGATGACTCACTGTGAAGCCTTCCCAACCCCCTTCCCTGCTCCAATTTGCCTGCGGCTTGAAGGAATGTATACTGAC
TTGAGTTAAAGAAGAAACTTGTGGTTTCCATTATAGTCTGTAAACATCGATTGCTGAAAGCCTCAAGTGTGCTATCATCTCCTGCCACACA
CATGCTTTATATACAGTTCATCTTGGAAAACAGATTGGGAGAGCACCGTGTTTACAACTCCAGAGTAACCCAACTAATGGGGGGCTGGTCT
GCCTGGGGGTCTCGAGACCATCACTGTTAACAGTGGAGGTCCCCTCTGATGTCTTTTAAGCCAAAGTATCAAGAGAGAAGCCAAAGACAAG
CCTCATTAAAGAAAGATCTTCCTGGTACCCACAGTGGTTCTTTGCTGCTTTATCATCAGCAAAAGAGACATAAAAAAATCTTTGCTTCTCT
TTTAGGTTGTATCCTTAGGCCTTATTCTTGATGTCATTGAAAAGTTGGTTTATCTATTATTTCCAAGGCCACTGGCAATGCTATCATTTTT
GATTAATGATTAGTTTATGGCTAAGTAGTATTCCATGGTTAGGTGATGGTTATATTGAAAGCCCAGATTTCAGCACTATGTAATATAGCCA
TGTAACAAATTGCATTTGTACCCCCTAAATCTATTTTTAAAAGTTTAATTTAATTATGCAGAGTGGCAGCAGATGCAAACCATCCAGGGAT
TATTCTGGTAGTGCCTGATGGATTTAATAATACCA
>gnl|ti|1731009828 name: 1095462037917 mate_pair: 1731009444
CCTAAGACATTTTGAATGTATACTATACATGACTTGGAATCTGACTAAAATTTAAGATAGAAAAAATGAGCCAAAAATGCTTAAAGCTTAG
CATTACTAAAATGTGGTACAGCCATTAATTAAAACTTGTATTAGAATTTGGGGGGATGGCAAAACAAATTTTATGTTGGACATGTTGATTT
GGGCAACAGTAGAAATAAAAGTGGGTTATTGACAATTTCAGCATGATGTCAGGTCAGGGCTGGAGTGTCATAAACTTTTGAGTCCCTGAAA
ATAAAGAAATCCCTAATGGTAGAGCATATGGAAAAAGAAAGTTTGGCAAAAGTTTTCAAGCGTTAAAACTTTGATTTTTTTTTCTCCAAGC
AATTGGAGTATAACAAAGCCCCAAACGTAGAAATGTATGTGTAAGTATATGTATGTGTGTGTATATATATGTGTGTGTGTGTATGTATA
TATGCGTGTGTGTGTGTGTGTGTGTGTGTGTGTTAATGACAGAAATGTTAGAGGCATAACAGATAAAAGAAAAGTATAGTTTTAAGAGGTTTC
TATGCAGGCAAGAAAGCAAGATAAGCCAAGGACTTTCTATTCCTTAGGCAATTAACAGTGTGTTGGTATATTAAAGATAATAGTAAATATG
ACAGCAGGAATTTGGATAAAGTATATCCATCACTTCATAATTTTAAACCTTTCTCTGCTTTAAGAAACTCTCTGAACTGCCTCTTAGGGTG
CCAGCTGATATAGAAGACTTCTTTTACTGACAGTATTGAC^C
```

It's Craig Venter!

