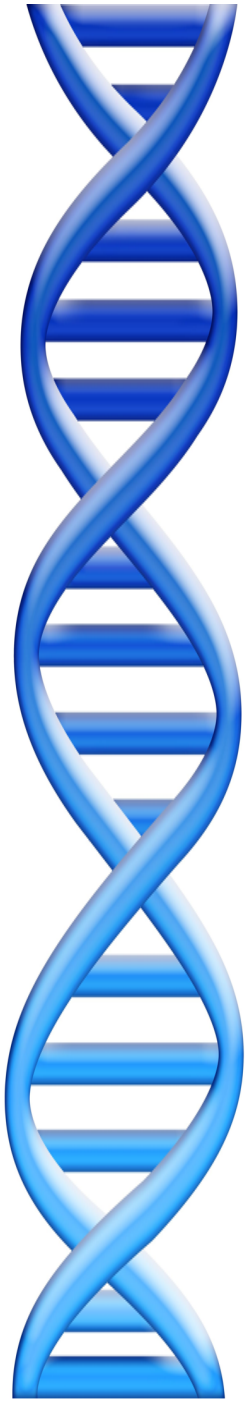
Lecture 18. Ancient and Modern Humans

Michael Schatz

April 13, 2017

JHU 600.649: Applied Comparative Genomics



Part 1:
Clustering Refresher

Clustering Refresher

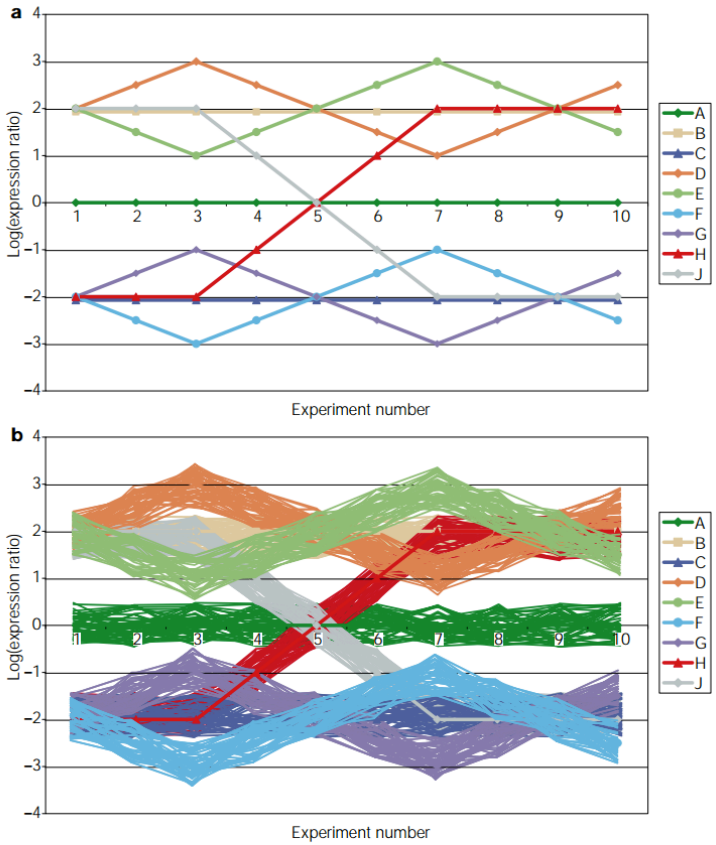
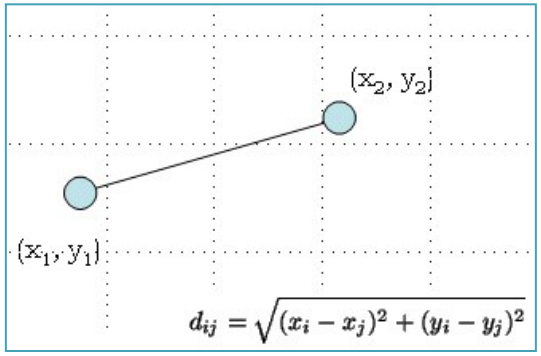
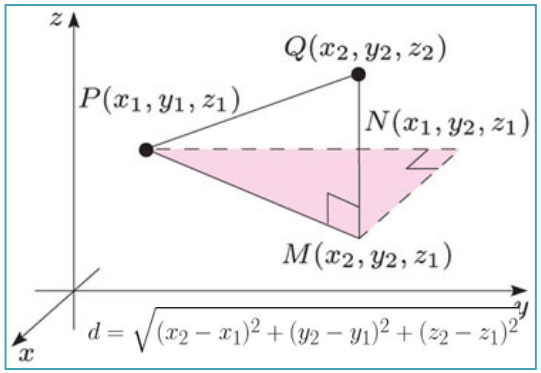


Figure 2 | **A synthetic gene-expression data set.** This data set provides an opportunity to evaluate how various clustering algorithms reveal different features of the data. **a** | Nine distinct gene-expression patterns were created with $\log_2$(ratio) expression measures defined for ten experiments.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

