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

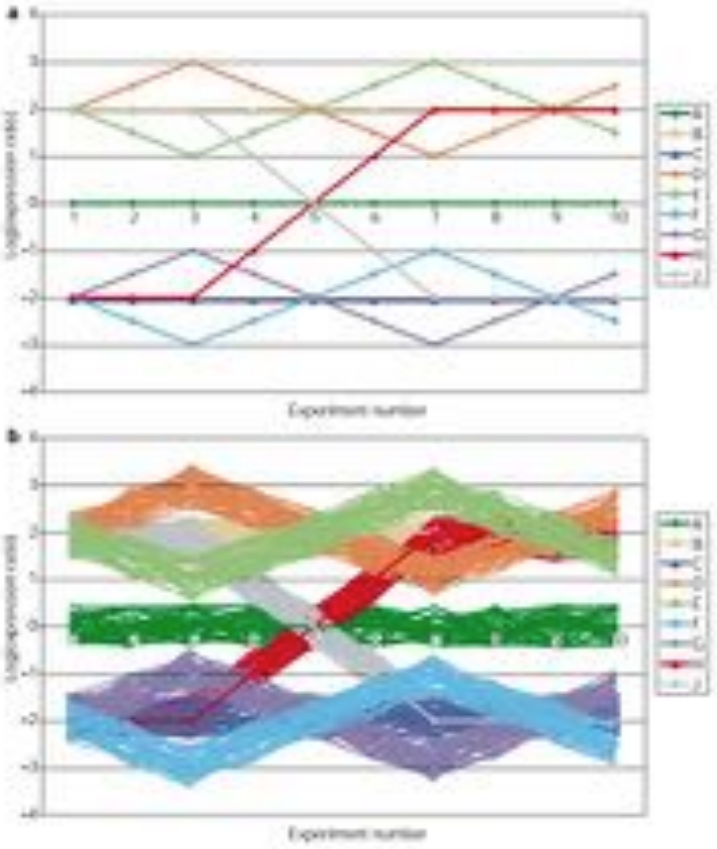
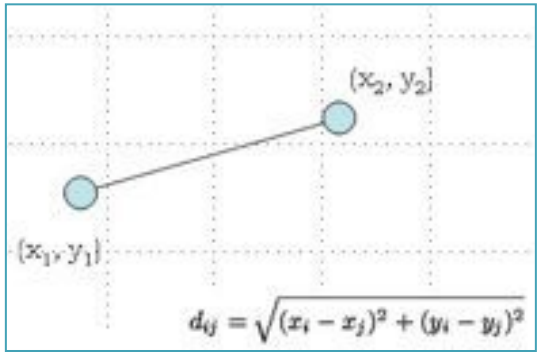
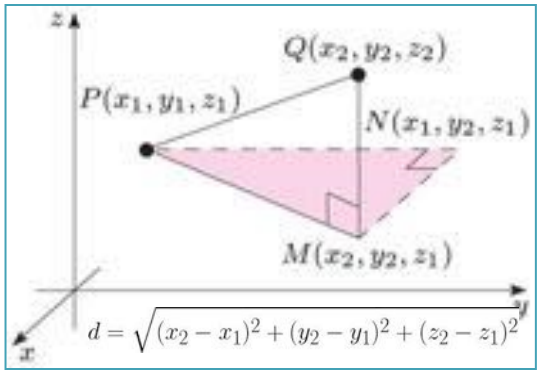


Figure 2 | **A synthetic gene-expression data set.** This data set provides an opportunity to evaluate how various clustering algorithms reveal different features of the data. **a** | Nine distinct gene-expression patterns were created with log (ratio) expression measures defined for ten experiments.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

Euclidean Distance

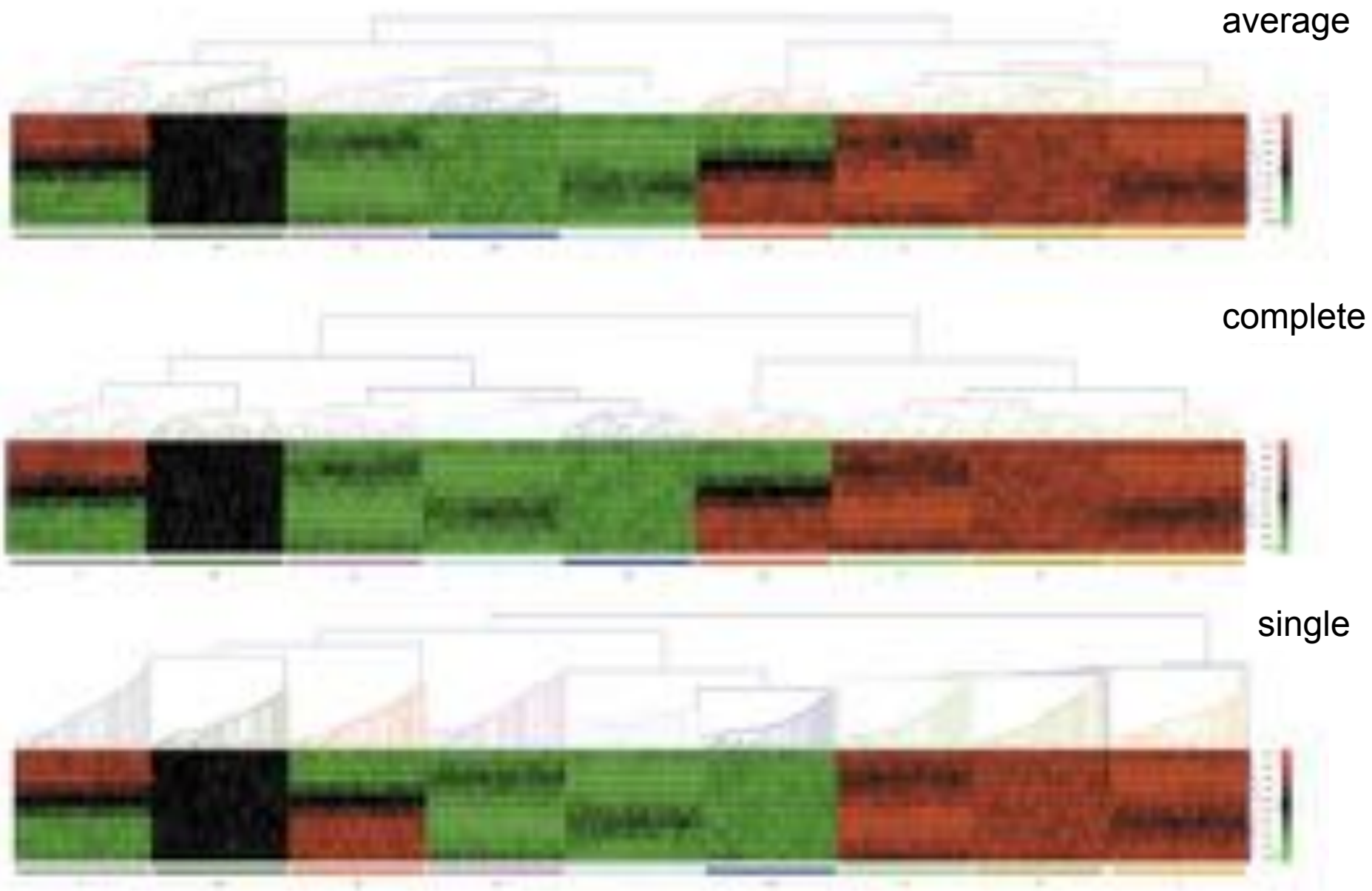




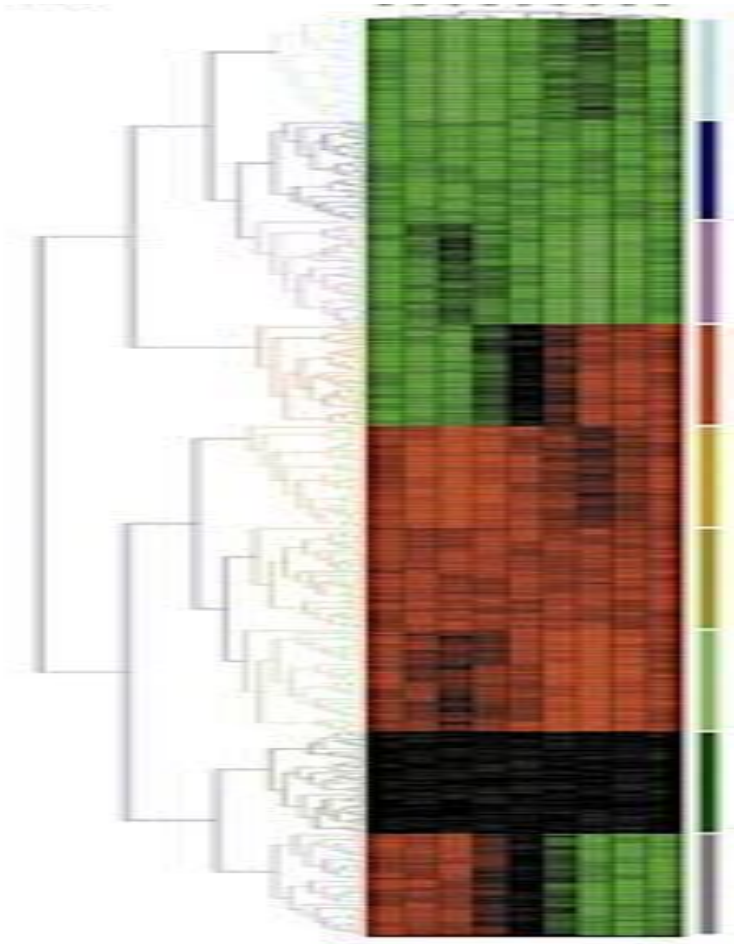
$$\mathrm{d}(\mathbf{p}, \mathbf{q}) = \mathrm{d}(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

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

Hierarchical Clustering



Principle Components Analysis (PCA)



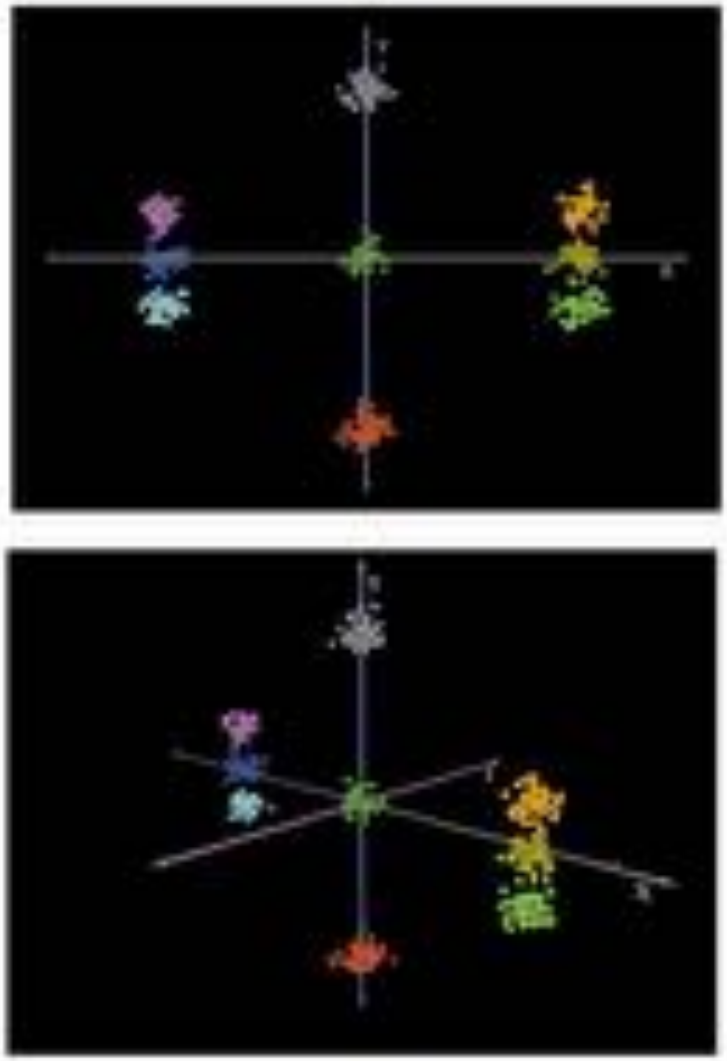


Figure 4 | **Principal component analysis.** The same demonstration data set was analysed using **a** | hierarchical (average-linkage) clustering and **b** | principal component analysis using Euclidean distance, to show how each treats the data, with genes colour coded on the basis of hierarchical clustering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

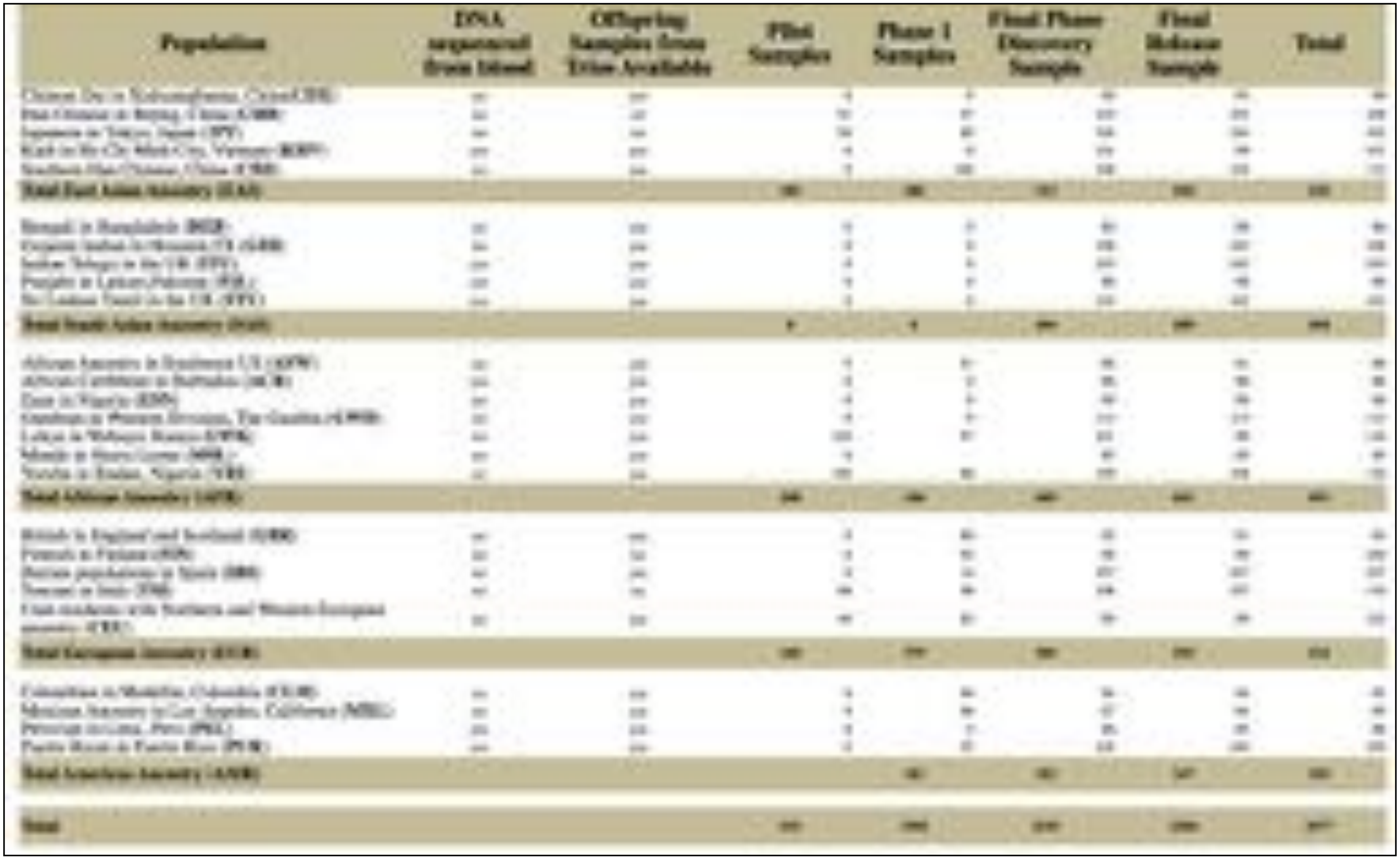
An integrated map of genetic variation from 1,092 human genomes

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 large 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 those disrupting clinically curated 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

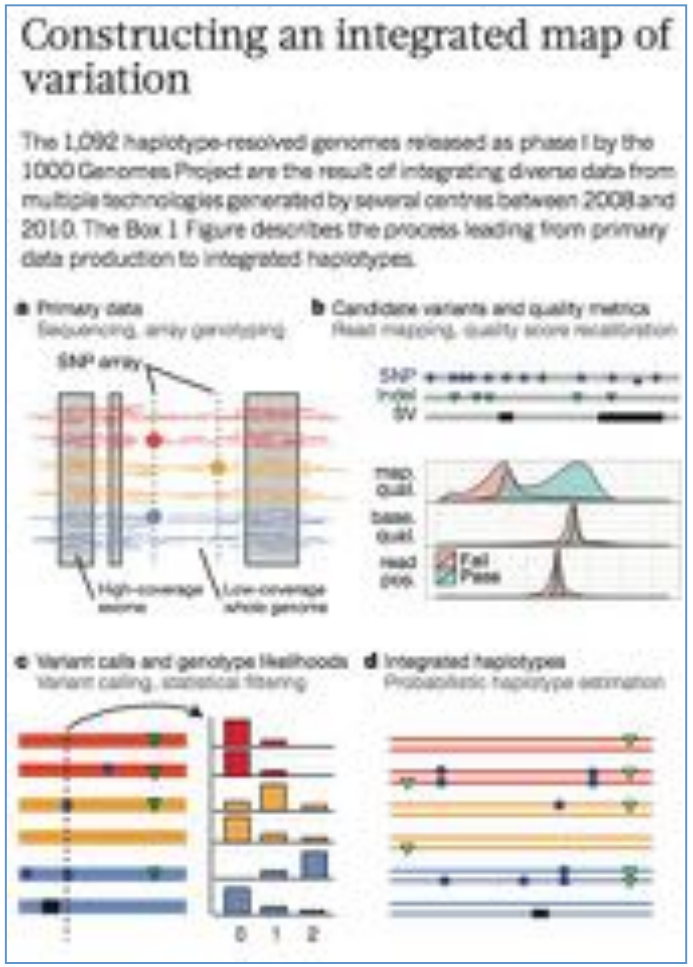
26 populations from 5 major population groups