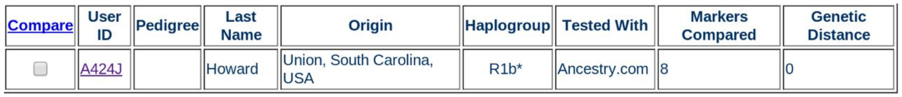


Identifying Personal Genomes by Surname Inference
Gymrek et al (2013) *Science*. doi: 10.1126/science.1229566

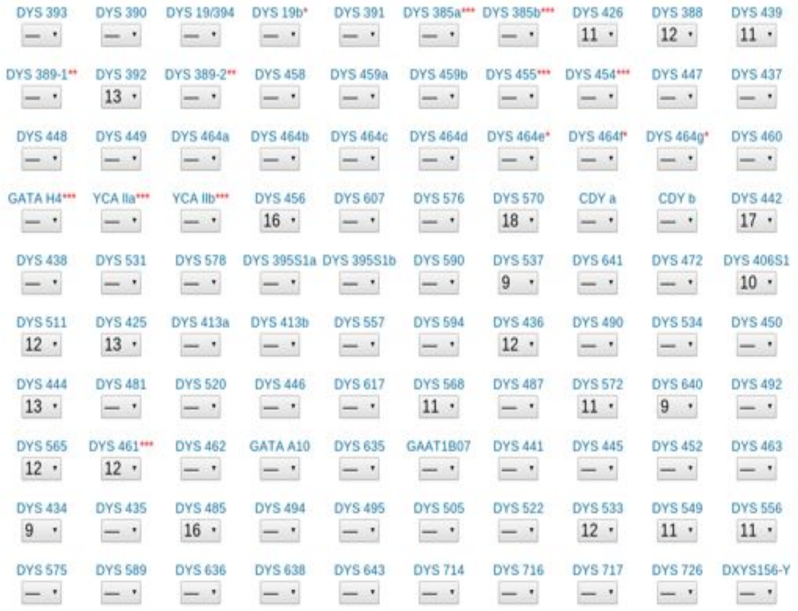
Can we identify Jim Watson?

- 187 fasta reads acquired from ftp://ftp.ncbi.nih.gov/pub/TraceDB/Personal_Genomics/Watson/
- 741,131,864 reads mapped.
- 24 markers identified.

- ySearch returns inconclusive search result:



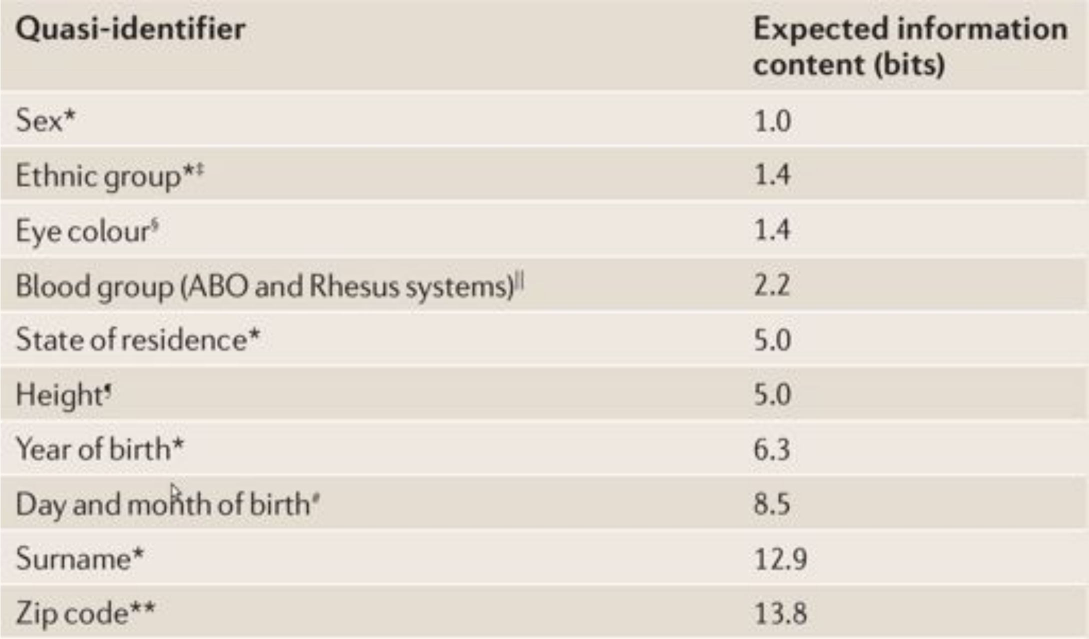
Compare	User ID	Pedigree	Last Name	Origin	Haplogroup	Tested With	Markers Compared	Genetic Distance
□	A424J		Howard	Union, South Carolina, USA	R1b*	Ancestry.com	8	0



DYS 393	DYS 390	DYS 19/394	DYS 19b*	DYS 391	DYS 385a***	DYS 385b***	DYS 426	DYS 388	DYS 439
—	—	—	—	—	—	—	11	12	11
DYS 389-1**	DYS 392	DYS 389-2**	DYS 458	DYS 459a	DYS 459b	DYS 455***	DYS 454***	DYS 447	DYS 437
—	13	—	—	—	—	—	—	—	—
DYS 448	DYS 449	DYS 464a	DYS 464b	DYS 464c	DYS 464d	DYS 464e*	DYS 464f*	DYS 464g*	DYS 460
—	—	—	—	—	—	—	—	—	—
GATA H4***	YCA IIa***	YCA IIb***	DYS 456	DYS 607	DYS 576	DYS 570	CDY a	CDY b	DYS 442
—	—	—	16	—	—	18	—	—	17
DYS 438	DYS 531	DYS 578	DYS 395S1a	DYS 395S1b	DYS 590	DYS 537	DYS 641	DYS 472	DYS 406S1
—	—	—	—	—	—	9	—	—	10
DYS 511	DYS 425	DYS 413a	DYS 413b	DYS 557	DYS 594	DYS 436	DYS 490	DYS 534	DYS 450
12	13	—	—	—	—	12	—	—	—