Euclidean Distance

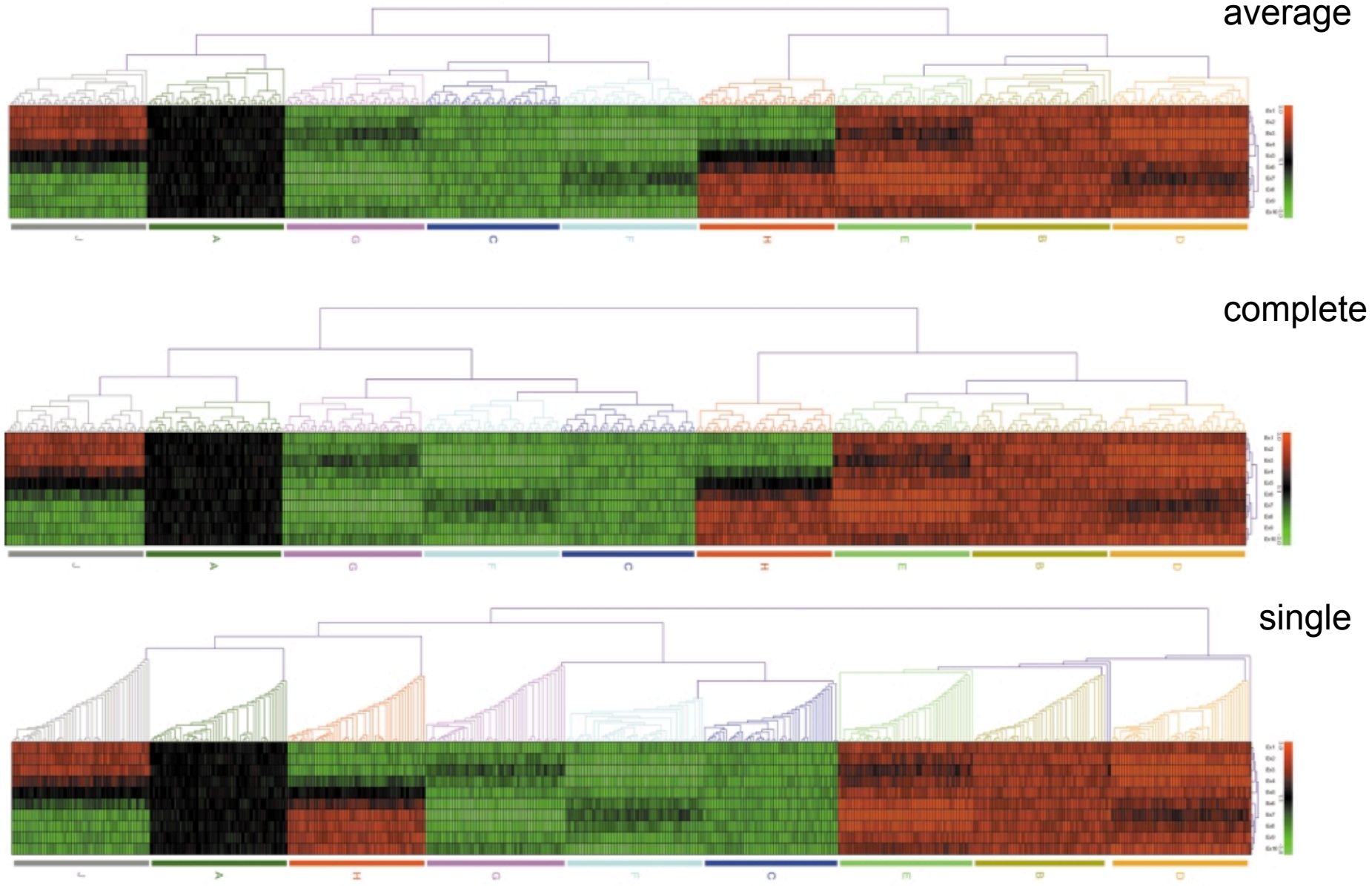




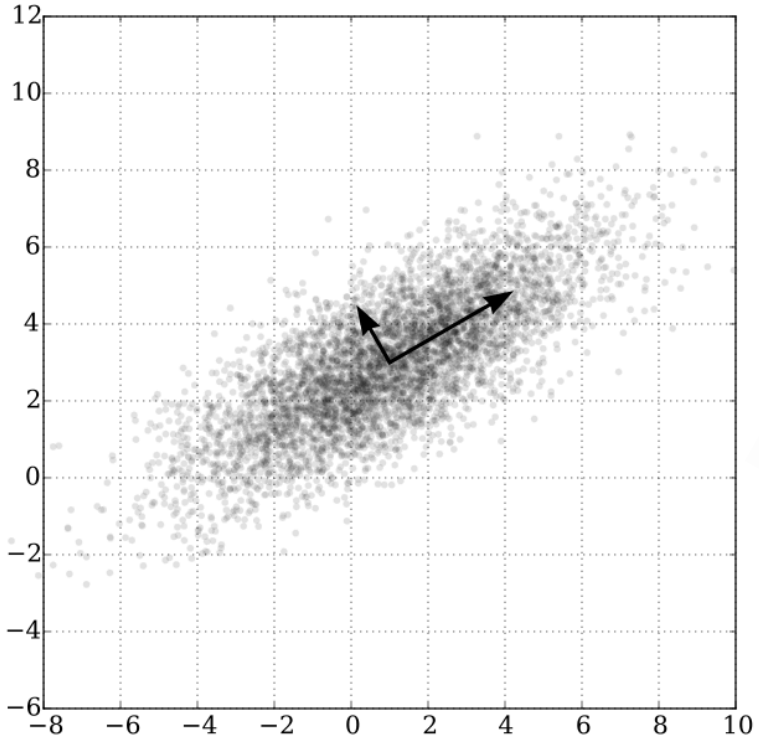
$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

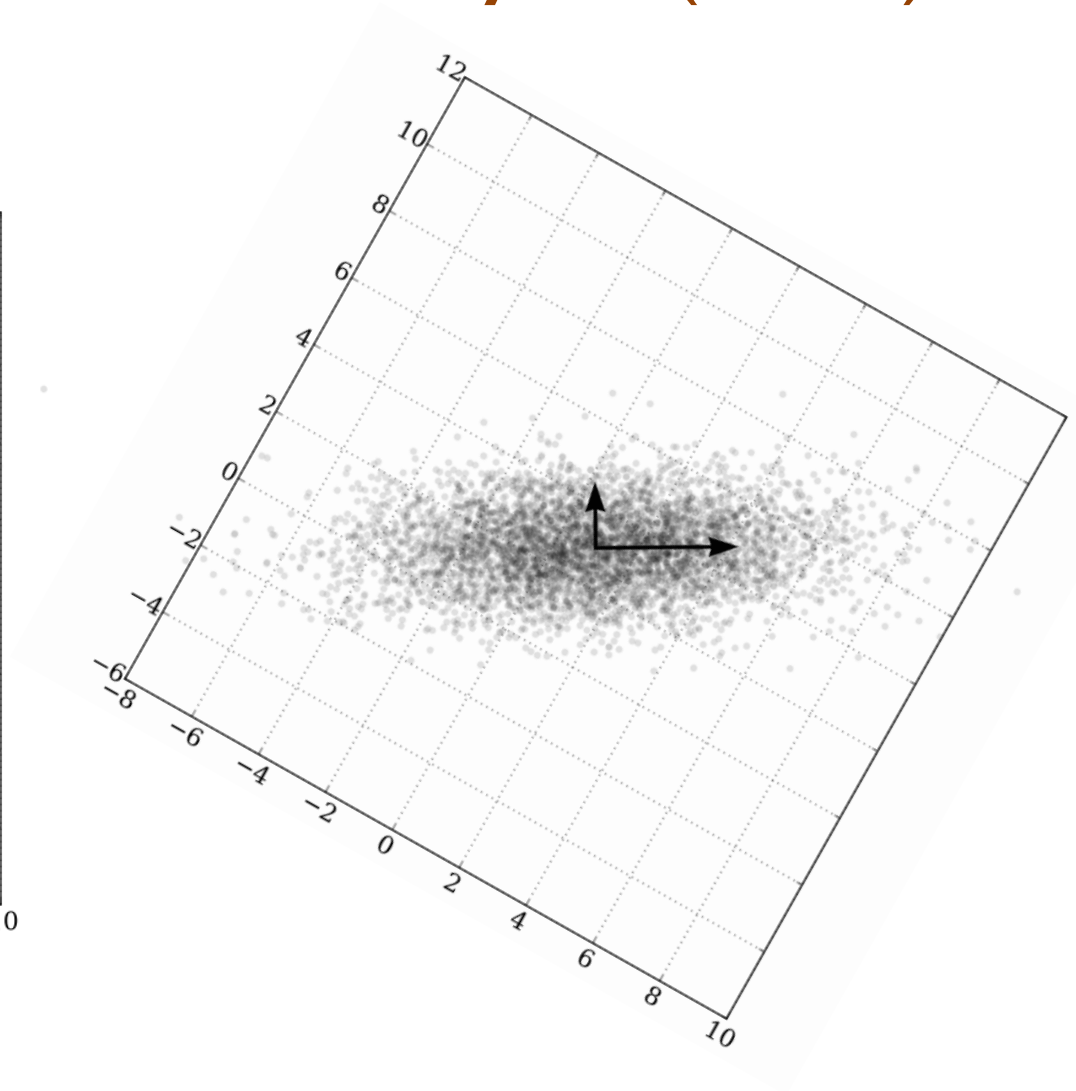
Computational genetics: Computational analysis of microarray data
Quackenbush (2001) *Nature Reviews Genetics*. doi:10.1038/35076576