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

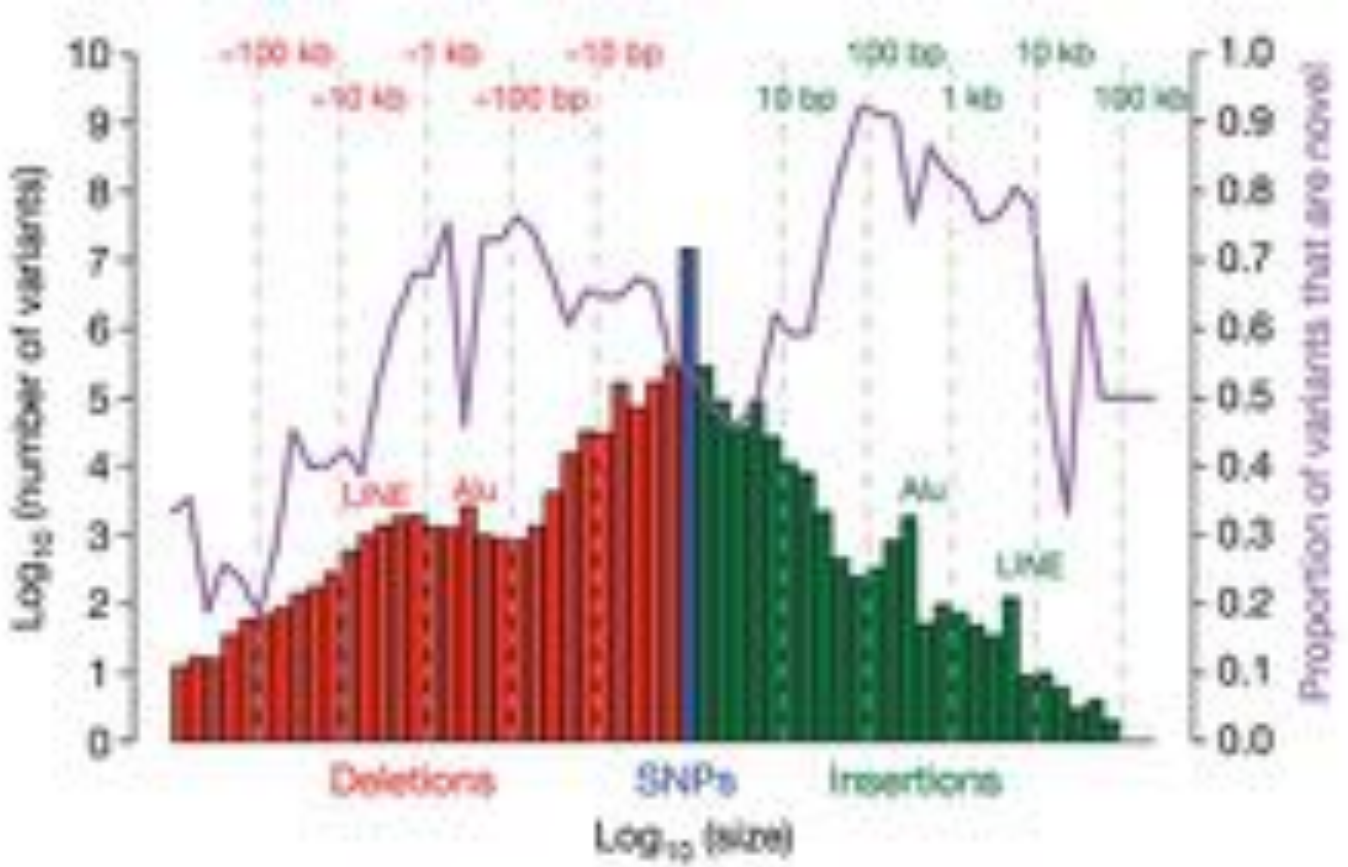
- Overall SNP rate between any two people is ~1/1200bp to ~1/1300
  - ~3M SNPs between me and you (.1%)
  - ~30M SNPs between human to Chimpanzees (1%)

- 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

Copy Number Variations

Large-Scale Copy Number Polymorphism in the Human Genome

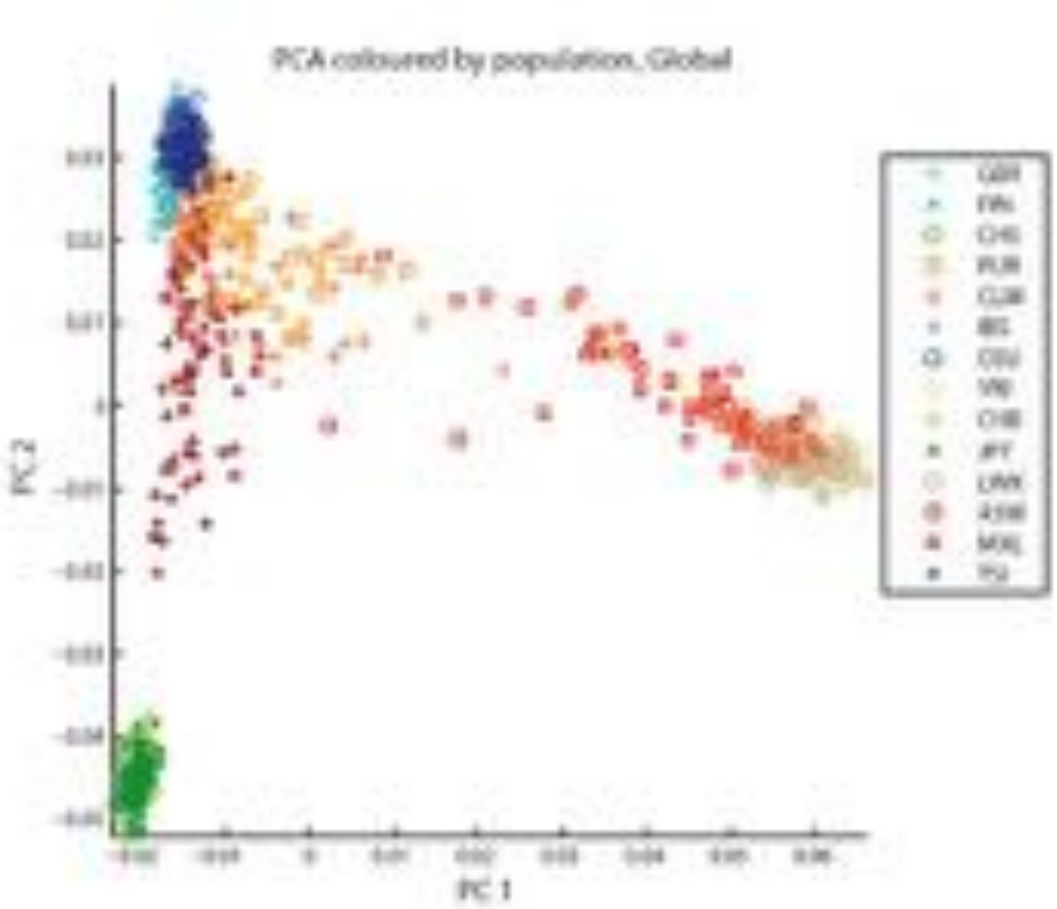
Strong Association of De Novo Copy Number Mutations with Autism

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

dbSNP

- 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

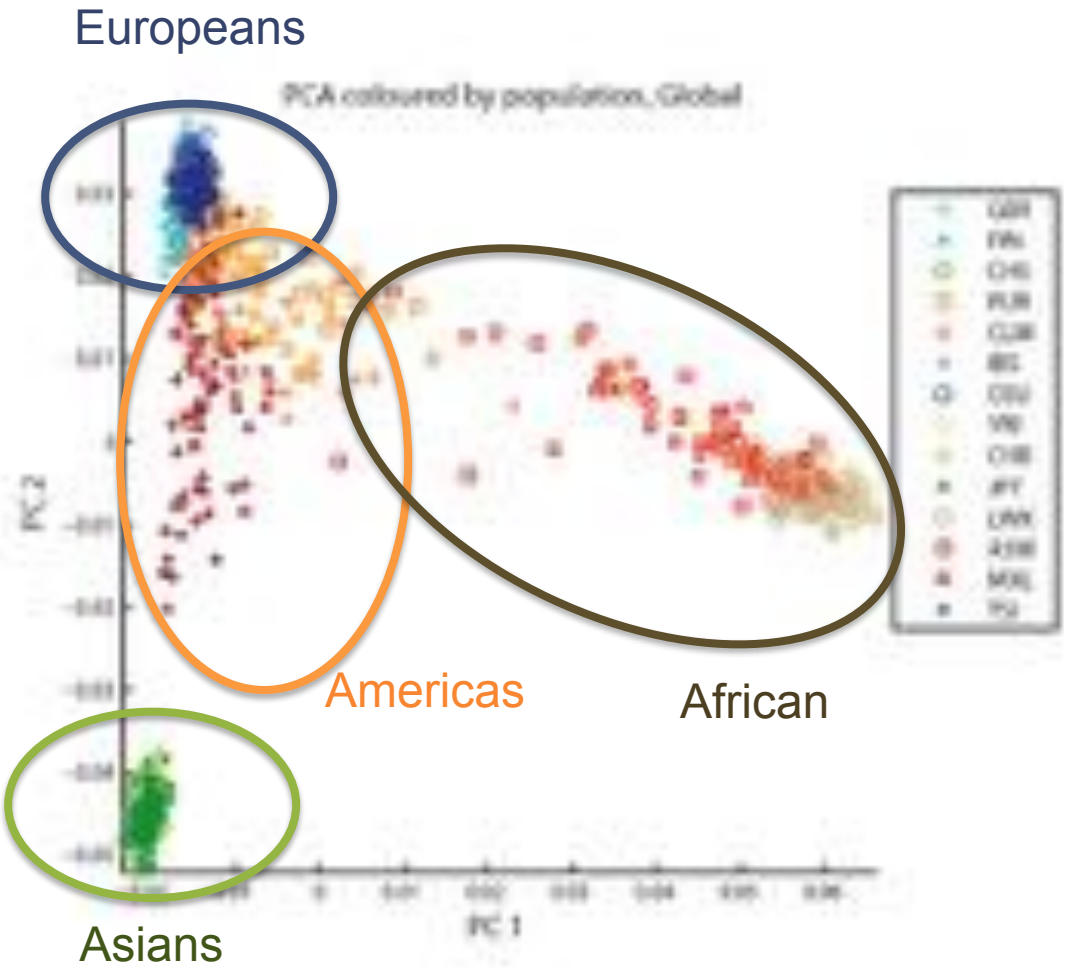


LEVEL	POP_PAIR	# of Highly differentiated SNPs	% in transcribed regions*
AFR	ASW-LWK	258	46.8
AFR	LWK-YRI	251	50.2
AFR	ASW-YRI	213	45.8
ASN	CHS-JPT	275	48.1
ASN	CHB-JPT	176	43.7
ASN	CHB-CHS	79	38.7
EUR	FIN-TSI	343	42.6
EUR	CEU-FIN	205	40.7
EUR	FIN-GBR	197	43.2
EUR	GBR-TSI	100	38.9
EUR	CEU-TSI	57	53.8
EUR	CEU-GBR	17	14.3
CON	AFR-EUR	348	52.2
CON	AFR-ASN	337	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

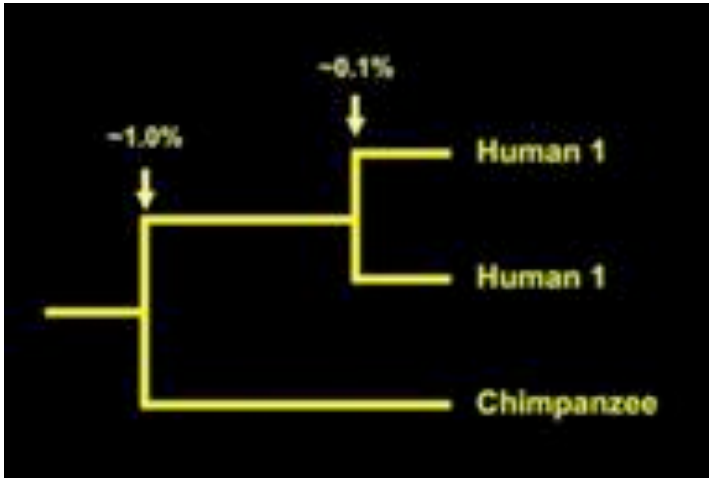


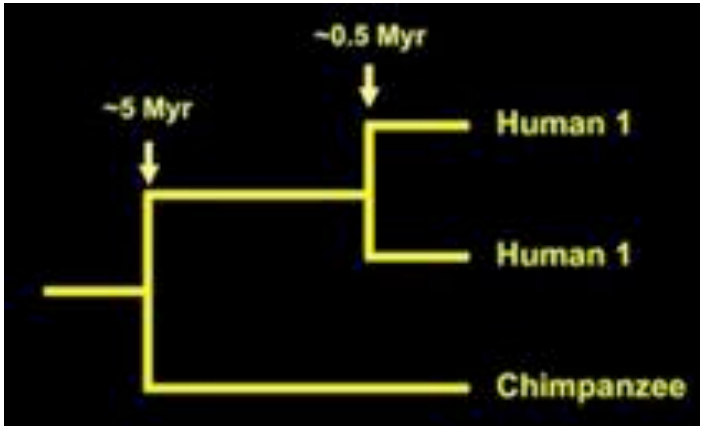
LEVEL	POP_PAIR	# of Highly differentiated SNPs	% in transcribed regions*
AFR	ASW-LWK	258	46.8
AFR	LWK-YRI	251	50.2
AFR	ASW-YRI	213	45.8
ASN	CHS-JPT	275	48.1
ASN	CHB-JPT	176	43.7
ASN	CHB-CHS	79	38.7
EUR	FIN-TSI	343	42.6
EUR	CEU-FIN	201	40.7
EUR	FIN-GBR	197	43.2
EUR	GBR-TSI	100	38.9
EUR	CEU-TSI	57	53.8
EUR	CEU-GBR	17	14.3
CON	AFR-EUR	348	52.2
CON	AFR-ASN	337	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 ☺

Science
AAAS

A Draft Sequence of the Neandertal Genome

Richard E. Green, *et al.*

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

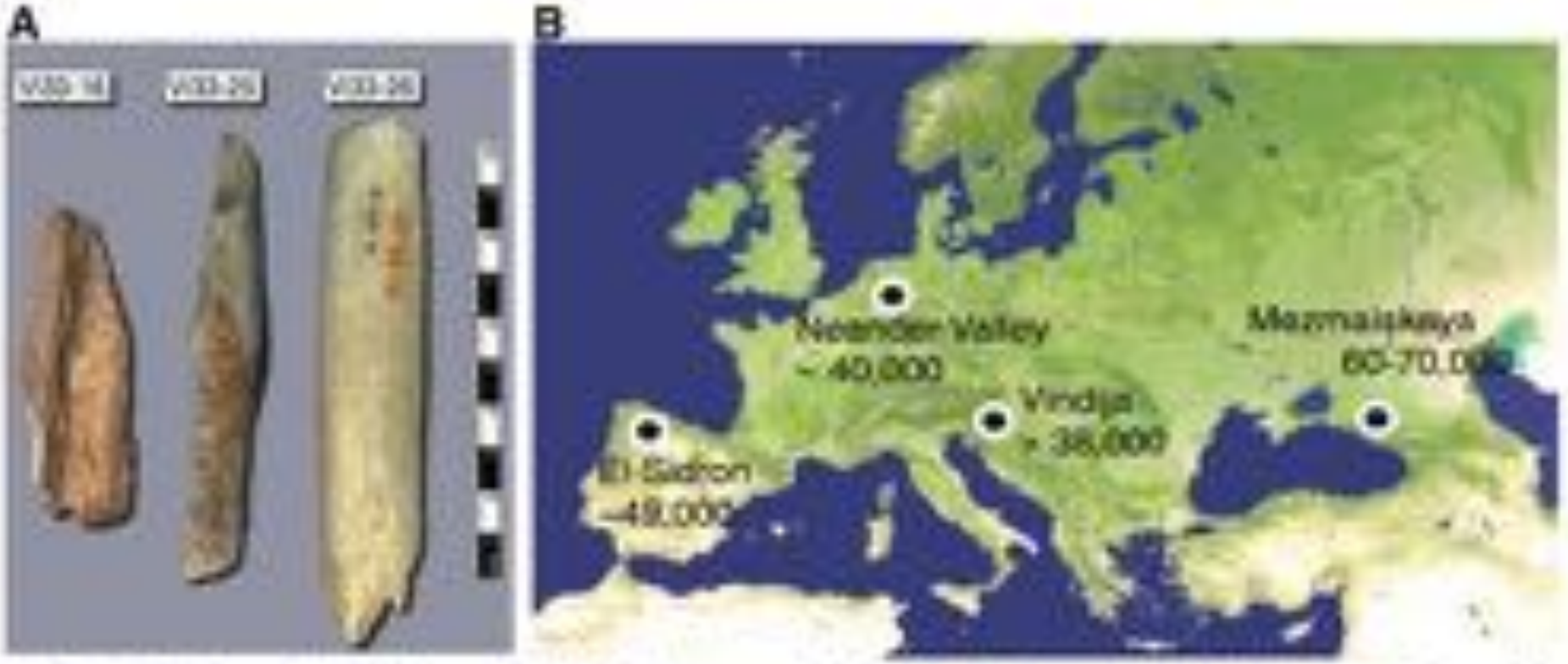
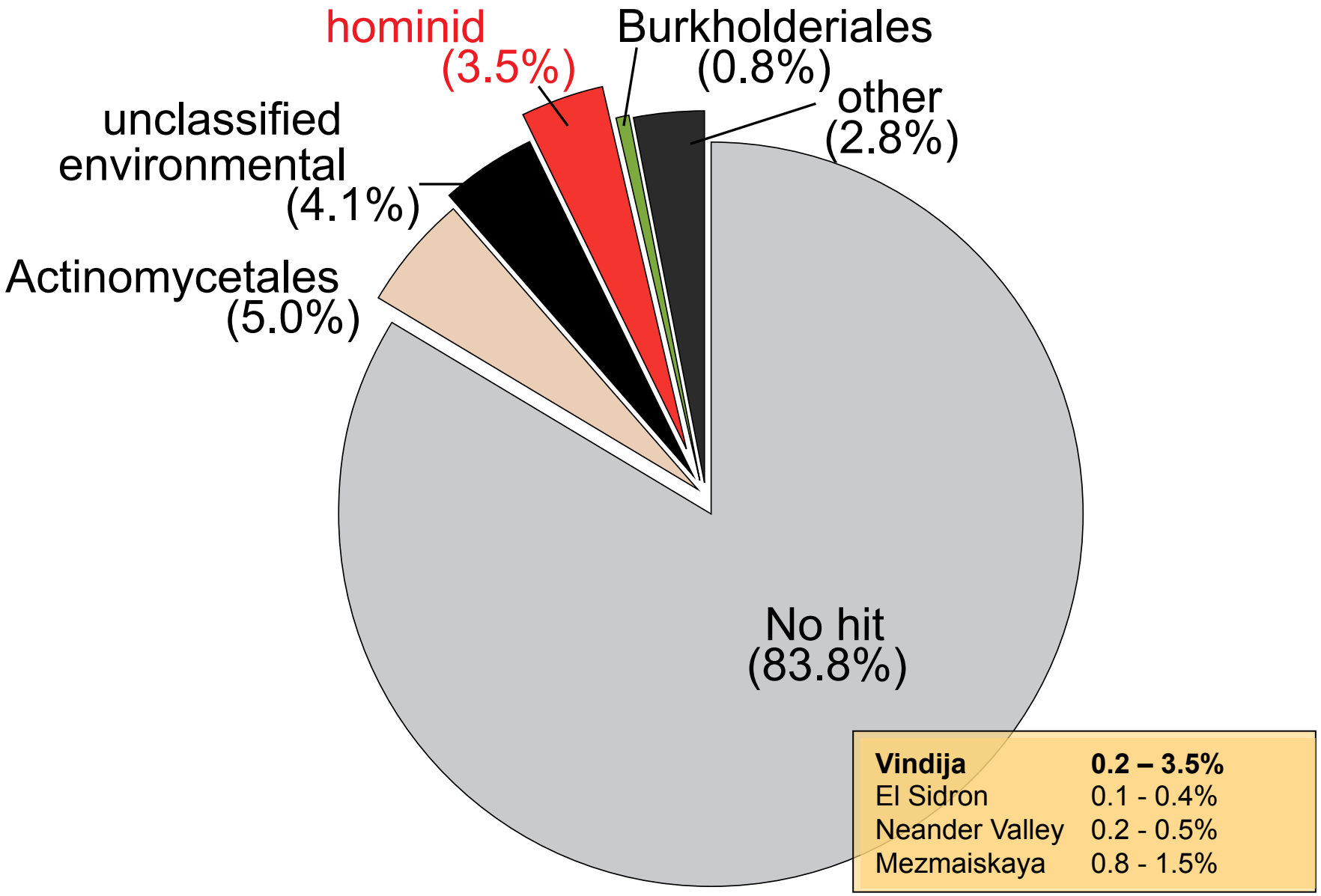


Fig. 1. Samples and sites from which DNA was retrieved. (**A**) The three bones from Vindija from which Neandertal DNA was sequenced. (**B**) Map showing the four archaeological sites from which bones were used and their approximate dates (years B.P.).