DYS 444	DYS 481	DYS 520	DYS 446	DYS 617	DYS 568	DYS 487	DYS 572	DYS 640	DYS 492
13	—	—	—	—	11	—	11	9	—
DYS 565	DYS 461***	DYS 462	GATA A10	DYS 635	GAAT1B07	DYS 441	DYS 445	DYS 452	DYS 463
12	12	—	—	—	—	—	—	—	—
DYS 434	DYS 435	DYS 485	DYS 494	DYS 495	DYS 505	DYS 522	DYS 533	DYS 549	DYS 556
9	—	16	—	—	—	—	12	11	11
DYS 575	DYS 589	DYS 636	DYS 638	DYS 643	DYS 714	DYS 716	DYS 717	DYS 726	DXYS156-Y
—	—	—	—	—	—	—	—	—	—

- Possible errors?
    - Insufficient family data for Watson’s relatives online
    - Unreliable sequence reads
    - Potential LobSTR mistake, mis-alignment error or not enough input data

Identifiers and Quasi-identifiers

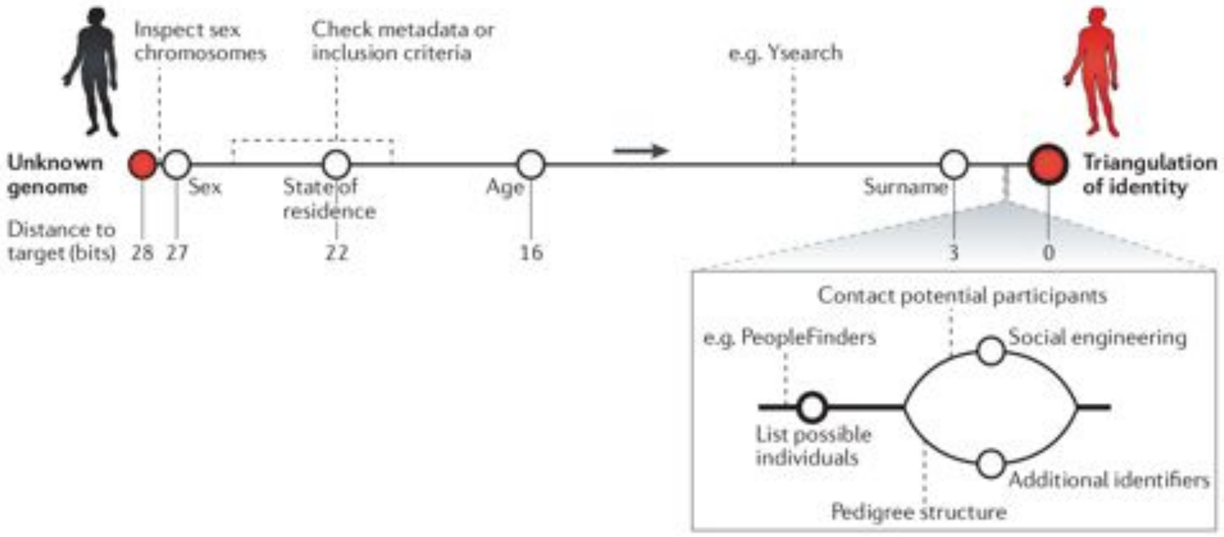


Quasi-identifier	Expected information content (bits)
Sex*	1.0
Ethnic group*‡	1.4
Eye colour§	1.4
Blood group (ABO and Rhesus systems)‖	2.2
State of residence*	5.0
Height¶	5.0
Year of birth*	6.3
Day and month of birth#	8.5
Surname*	12.9
Zip code**	13.8

- What are Quasi-Identifiers?
  - Pieces of information that are not unique by themselves, but when combined with other quasi-identifiers, may create a unique identifier.
- What is Entropy?
  - Entropy measures the degree of uncertainty in the outcome of a random variable, where 1 bit equates to the chances of tossing a single fair coin.