Hierarchical Clustering



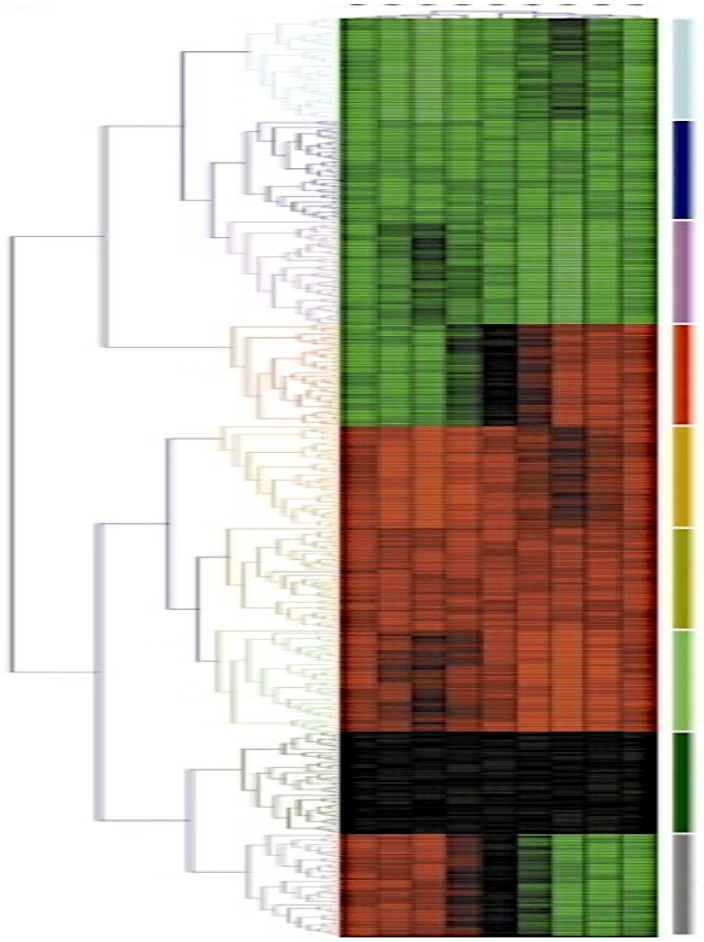
Principle Components Analysis (PCA)

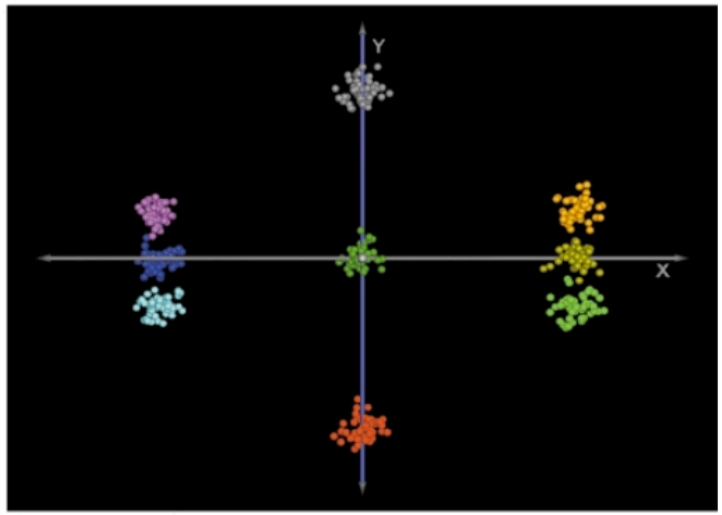




PC1: “New X”- The dimension with the most variability
PC2: “New Y”- The dimension with the second most variability

Principle Components Analysis (PCA)