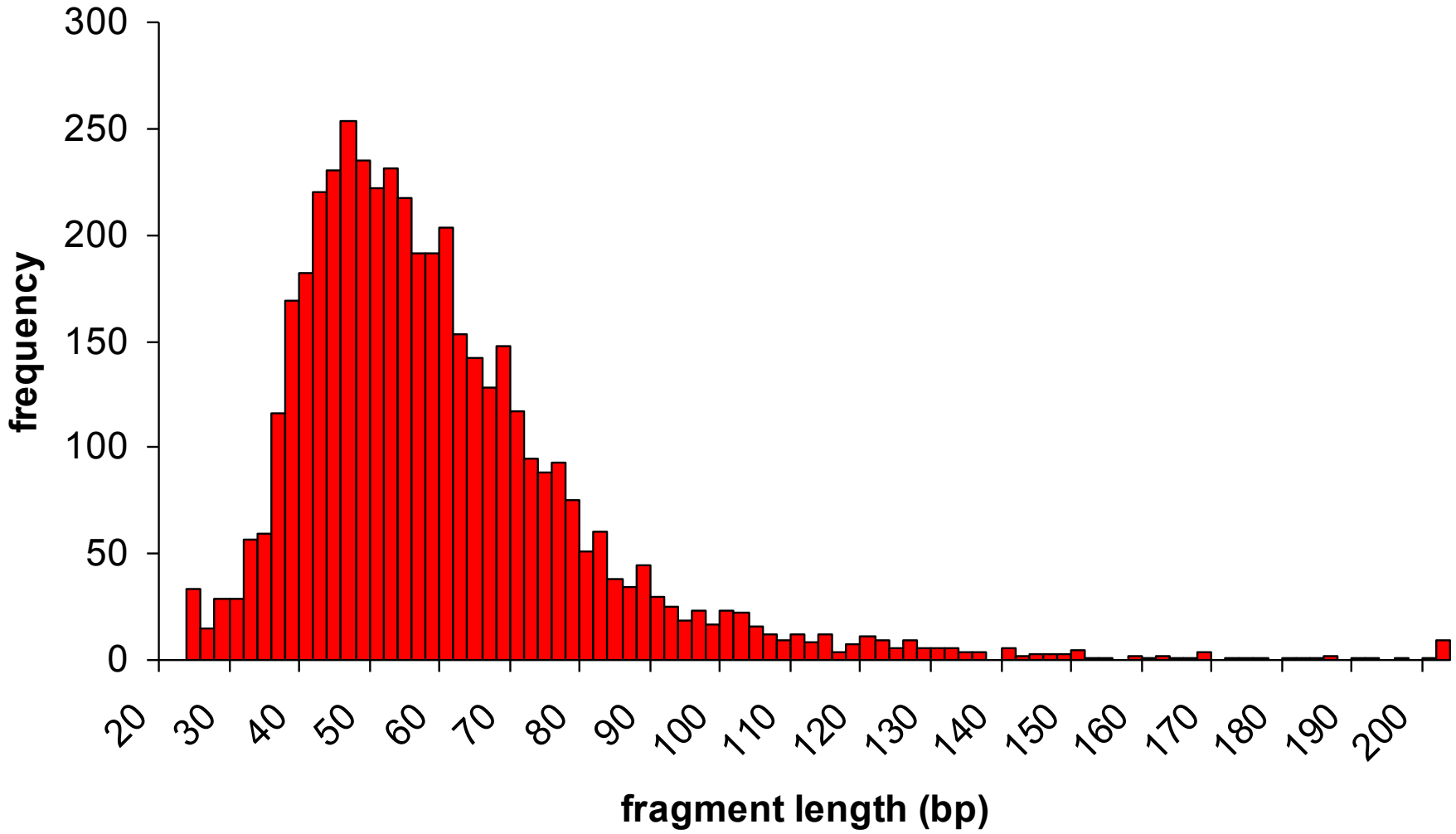
Extracting Ancient DNA

DNA is from mixed sources

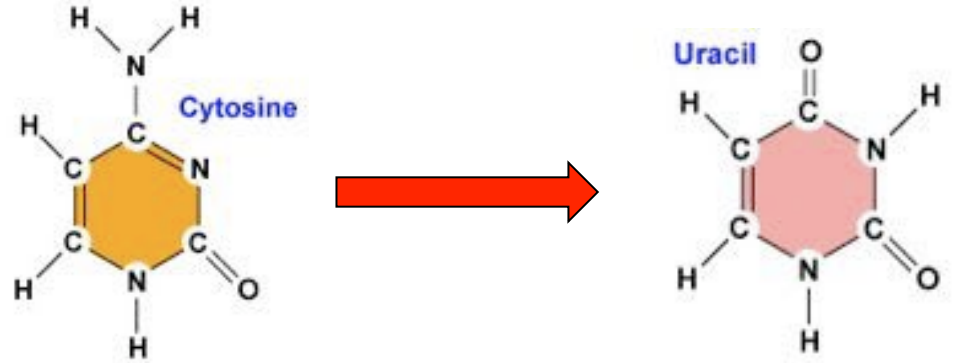


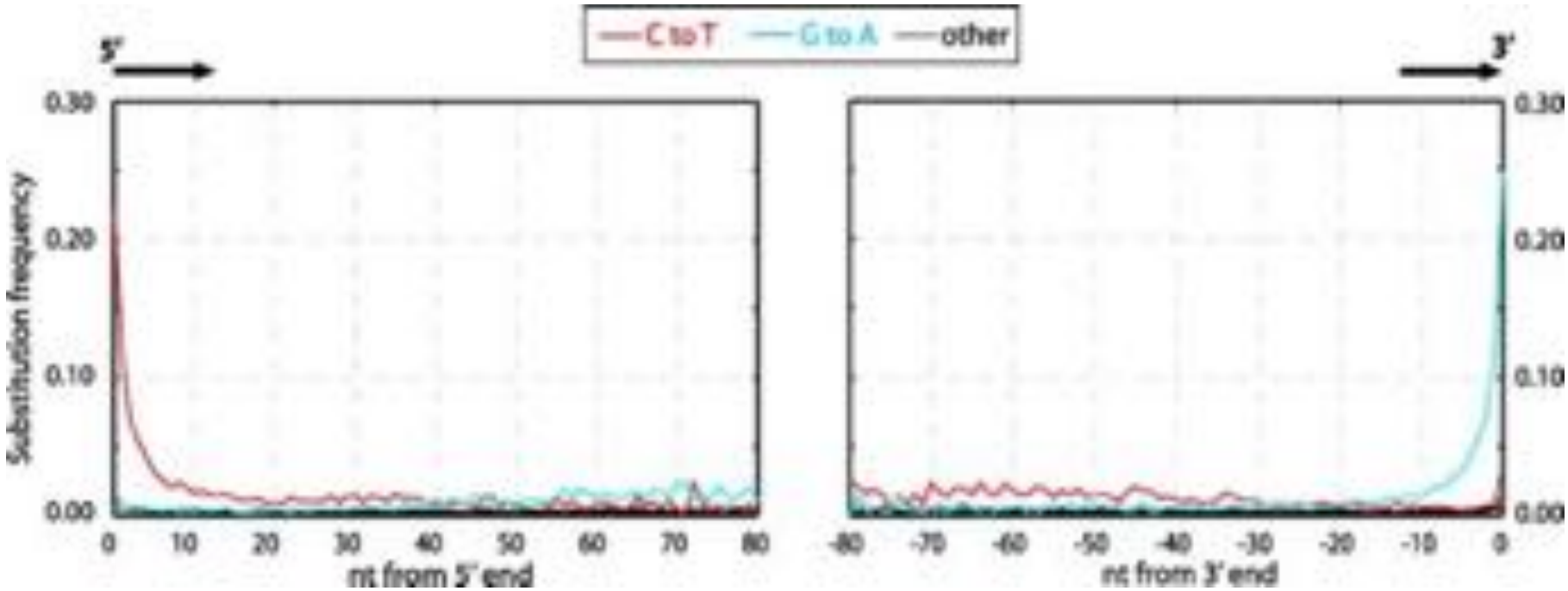
Vindija	**0.2 – 3.5%**
El Sidron	0.1 - 0.4%
Neander Valley	0.2 - 0.5%
Mezmaiskaya	0.8 - 1.5%

DNA is degraded



DNA is chemically damaged





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

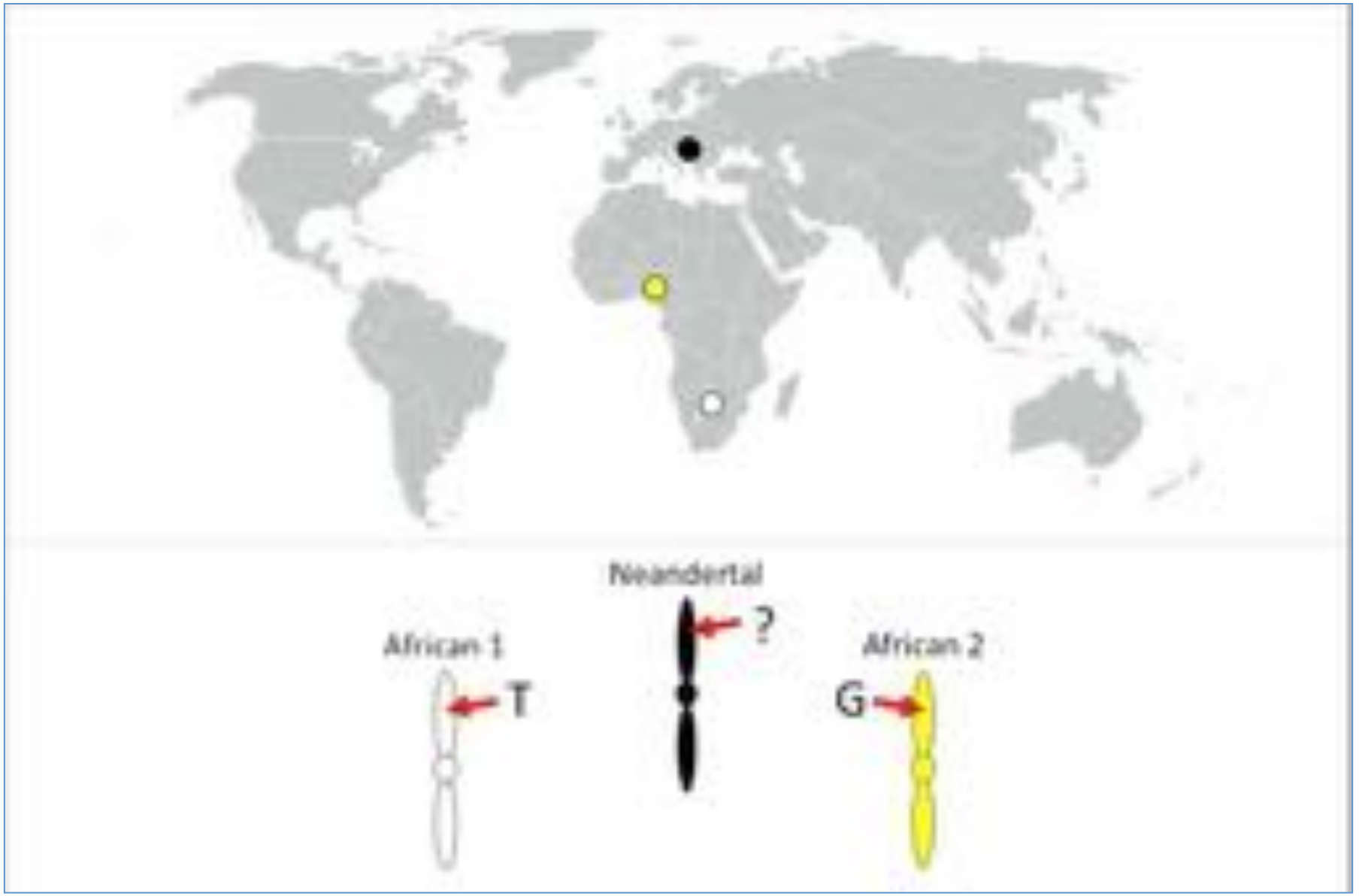
Green et al. 2010

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

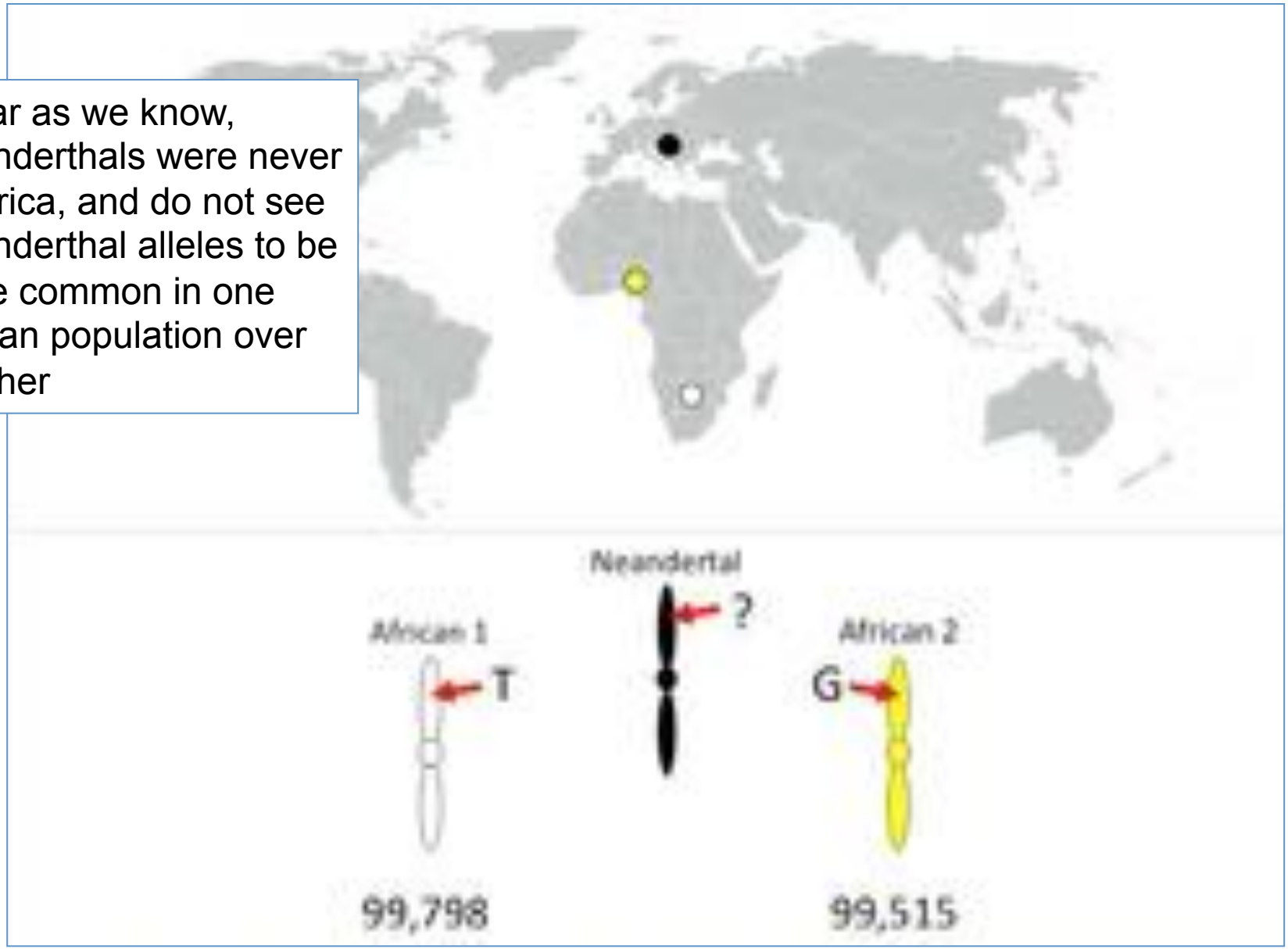
Genome coverage ~1.3 X

Did we mix?