  - Complete identification is guaranteed when expected information bits reaches 0.

Routes for breaching and protecting genetic privacy
Erlich and Narayanan (2014) *Nature Reviews Genetics*. doi: 10.1038/nrg3723

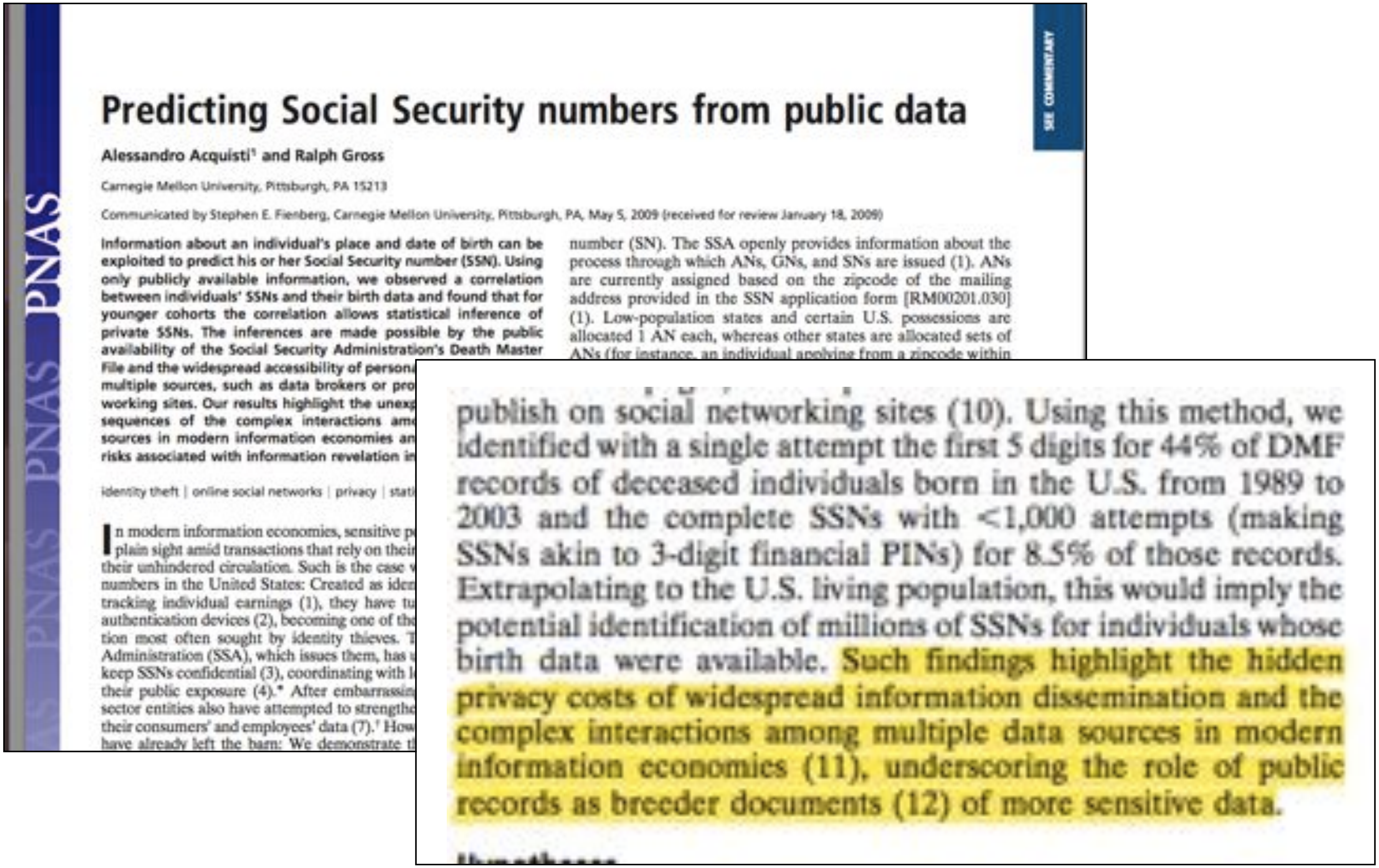
Possible route for identity tracing



- *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 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 amo
sources in modern information economies an
risks associated with information revelation in**

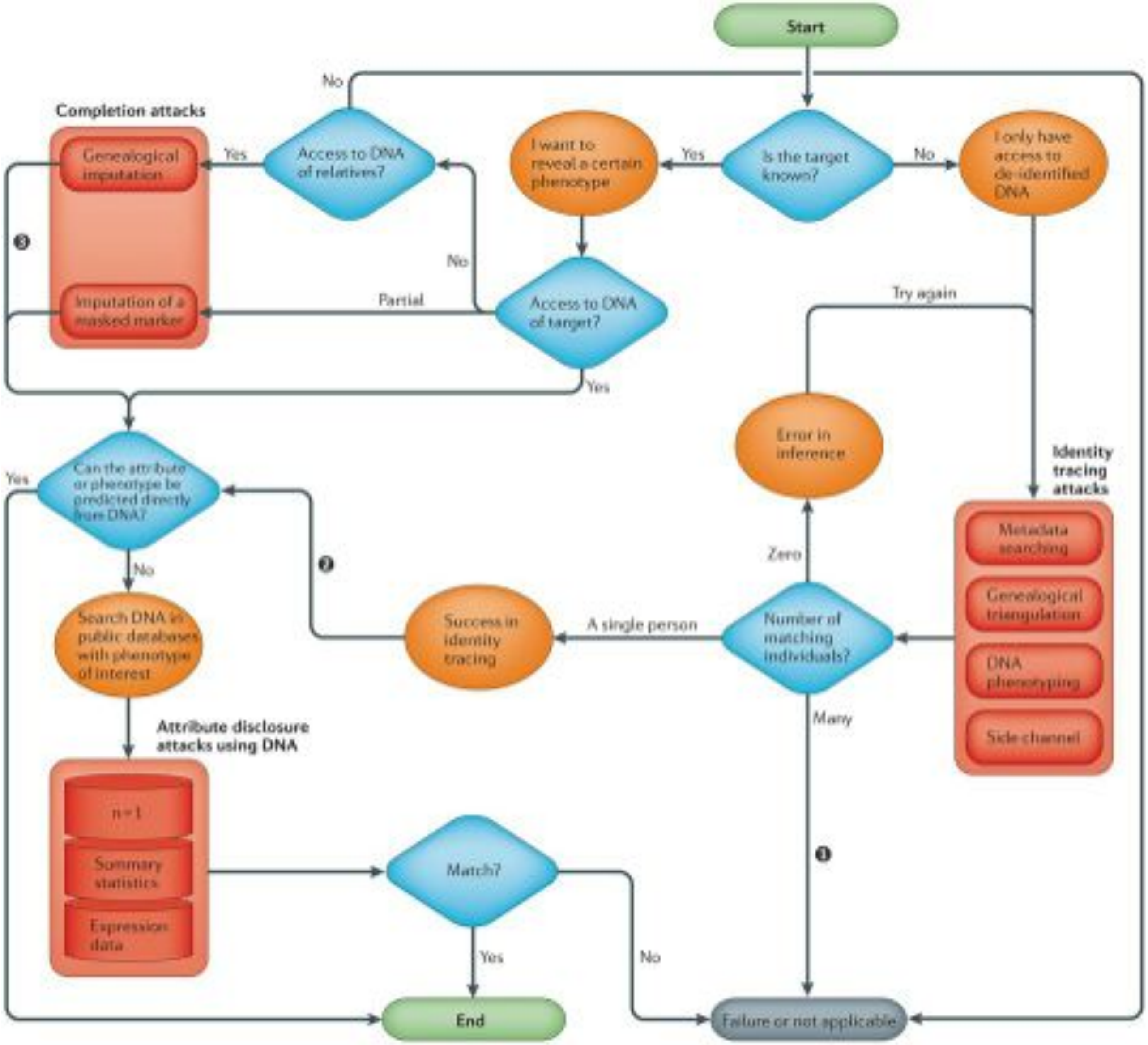
identity theft | online social networks | privacy | stati

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

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.

Broader Privacy Implications


Start
No
Completion attacks
Genealogical imputation
Yes
Access to DNA of relatives?
I want to reveal a certain phenotype
Yes
Is the target known?
No
I only have access to de-identified DNA
3
No
Try again
Partial
Imputation of a masked marker
Access to DNA of target?
Yes
Error in inference
Identity tracing attacks
Yes
Can the attribute or phenotype be predicted directly from DNA?
Metadata searching
2
Zero
Genealogical triangulation
No
Search DNA in public databases with phenotype of interest
Success in identity tracing
A single person
Number of matching individuals?
DNA phenotyping
Side channel
Many
Attribute disclosure attacks using DNA
n=1
Summary statistics
1
Match?
Expression data
Yes
No
End
Failure or not applicable

Nature Reviews | Genetics

Next class

- Gene Finding and HMMs
- Review!
- Homework due Monday