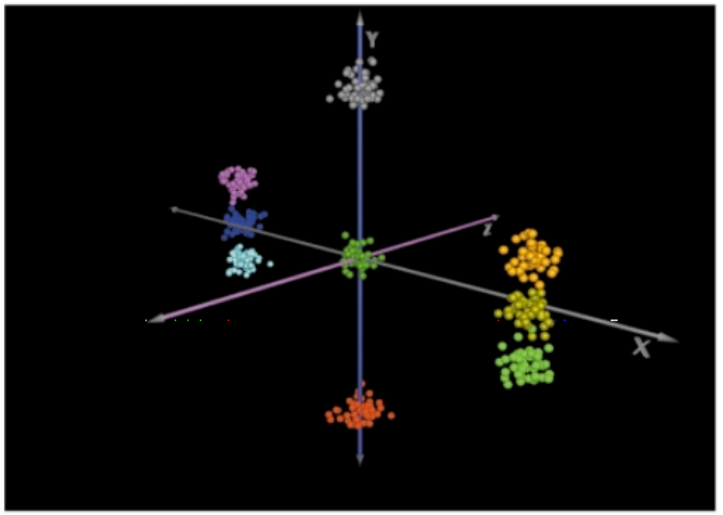
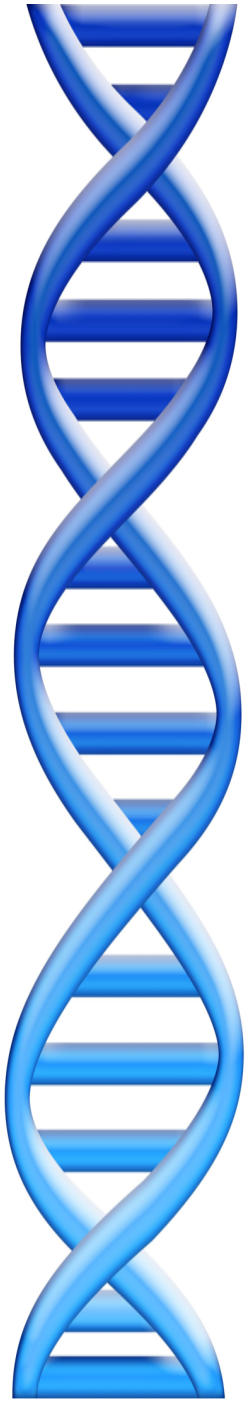


Figure 4 | **Principal component analysis.** The same demonstration data set was analysed using **a** | hierarchical (average-linkage) clustering and **b** | principal component analysis using Euclidean distance, to show how each treats the data, with genes colour coded on the basis of hierarchical clustering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