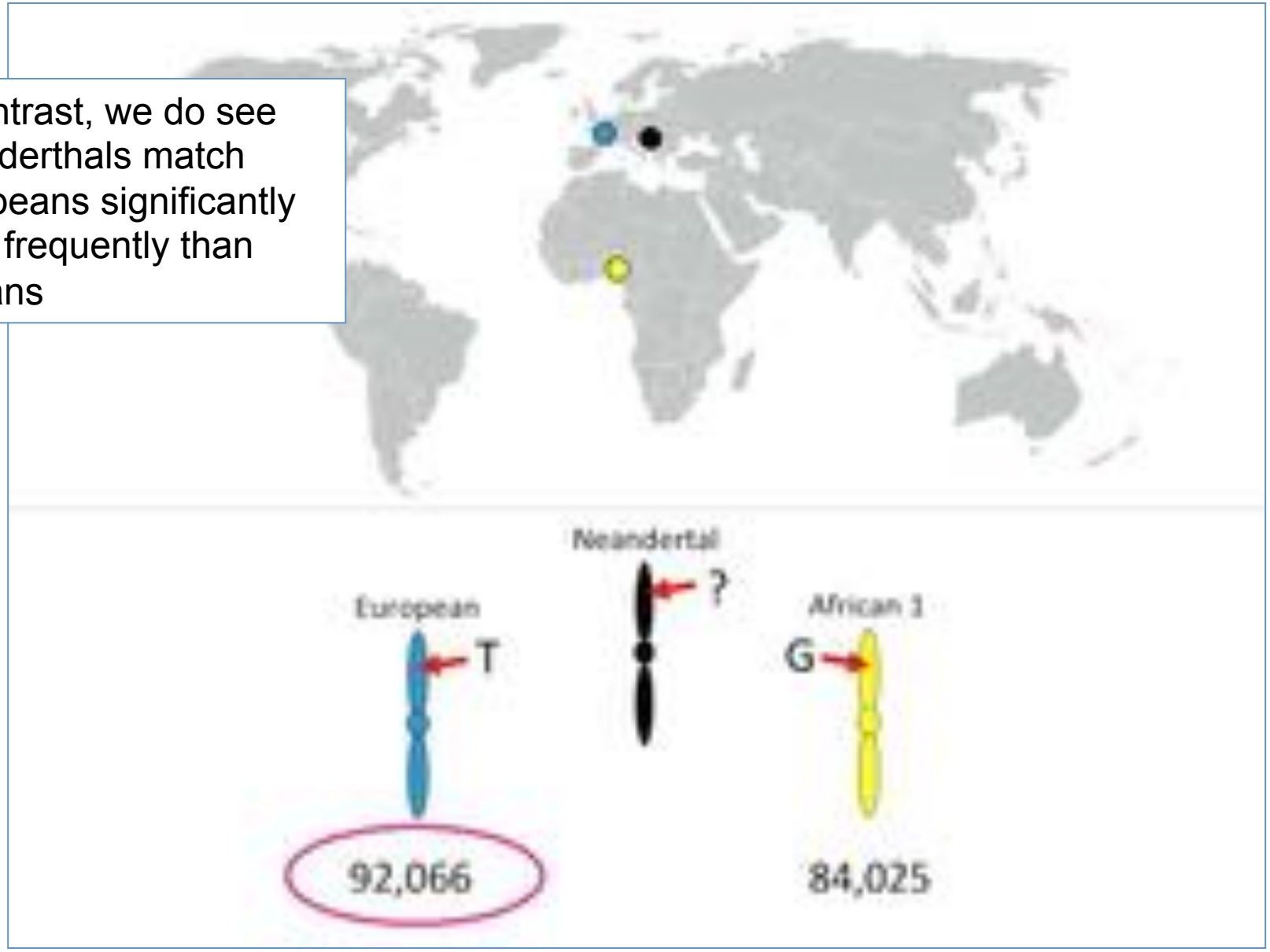
Did we mix?

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?

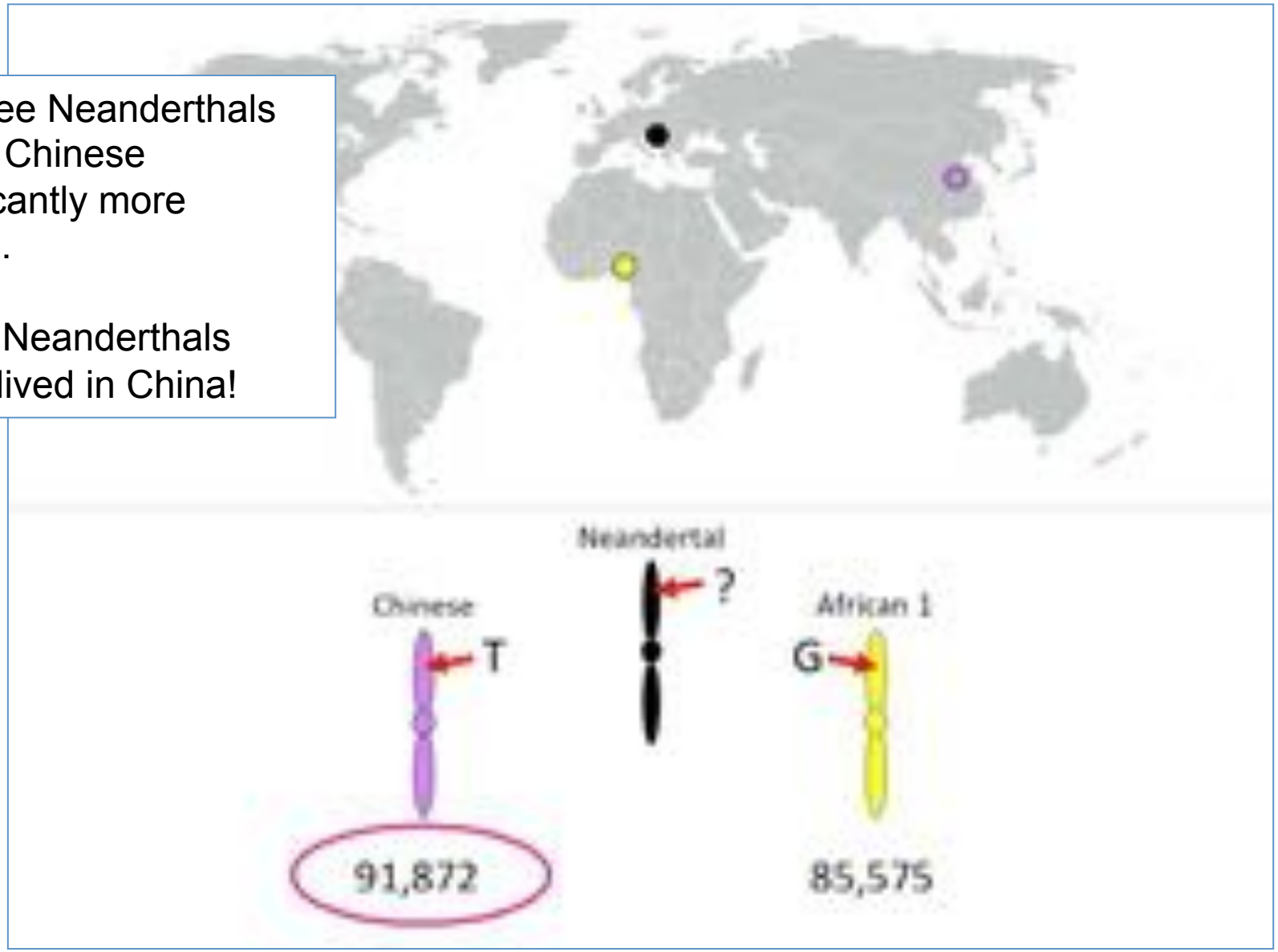
In contrast, we do see Neanderthals match Europeans significantly more frequently than Africans



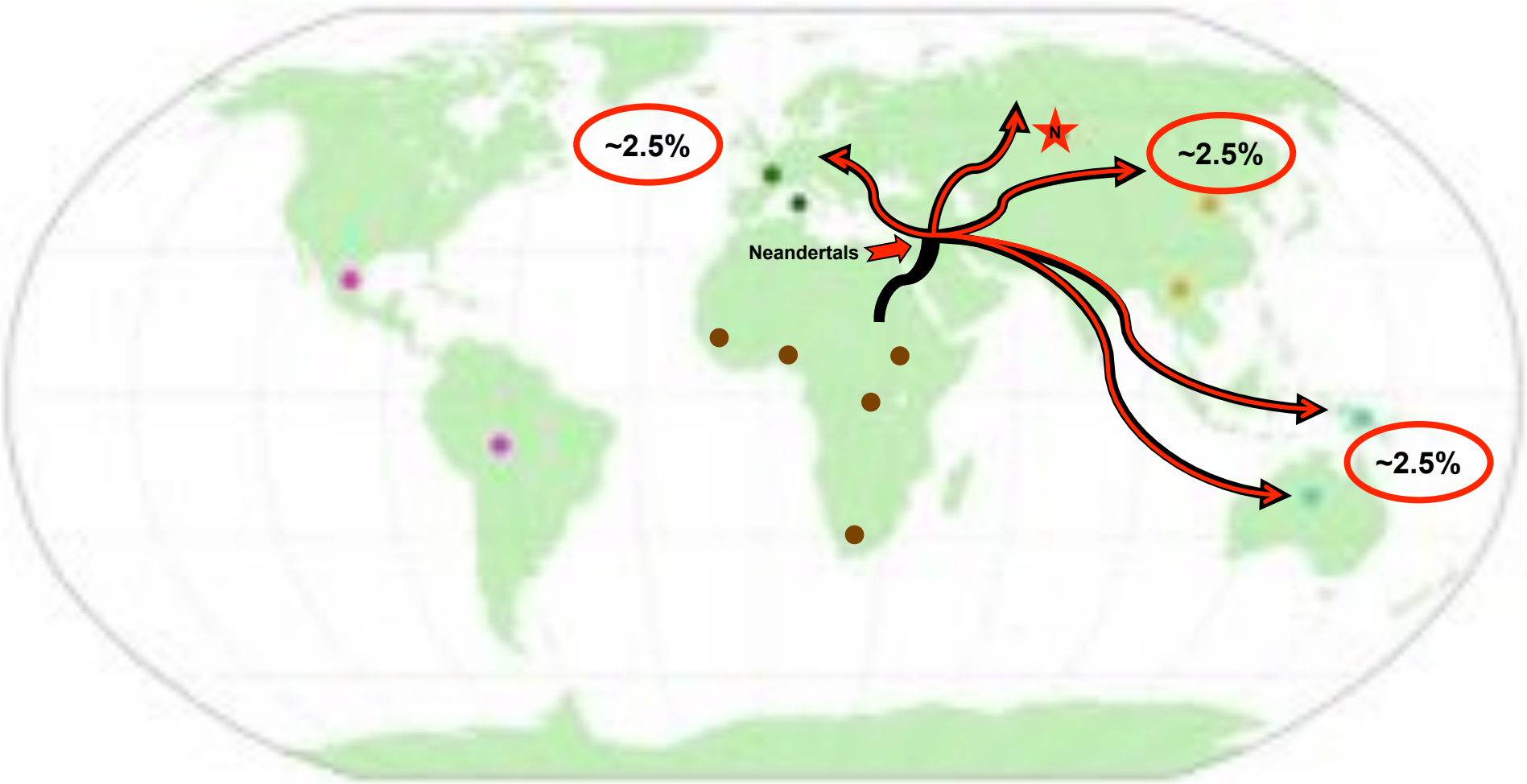
Did we mix?

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!

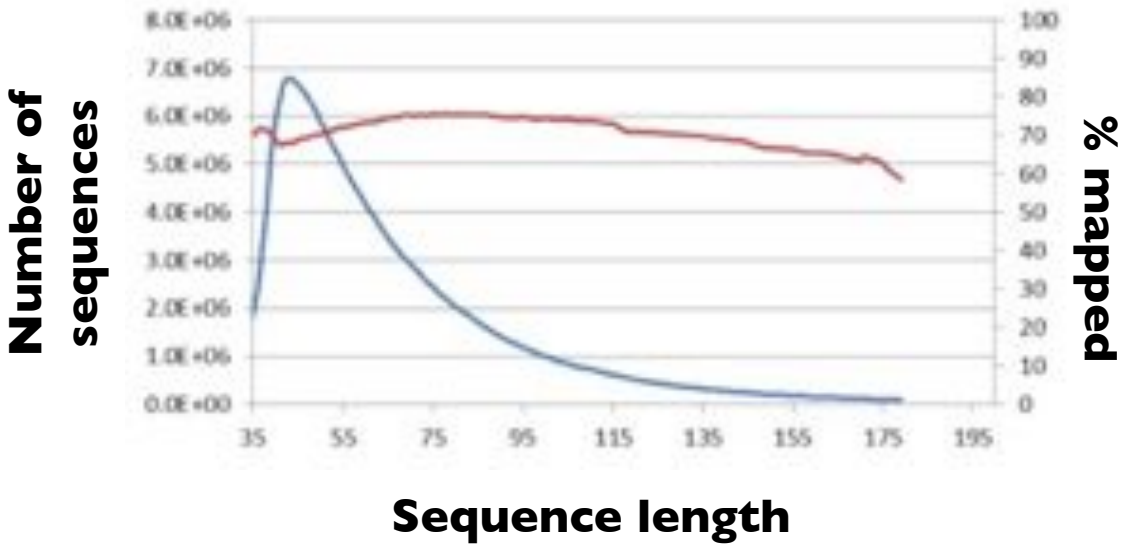
What about other ancient hominids?

Denisova cave
Altai mountains
Russia

Academician A.P. Derevianko

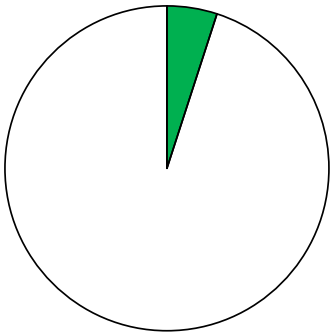

ONE CENT

Extraordinary preservation

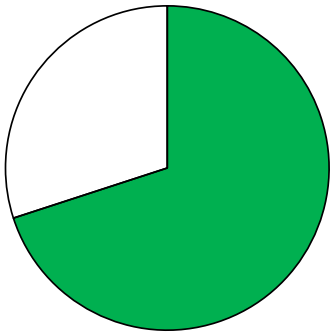


5% endogenous DNA

Best Neandertal bone



>70% endogenous DNA



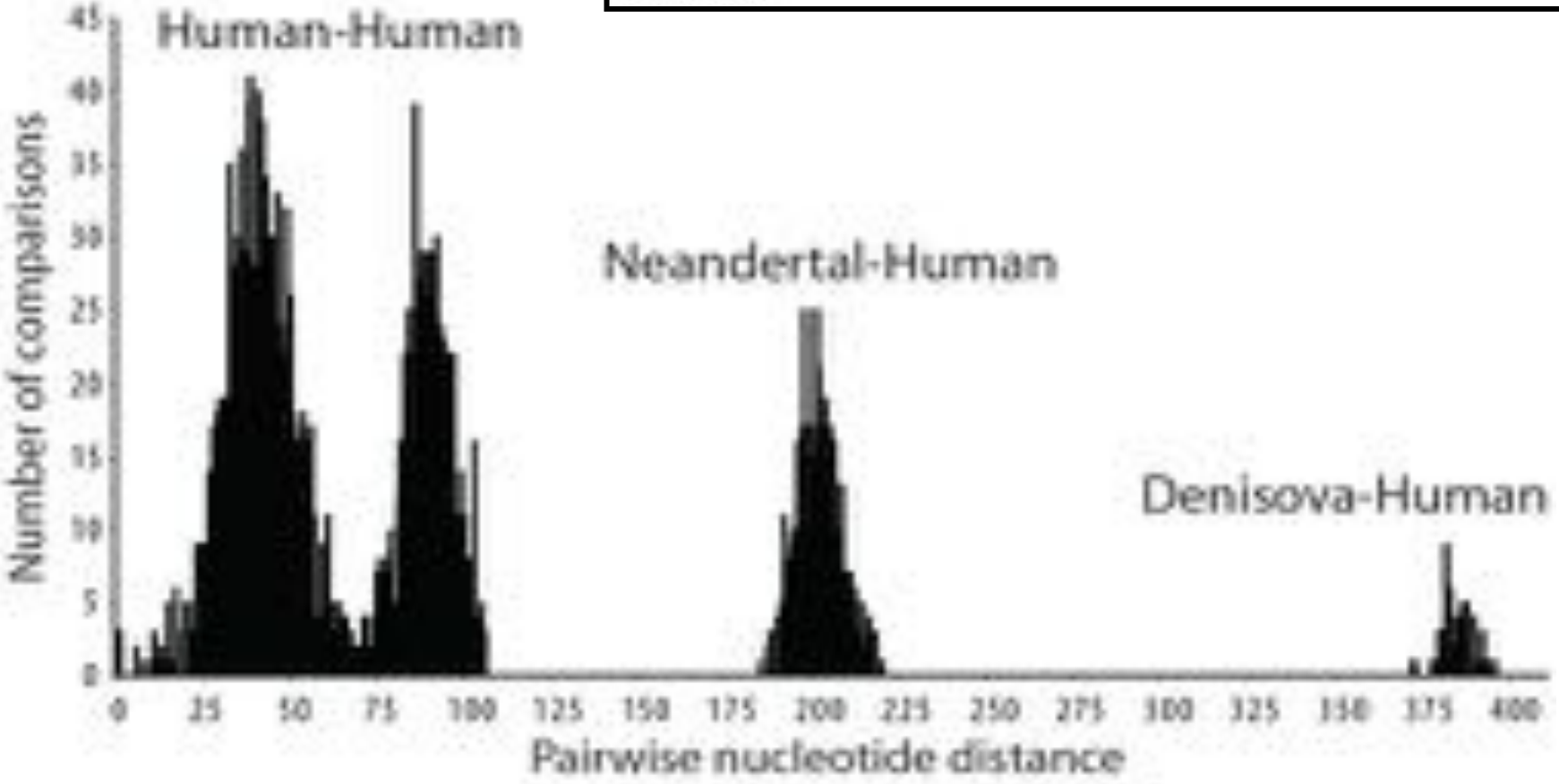
Denisova bone

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

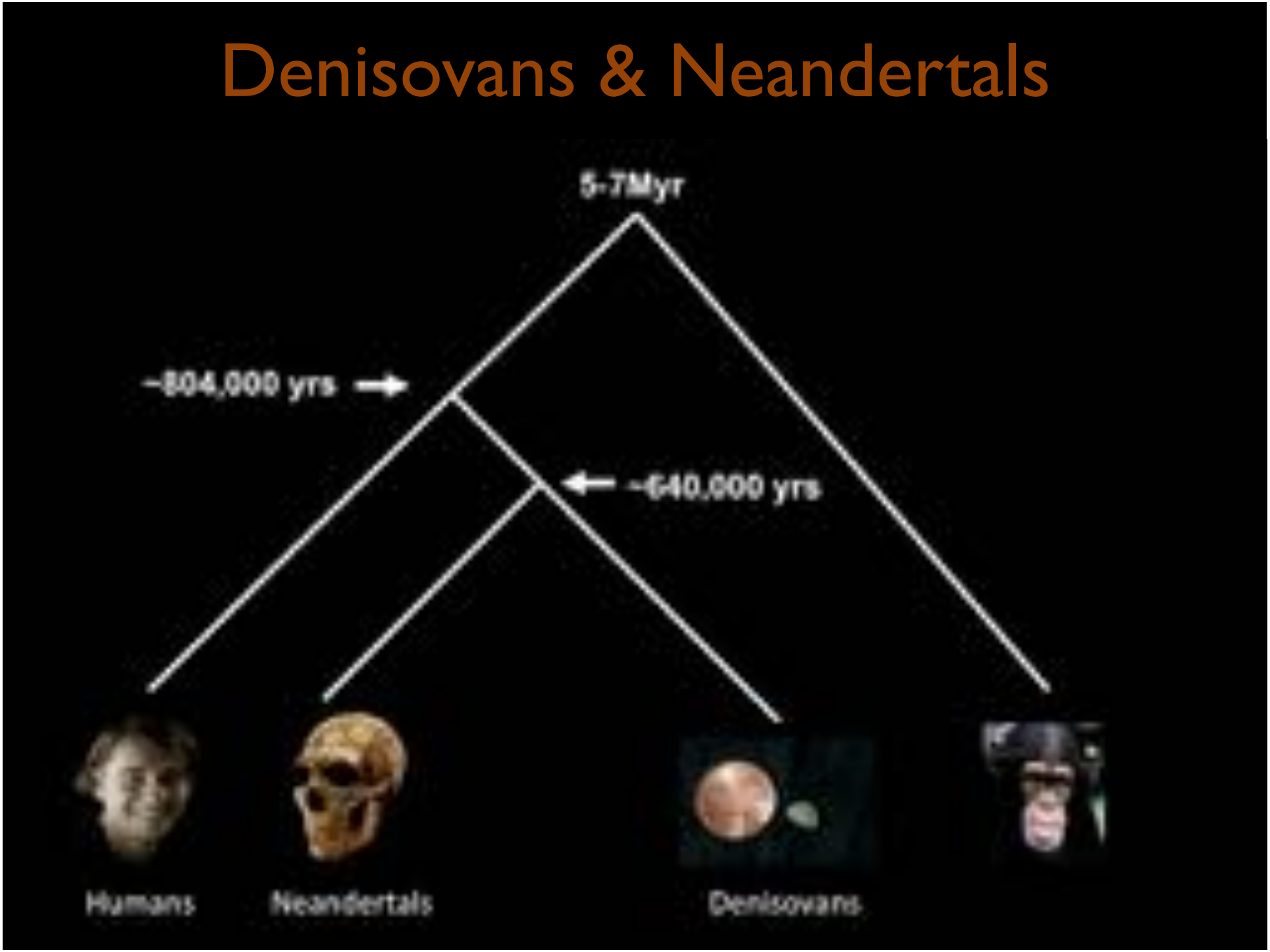
LETTERS

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]