Part 2:
Modern Humans

ARTICLE

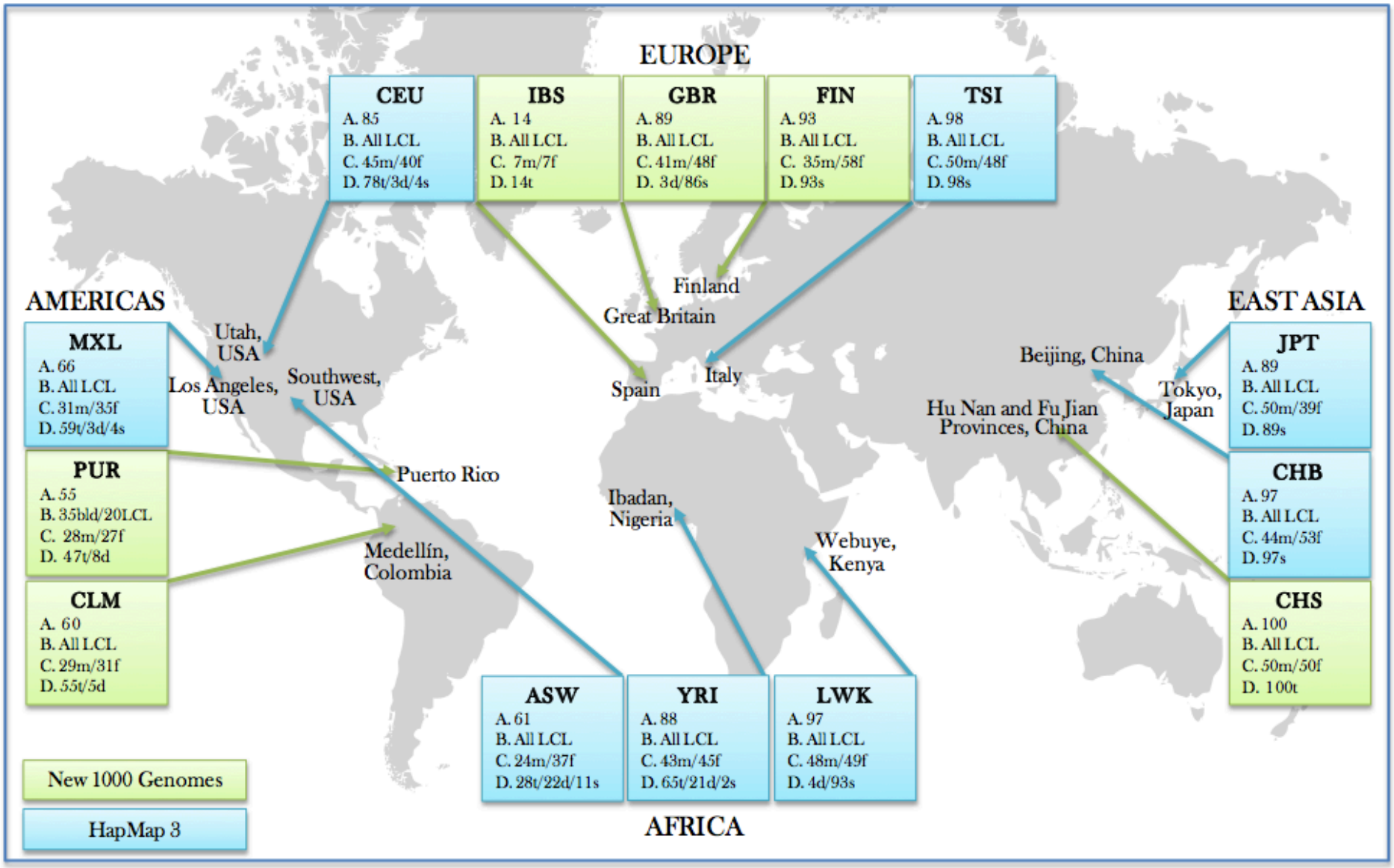
doi:10.1038/nature11632

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

The 1000 Genomes Project Consortium*

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

1000 Genomes Populations



1000 Genomes Populations

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

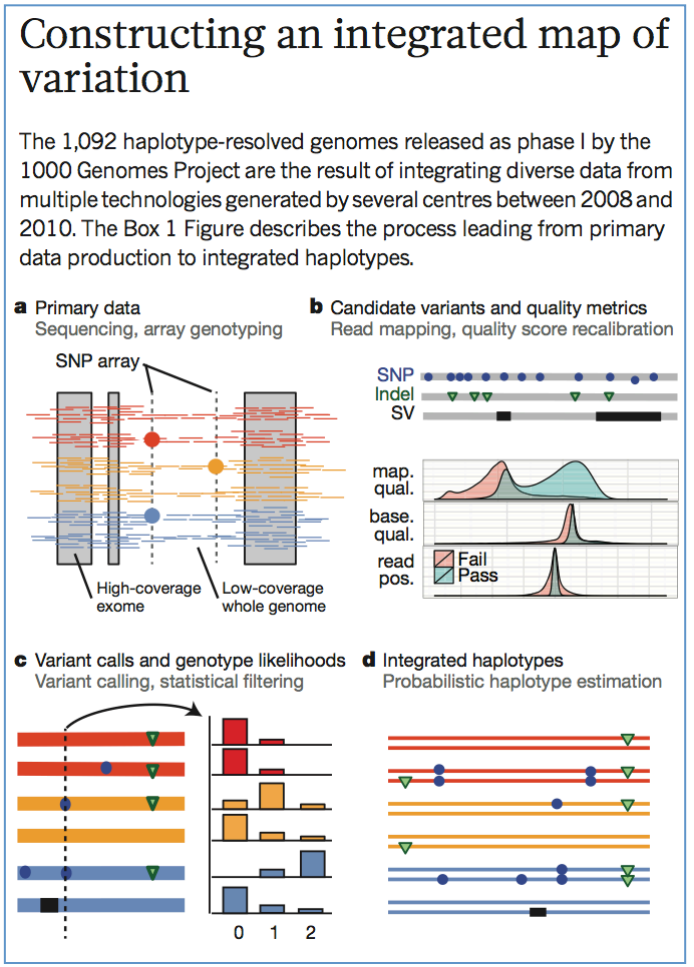
26 populations from 5 major population groups

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



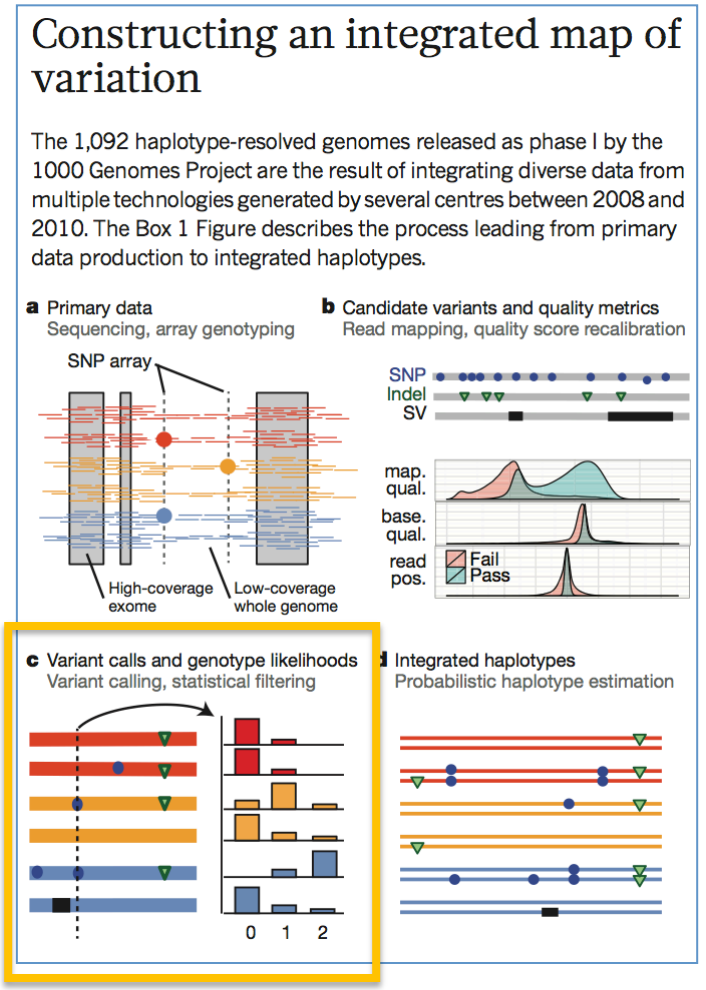
An integrated map of genetic variation from 1,092 human genomes
1000 genomes project (2012) *Nature*. doi:10.1038/nature11632

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



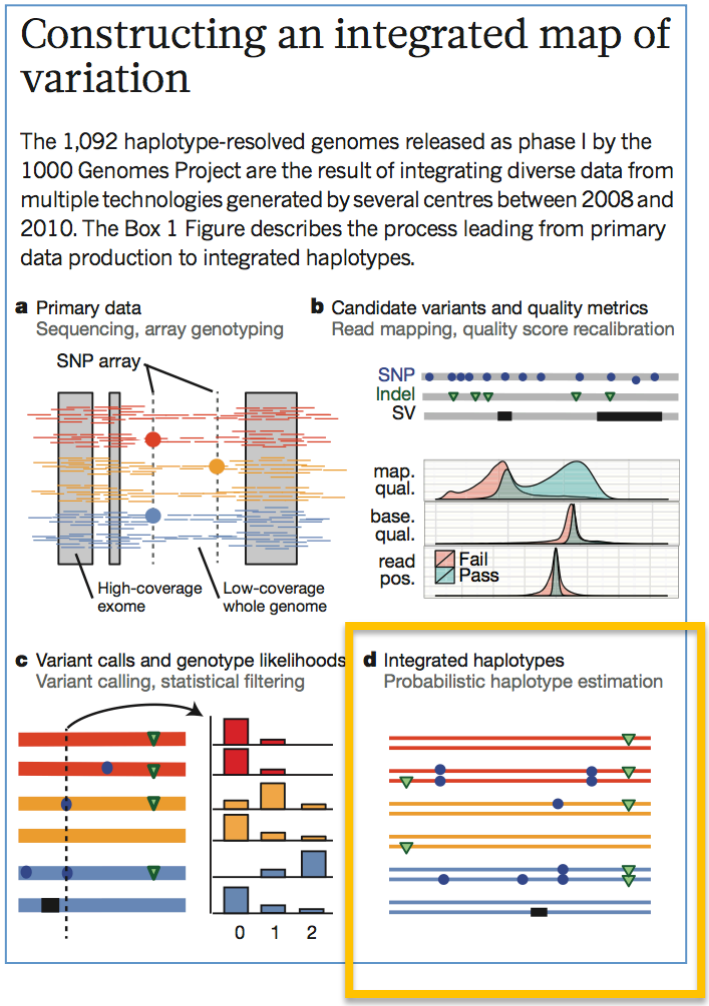
An integrated map of genetic variation from 1,092 human genomes
1000 genomes project (2012) *Nature*. doi:10.1038/nature11632

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

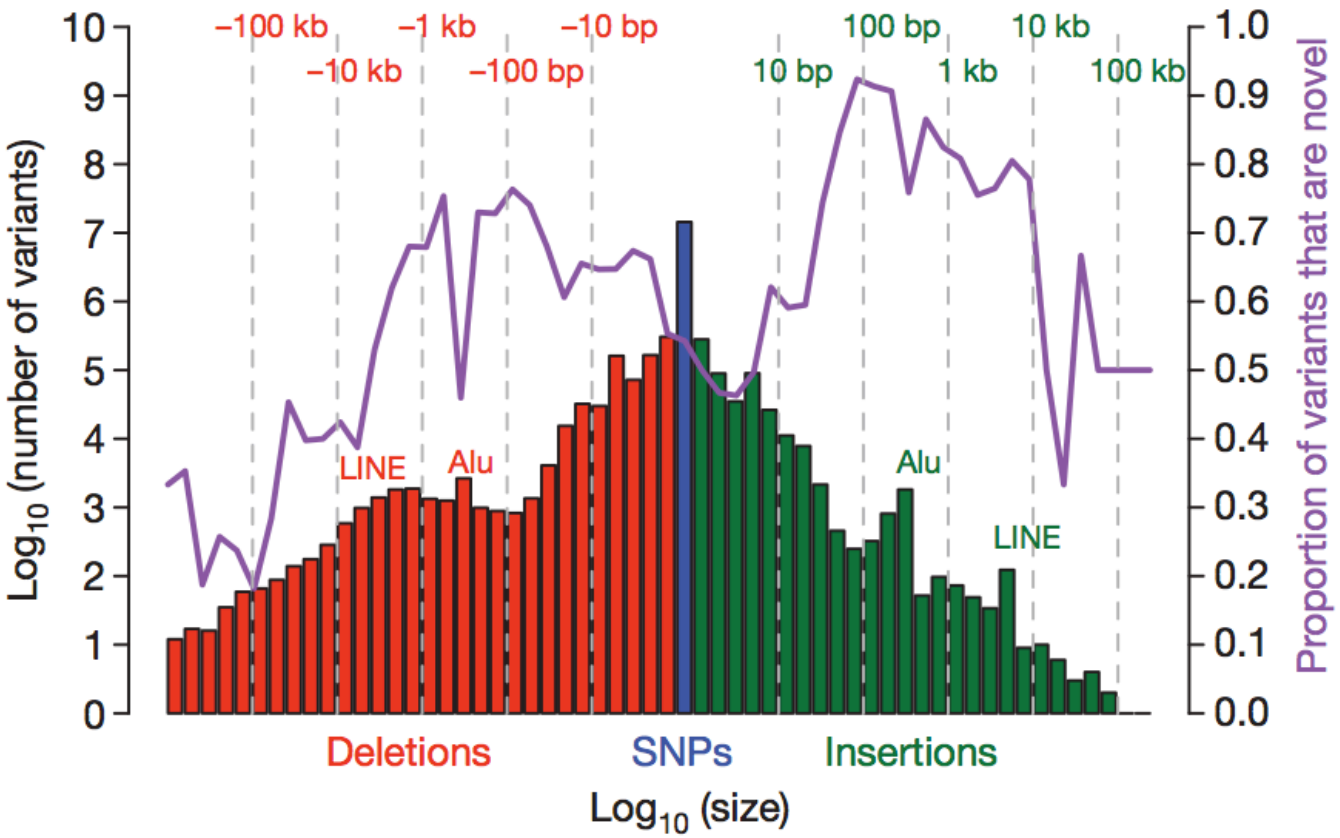
- Overall SNP rate between any two people is ~1/1200bp to ~1/1300
  - ~3M SNPs between me and you (.1%)
  - ~30M SNPs between human to Chimpanzees (1%)

- 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

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

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

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

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

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

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

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

Homozygous LoF Mutations

LETTER

doi:10.1038/nature22034

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

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

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

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

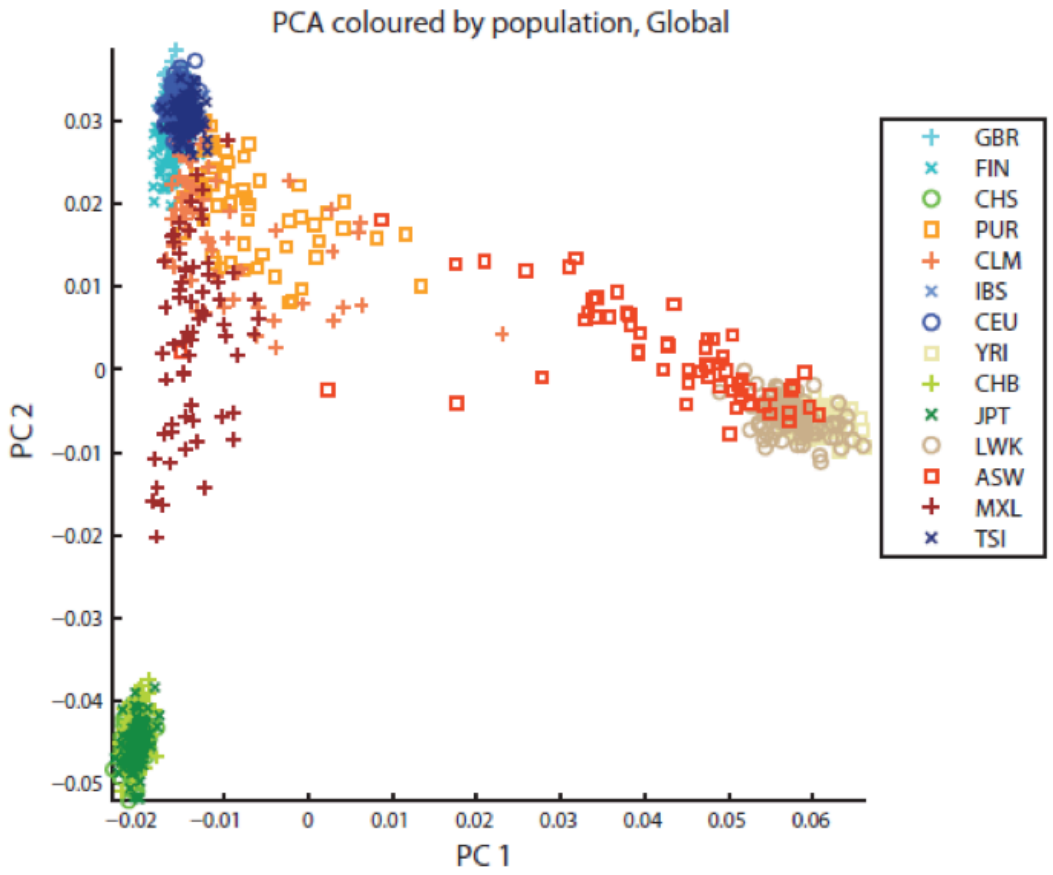
Across all 10,503 PROMIS participants, both copies of 1,317 distinct genes were predicted to be inactivated owing to pLoF mutations. A full listing of all 1,317 genes knocked out, the number of knockout participants for each gene, and the specific pLoF mutation(s) are provided in Supplementary Table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