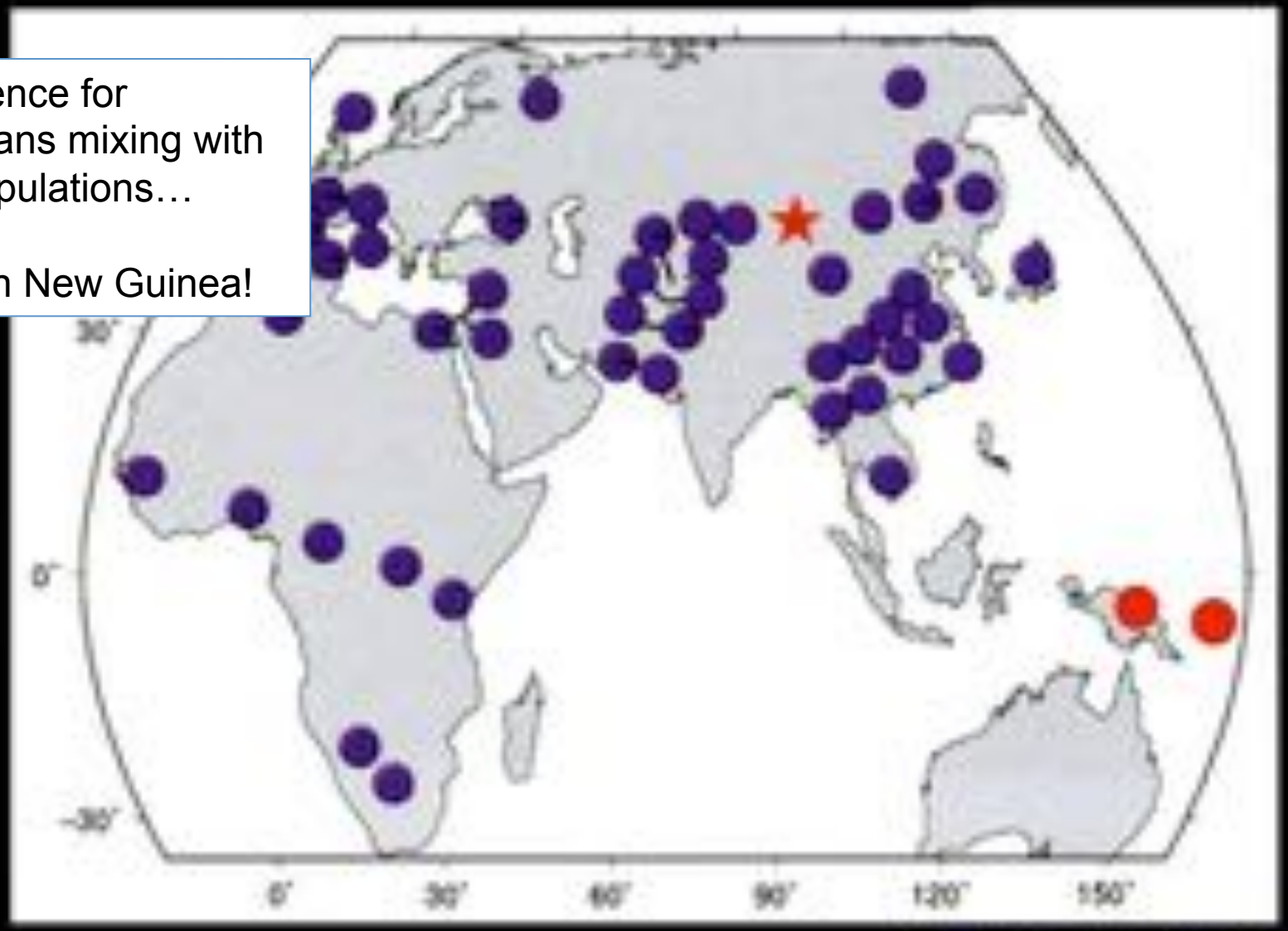
Denisovans & Neandertals



Did we mix?

No evidence for Denisovans mixing with other populations…

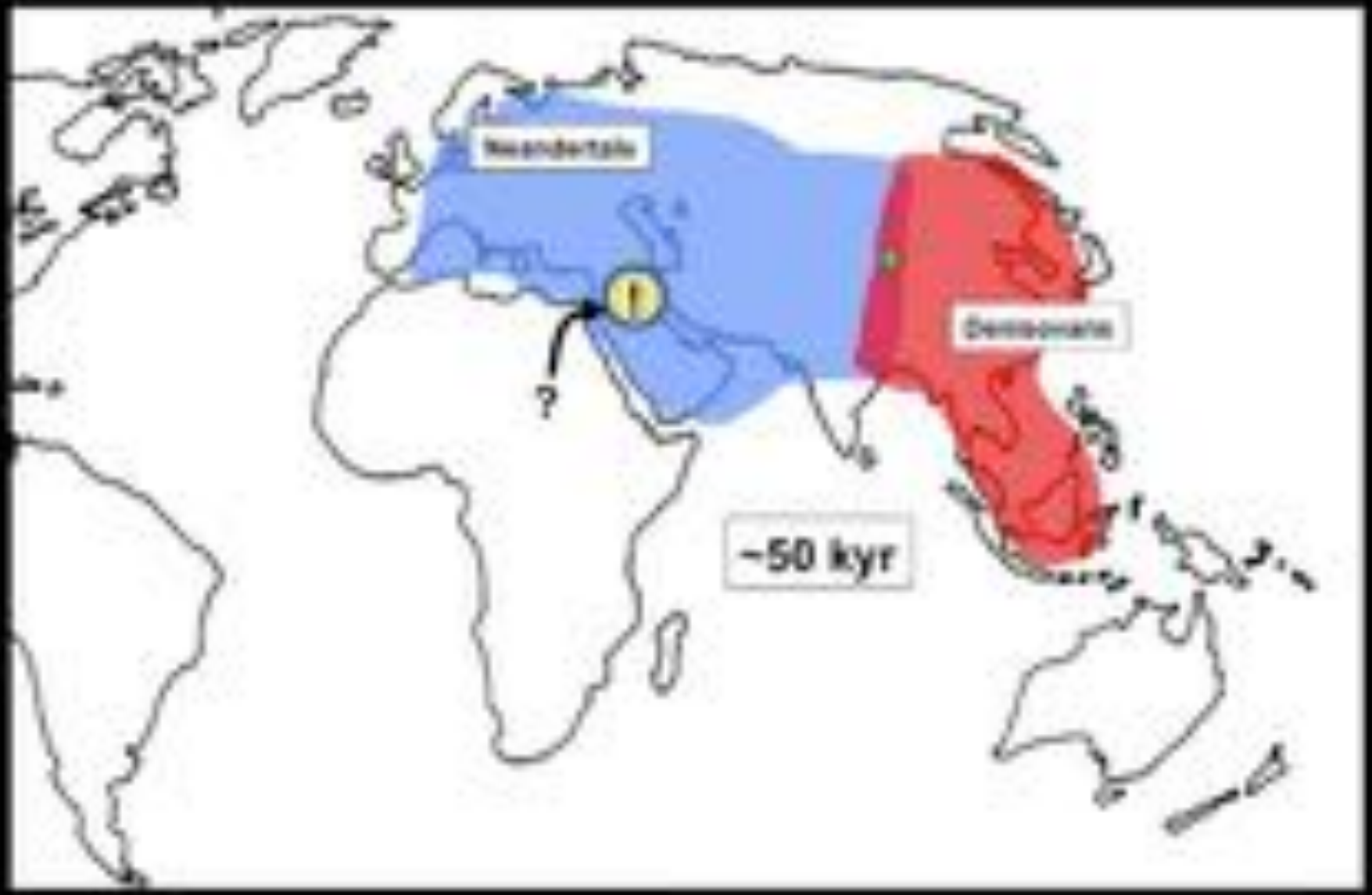
Except in New Guinea!



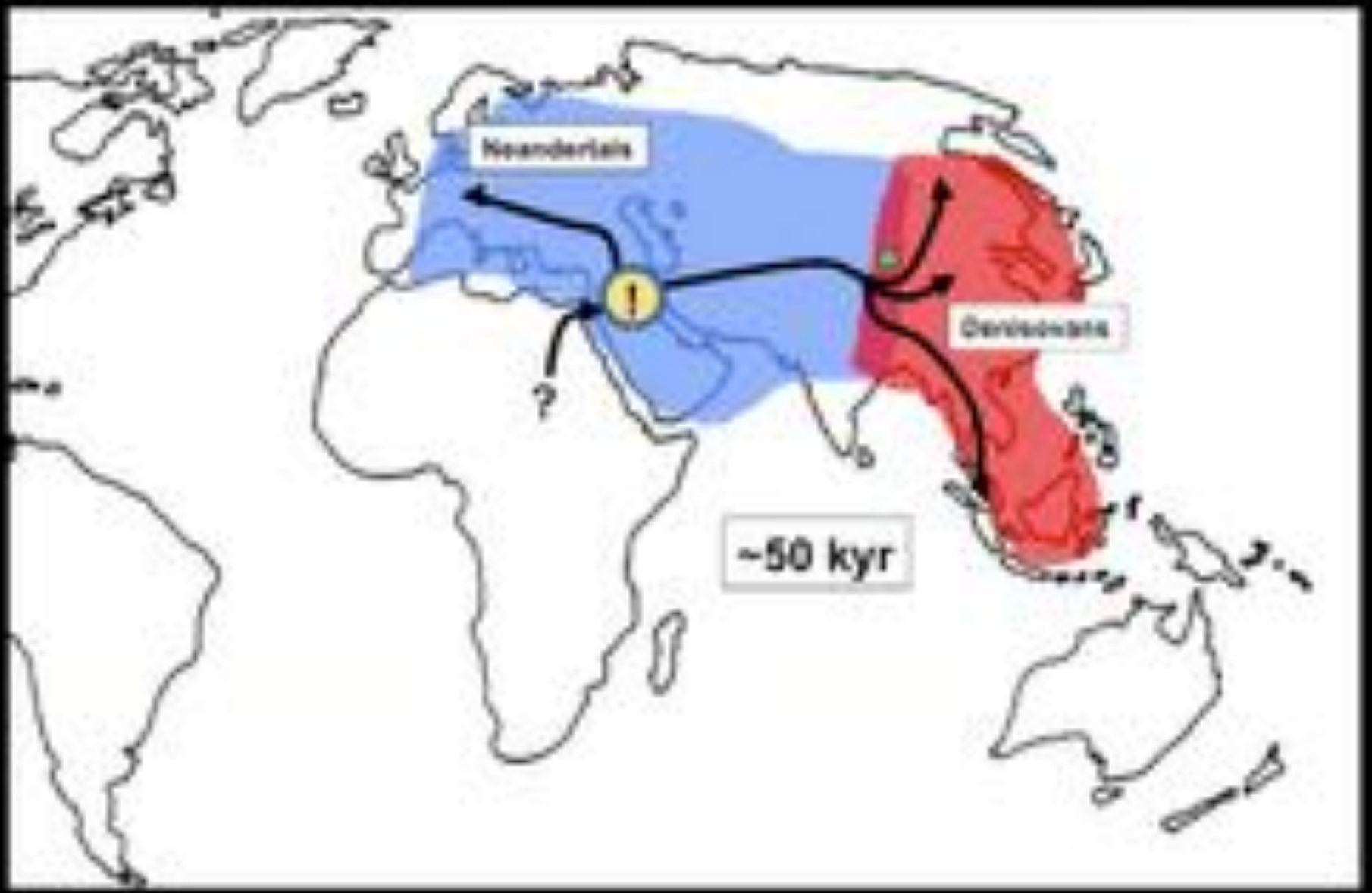
Map after Pickrell et al., 2009

Timeline of ancient hominids

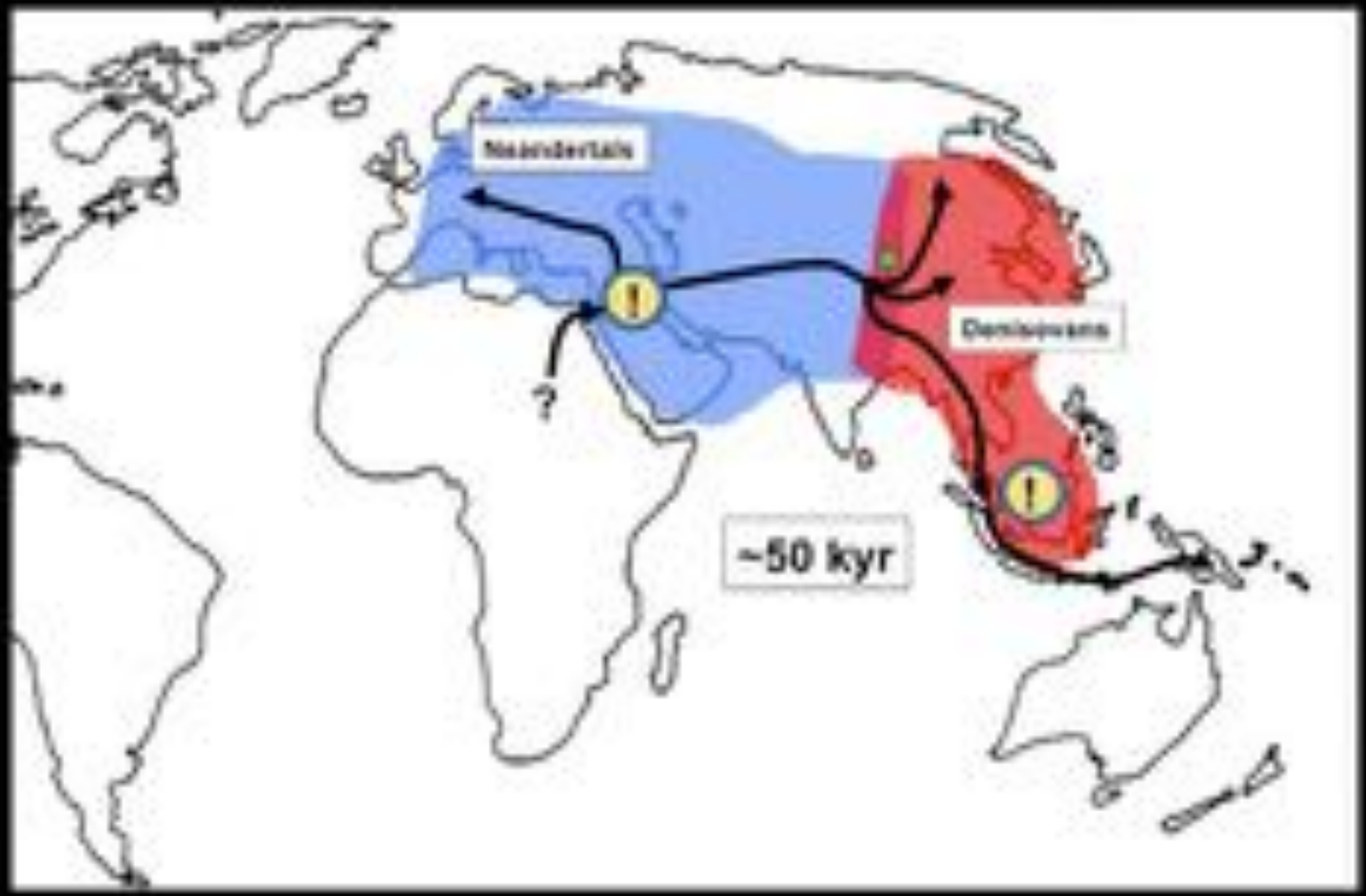
Timeline of ancient hominids



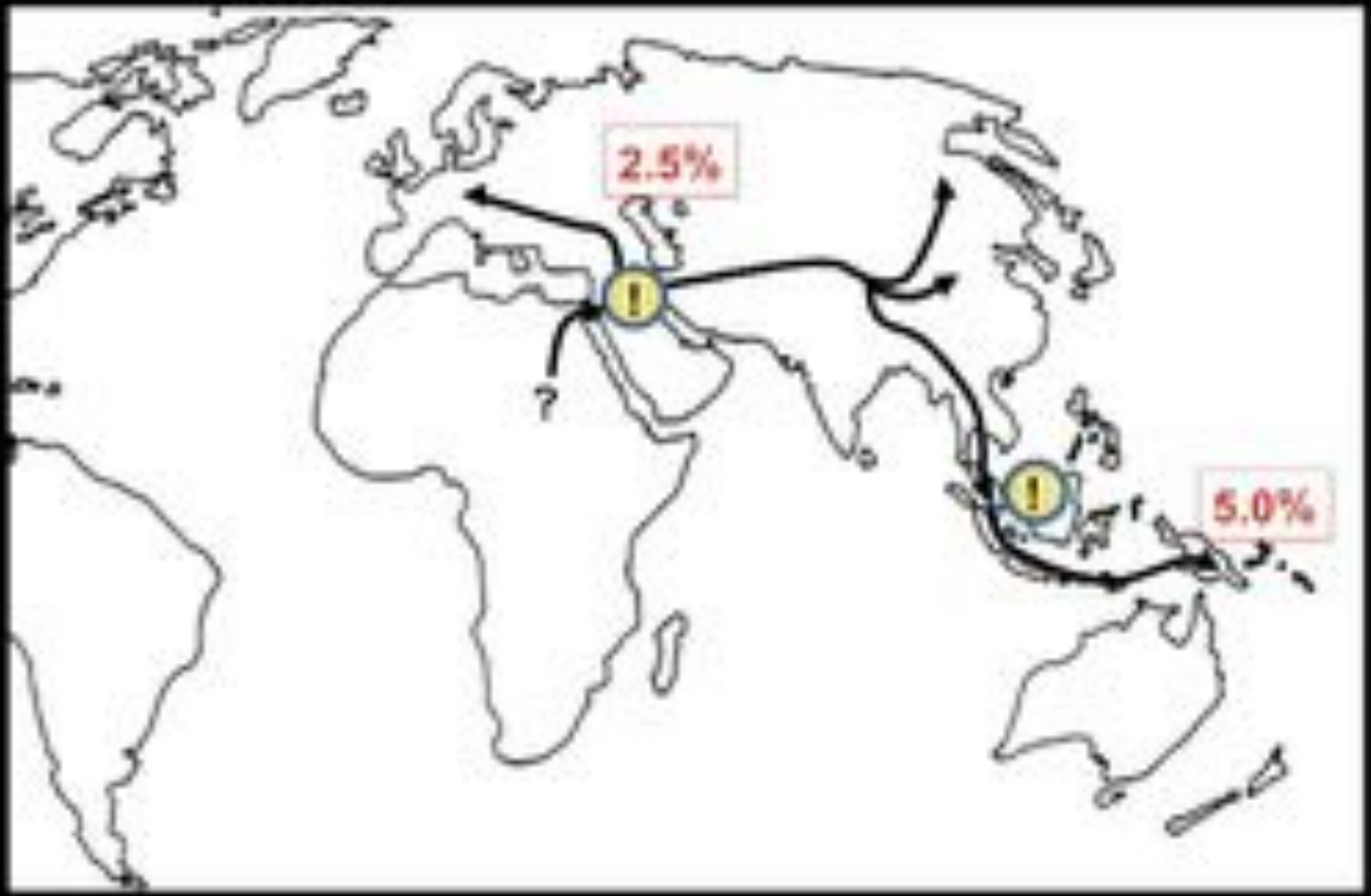
Timeline of ancient hominids



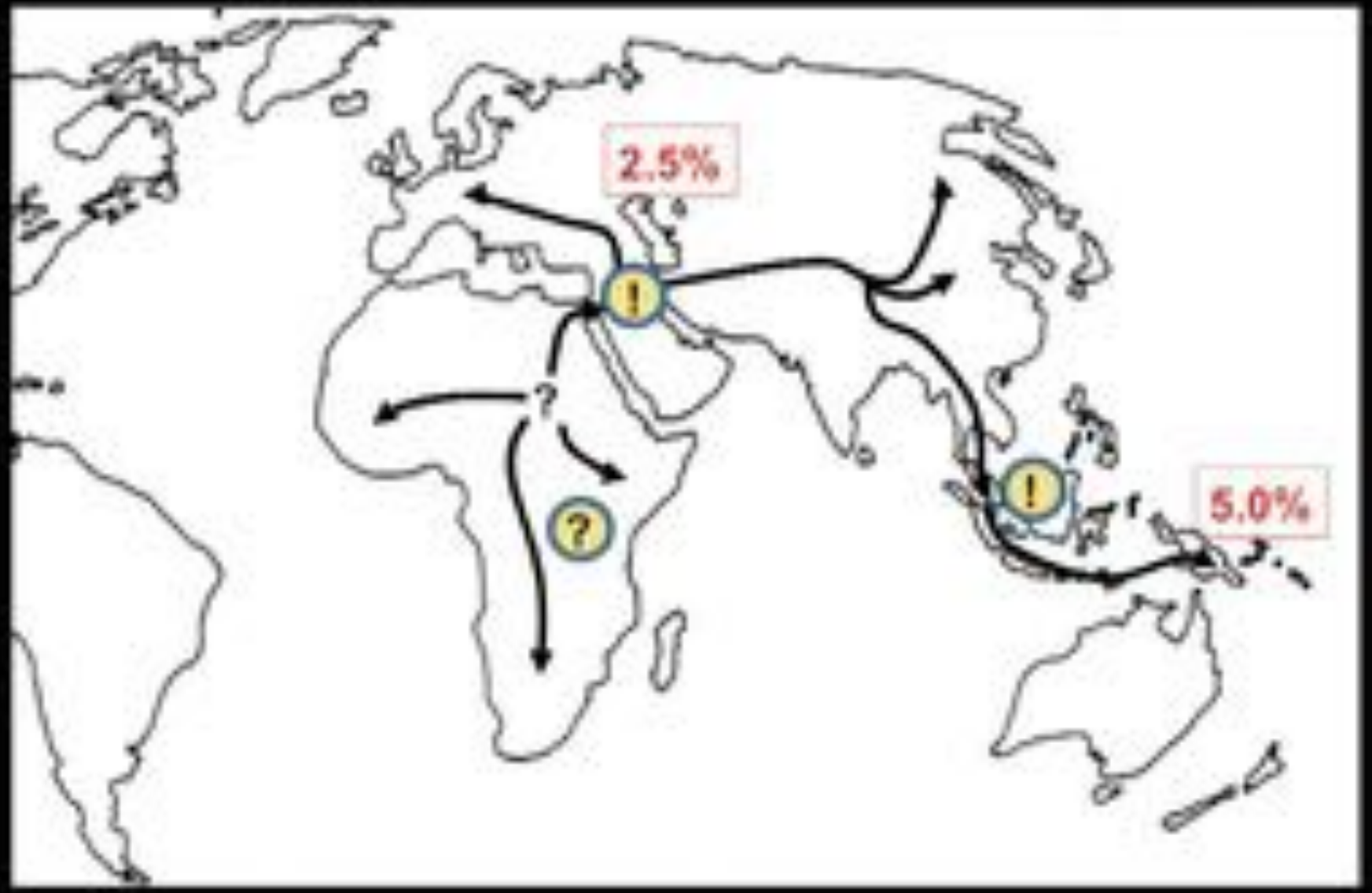
Timeline of ancient hominids



Timeline of ancient hominids

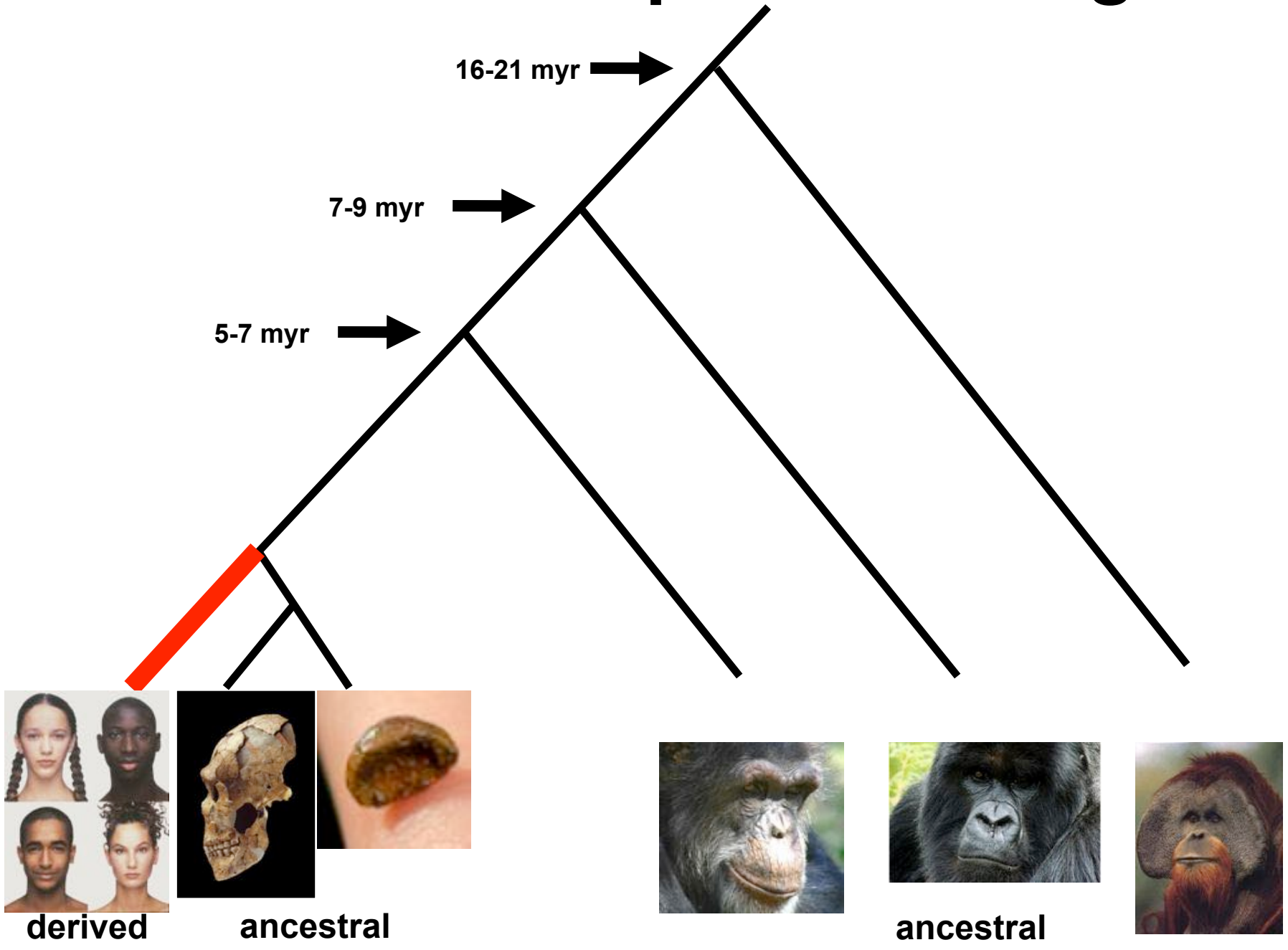


Timeline of ancient hominids



We have always mixed!

Modern human-specific changes



Recipe for a modern human

109,295 single nucleotide changes (SNCs)

7,944 insertions and deletions

Changes in protein coding genes

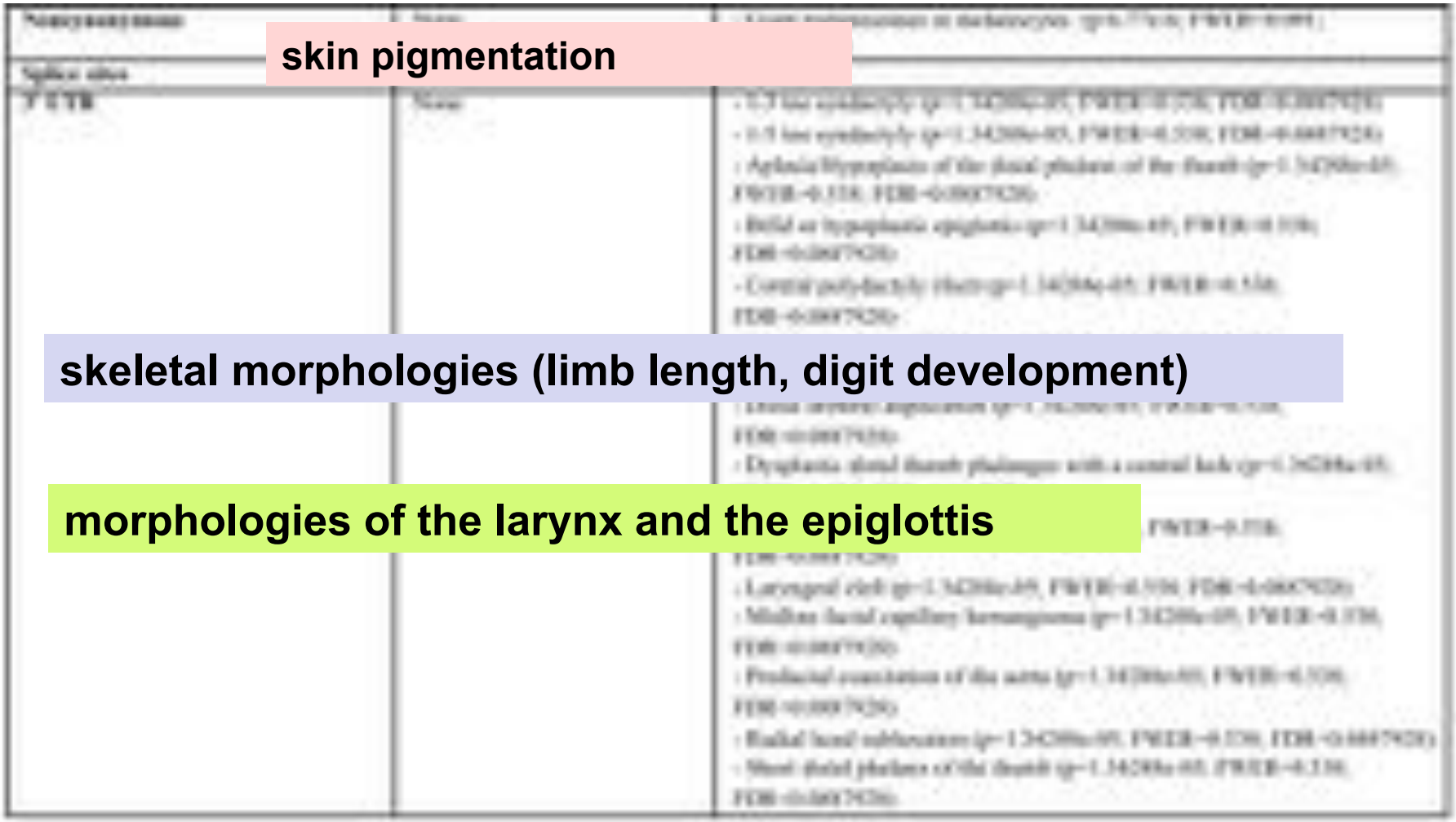
277 cause fixed amino acid substitutions

87 affect splice sites

Changes in Non-coding & regulatory sequences

26 affect well-defined motifs inside regulatory regions

Enrichment analysis

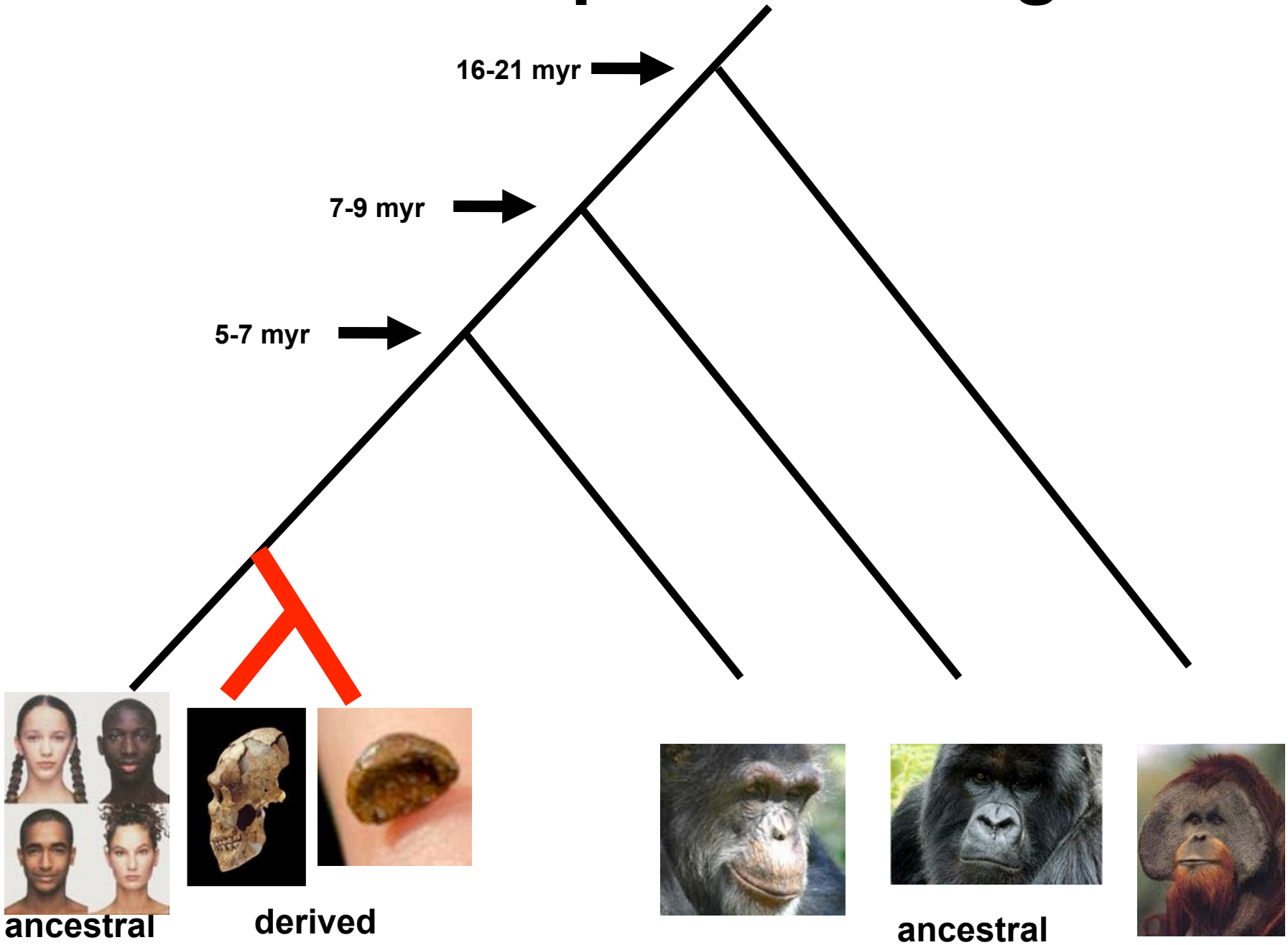


skin pigmentation

skeletal morphologies (limb length, digit development)