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

(2017) Nature. doi:10.1038/nature22034

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

Variation across populations

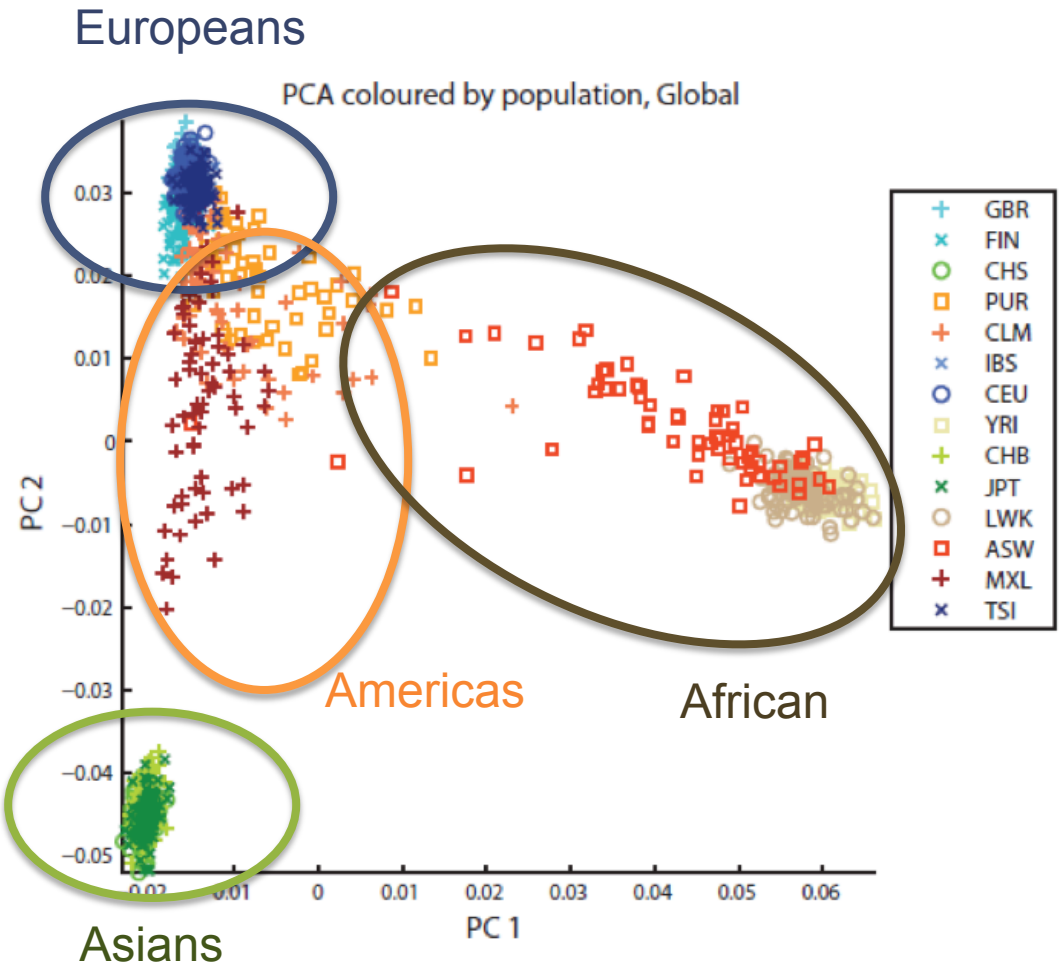


LEVEL	POP_PAIR	# of Highly differentiated SNPs	% in transcribed regions*
AFR	ASW-LWK	258	46.8
AFR	LWK-YRI	251	50.2
AFR	ASW-YRI	213	45.8
ASN	CHS-JPT	275	48.1
ASN	CHB-JPT	176	43.7
ASN	CHB-CHS	79	38.7
EUR	FIN-TSI	343	42.6
EUR	CEU-FIN	201	40.7
EUR	FIN-GBR	197	43.2
EUR	GBR-TSI	100	38.9
EUR	CEU-TSI	57	53.8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