morphologies of the larynx and the epiglottis

Neandertal-specific changes



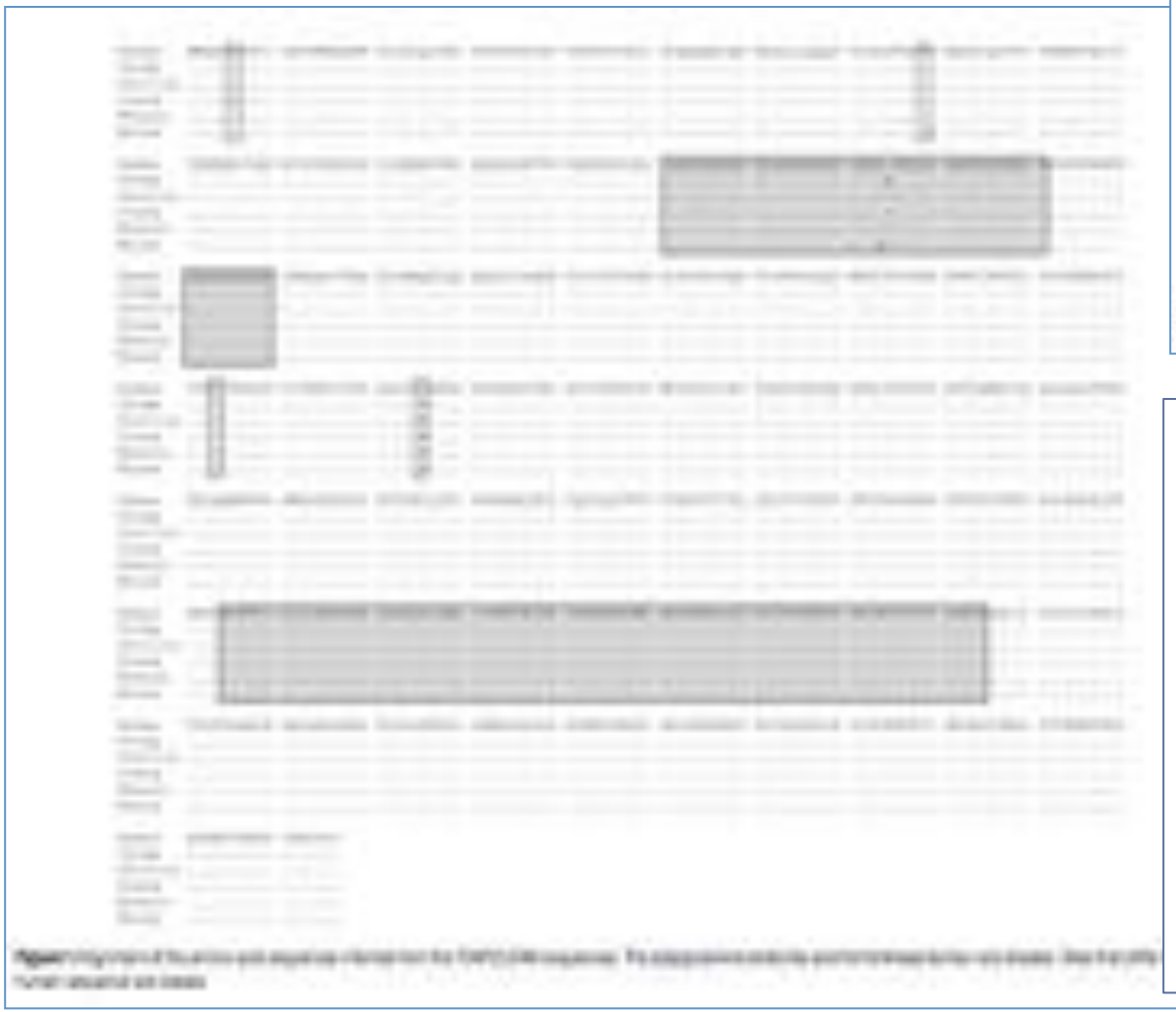
Enrichment analysis

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

Skeletal and hair morphology

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

FOXP2 Analysis



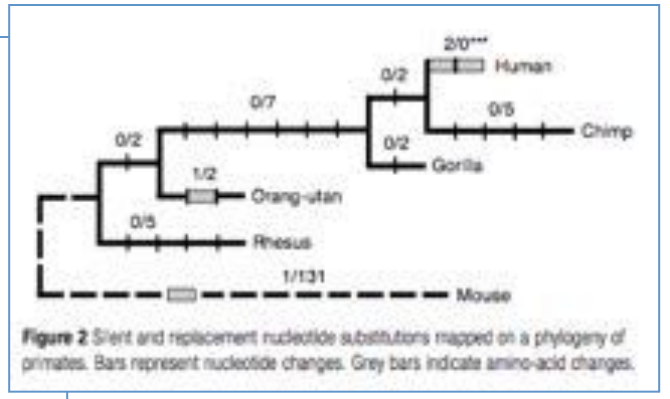


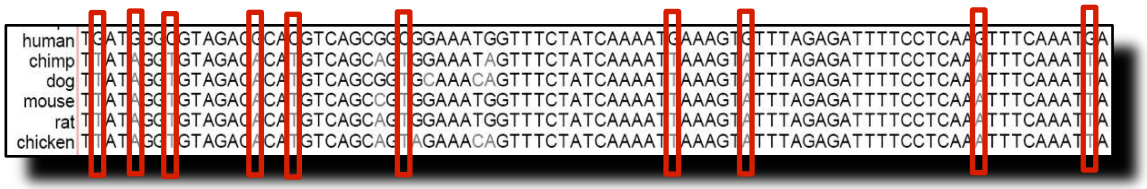
Figure 2 Silent and replacement nucleotide substitutions mapped on a phylogeny of primates. Bars represent nucleotide changes. Grey bars indicate amino-acid changes.

- 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

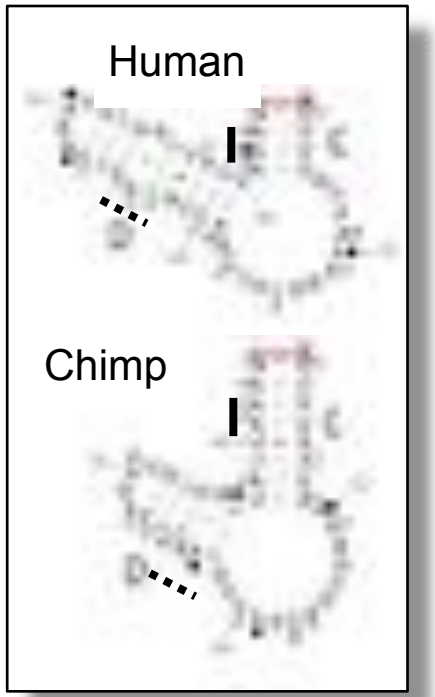
What makes us human?

"*Human Accelerated Regions*"



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

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 ➔ $(CAA)_7$

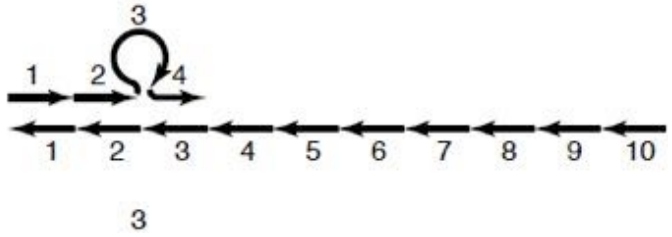
tTTGTCTTGTCTTGTCTTGTCTTGTCTTGTCc ➔ $(TTGTC)_6$ cCATTCATTCATTCATTa ➔ $(CATT)_4$

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

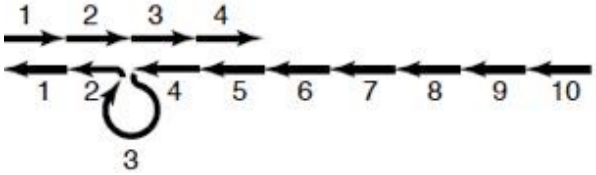
Replication slippage

- **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:



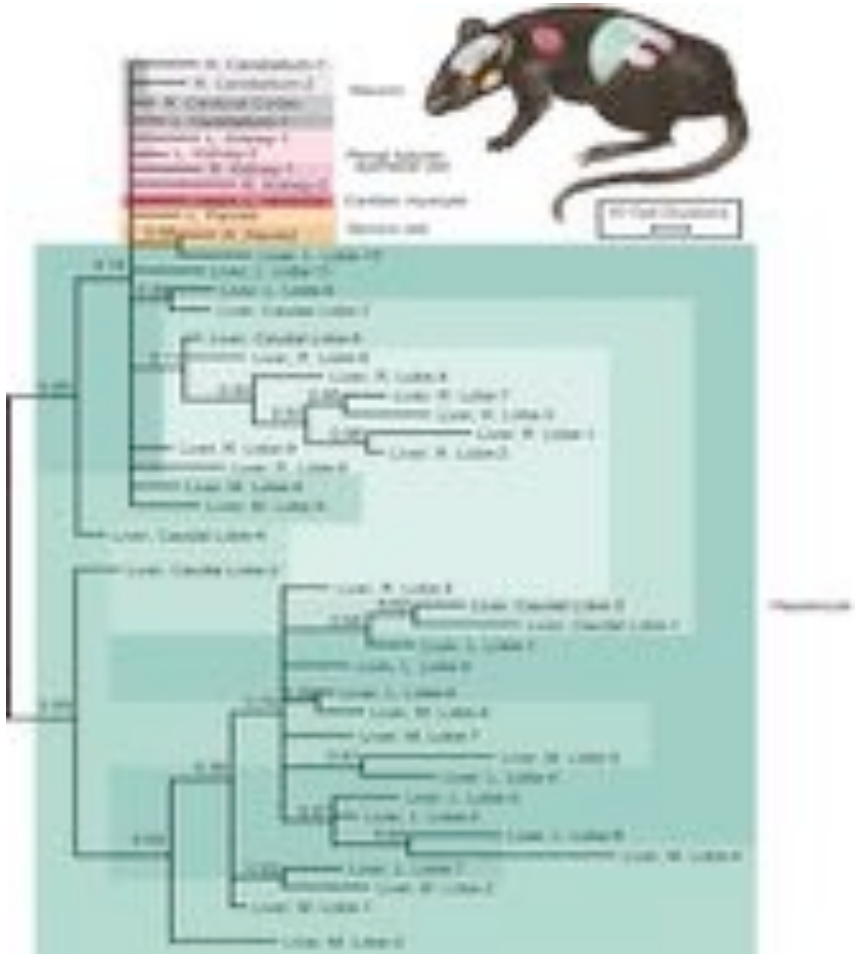
Contraction:



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

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

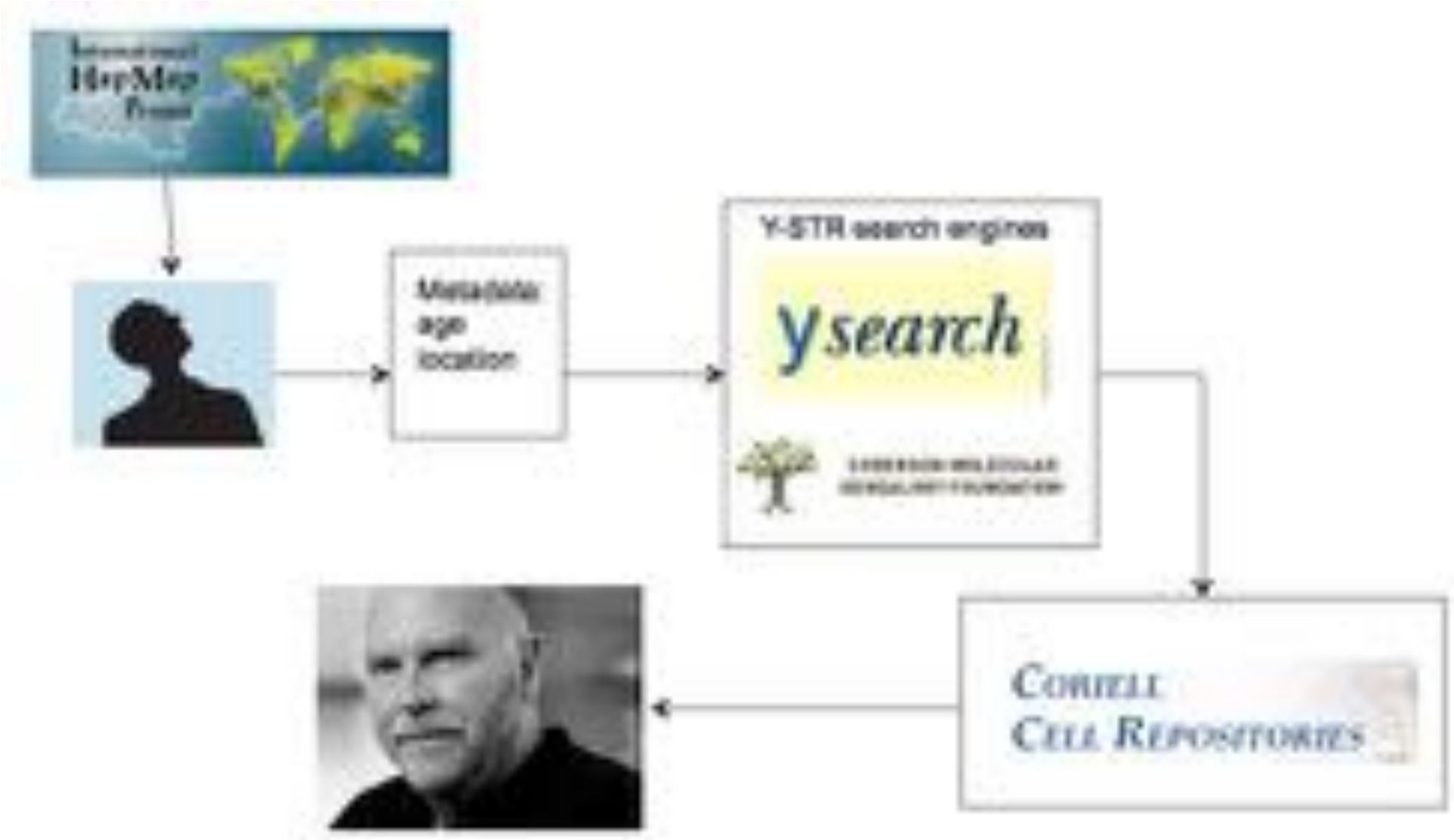


Phylogenetic fate mapping
Salipante (2006) *PNAS*. doi: 10.1073/pnas.0601265103

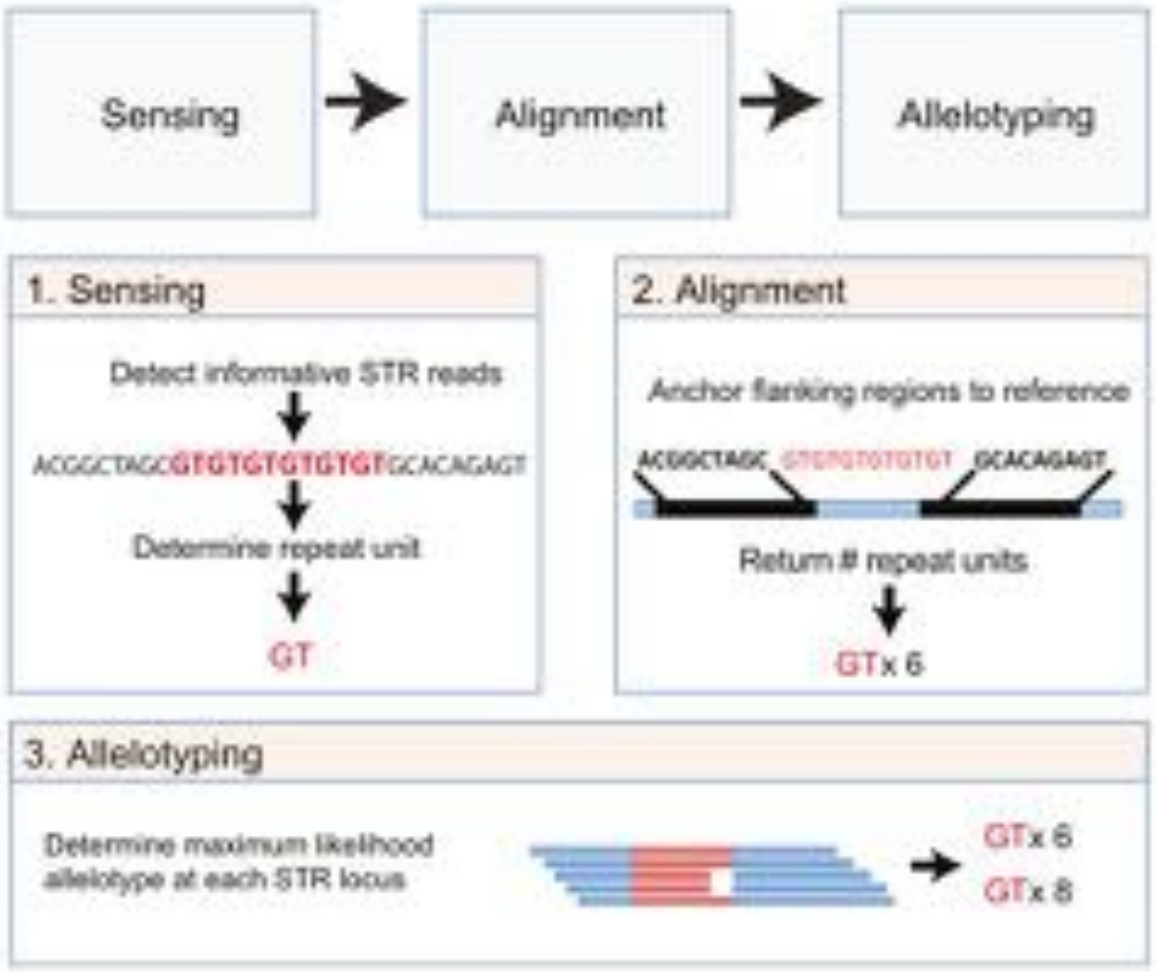
Genealogy Databases

Genealogy Databases Enable Naming Of Anonymous DNA Donors

Surname Inference Overview



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

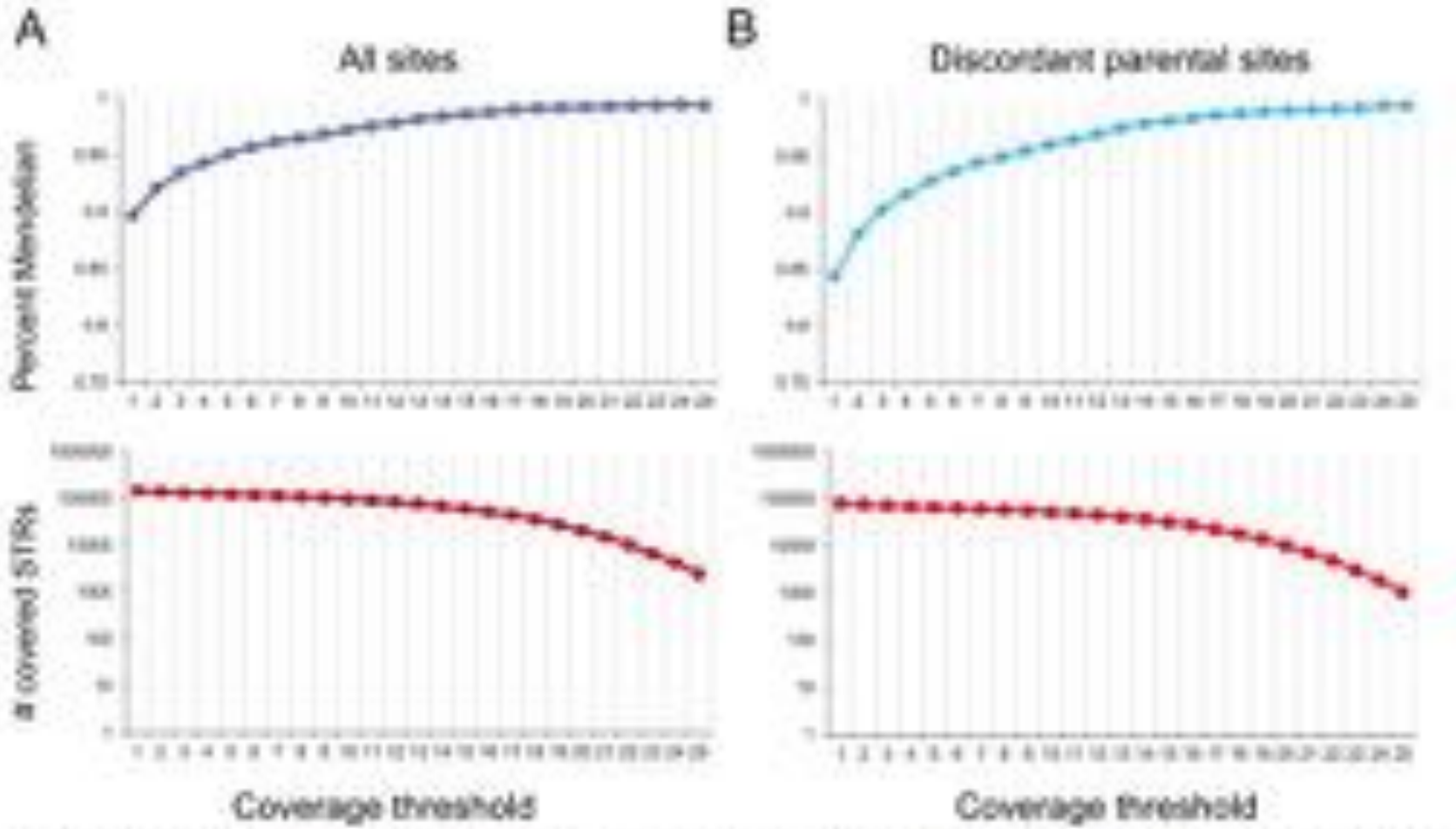
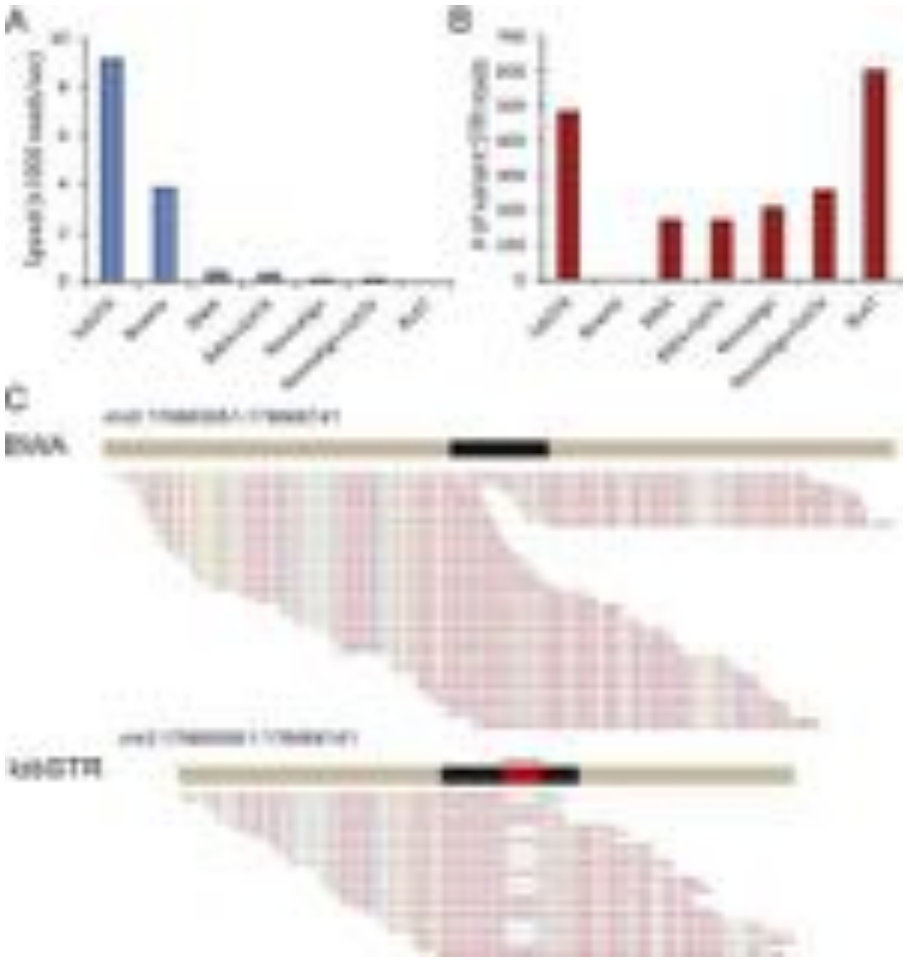


Figure 4. Validating lobSTR by Mendelian inheritance in a HapMap trio. Mendelian inheritance (blue and cyan) rose to 99% above 17× coverage. (Dark and light red) The number of covered loci at each coverage threshold. (*A*) Mendelian inheritance of all covered loci. (*B*) Mendelian inheritance of loci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

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.

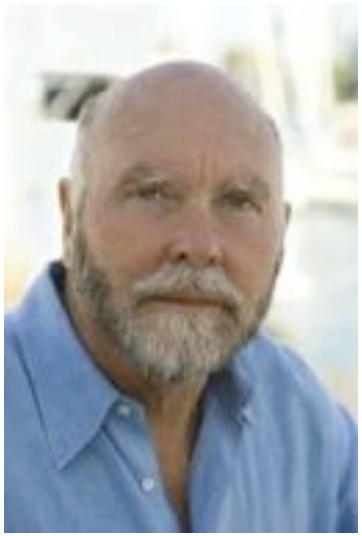
http://www.ysearch.org

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

http://www.ussearch.com

Surname Inference

It's Craig Venter!



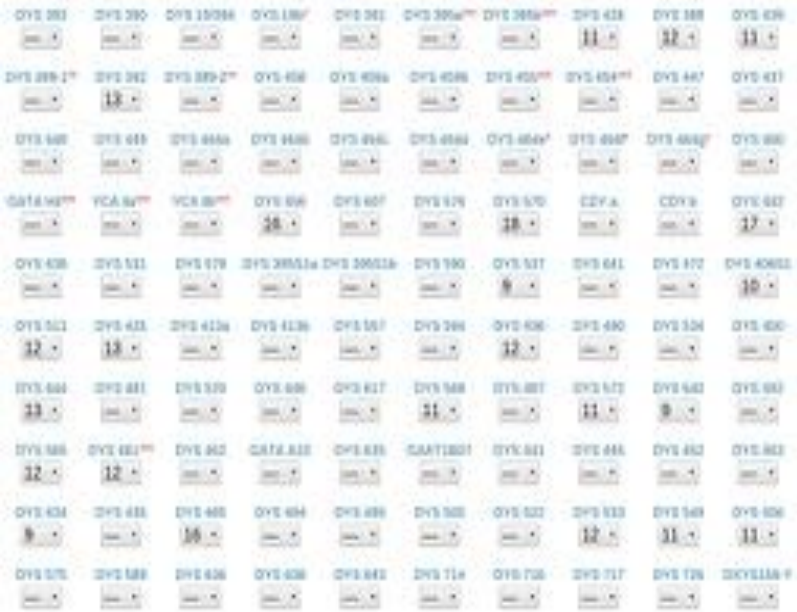
Identifying Personal Genomes by Surname Inference
Gymrek et al (2013) *Science*. doi: 10.1126/science.1229566

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



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

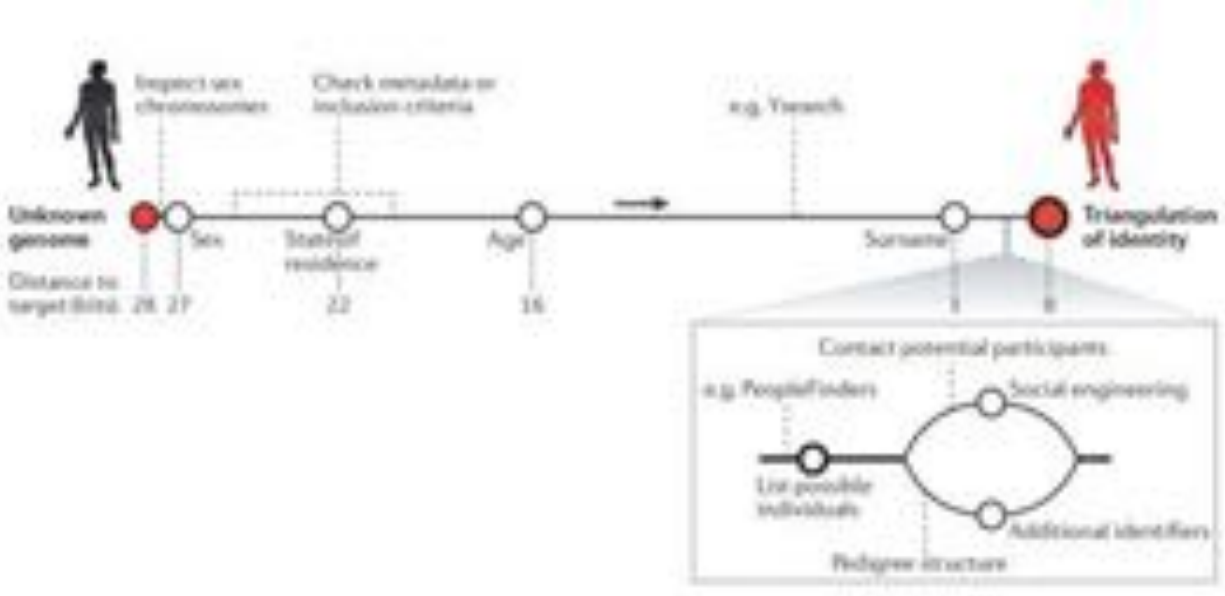
Identifiers and Quasi-identifiers

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

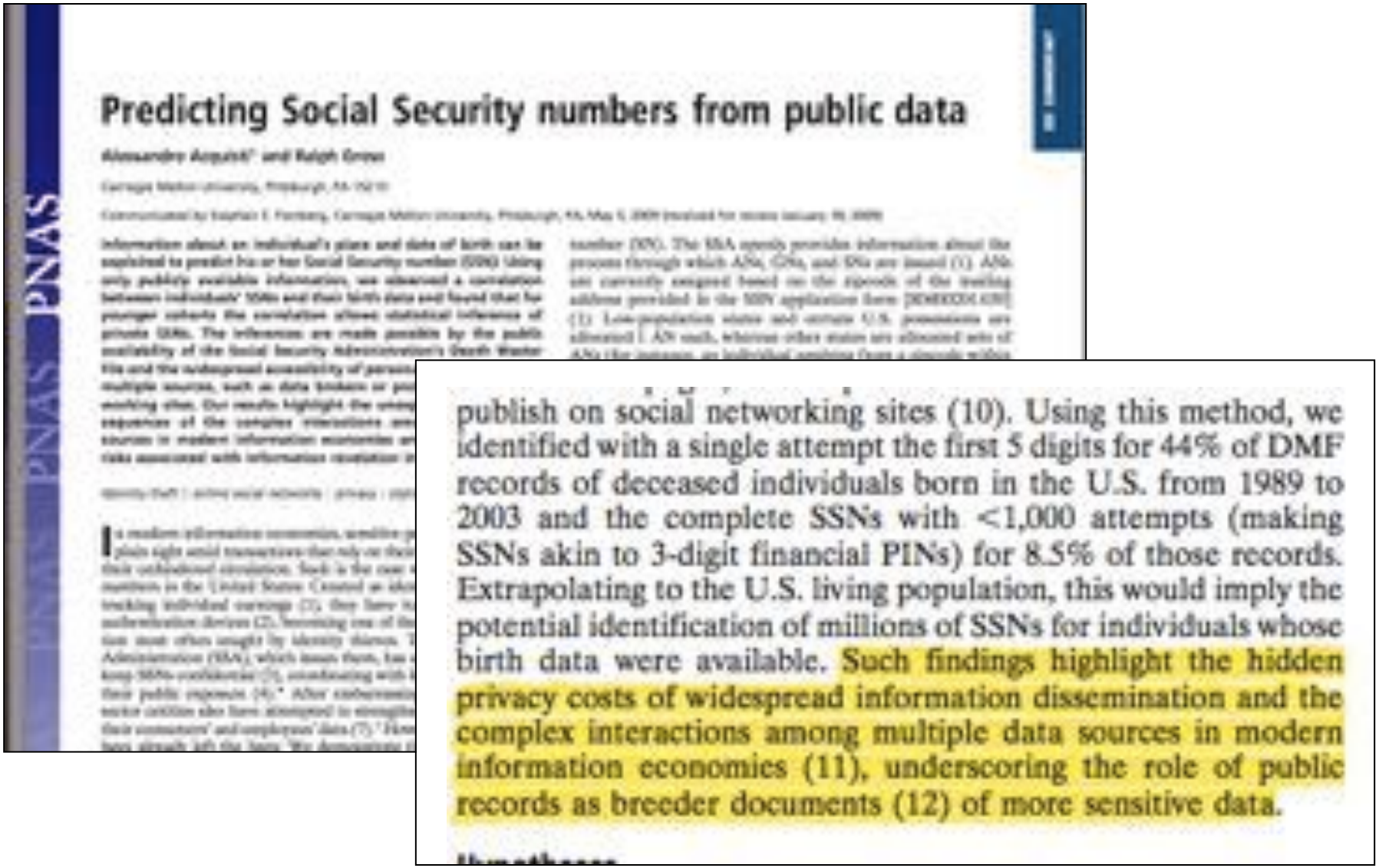
Possible route for identity tracing



- *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.

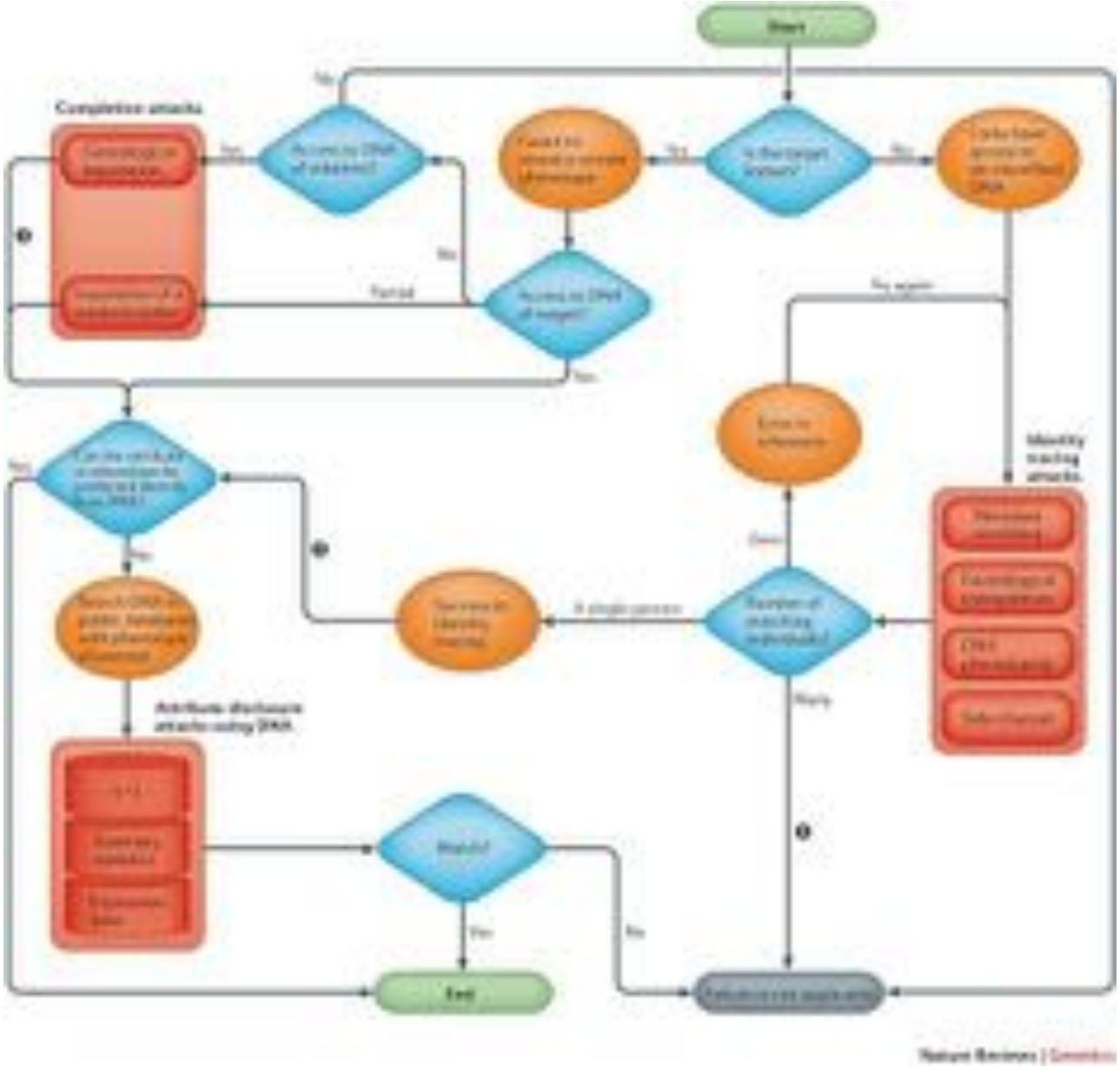
The risks of big data?



Predicting Social Security numbers from public data

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



Next class

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