Table S12A Summary of sites showing high levels of population differentiation

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

Variation across populations

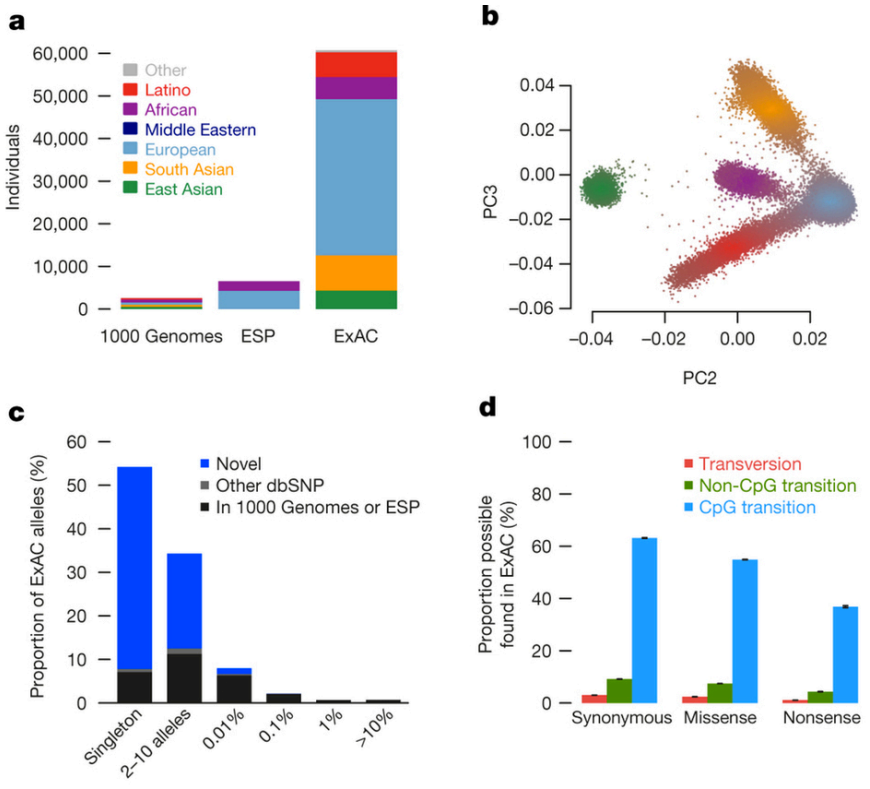


LEVEL	POP_PAIR	# of Highly differentiated SNPs	% in transcribed regions*
AFR	ASW-LWK	258	46.8
AFR	LWK-YRI	251	50.2
AFR	ASW-YRI	213	45.8
ASN	CHS-JPT	275	48.1
ASN	CHB-JPT	176	43.7
ASN	CHB-CHS	79	38.7
EUR	FIN-TSI	343	42.6
EUR	CEU-FIN	201	40.7
EUR	FIN-GBR	197	43.2
EUR	GBR-TSI	100	38.9
EUR	CEU-TSI	57	53.8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

Table S12A Summary of sites showing high levels of population differentiation

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

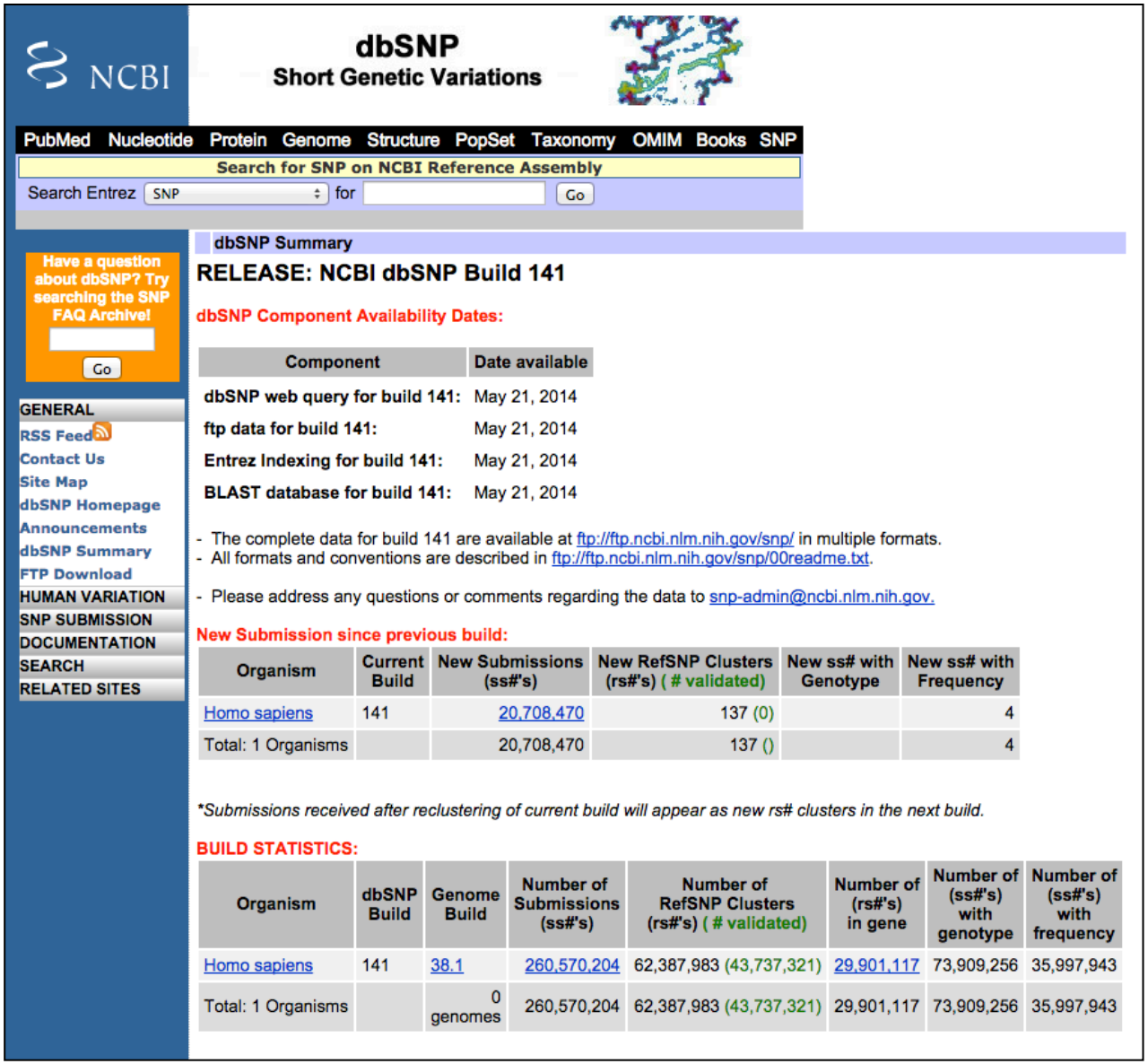
ExAC: Exome Aggregation Consortium



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

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

dbSNP



NCBI

dbSNP
Short Genetic Variations

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books SNP

Search for SNP on NCBI Reference Assembly

Search Entrez SNP for Go

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

Go

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

dbSNP Summary

RELEASE: NCBI dbSNP Build 141

dbSNP Component Availability Dates:

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

- The complete data for build 141 are available at ftp://ftp.ncbi.nlm.nih.gov/snp/ in multiple formats.
- All formats and conventions are described in ftp://ftp.ncbi.nlm.nih.gov/snp/00readme.txt.

- Please address any questions or comments regarding the data to snp-admin@ncbi.nlm.nih.gov.

New Submission since previous build:

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

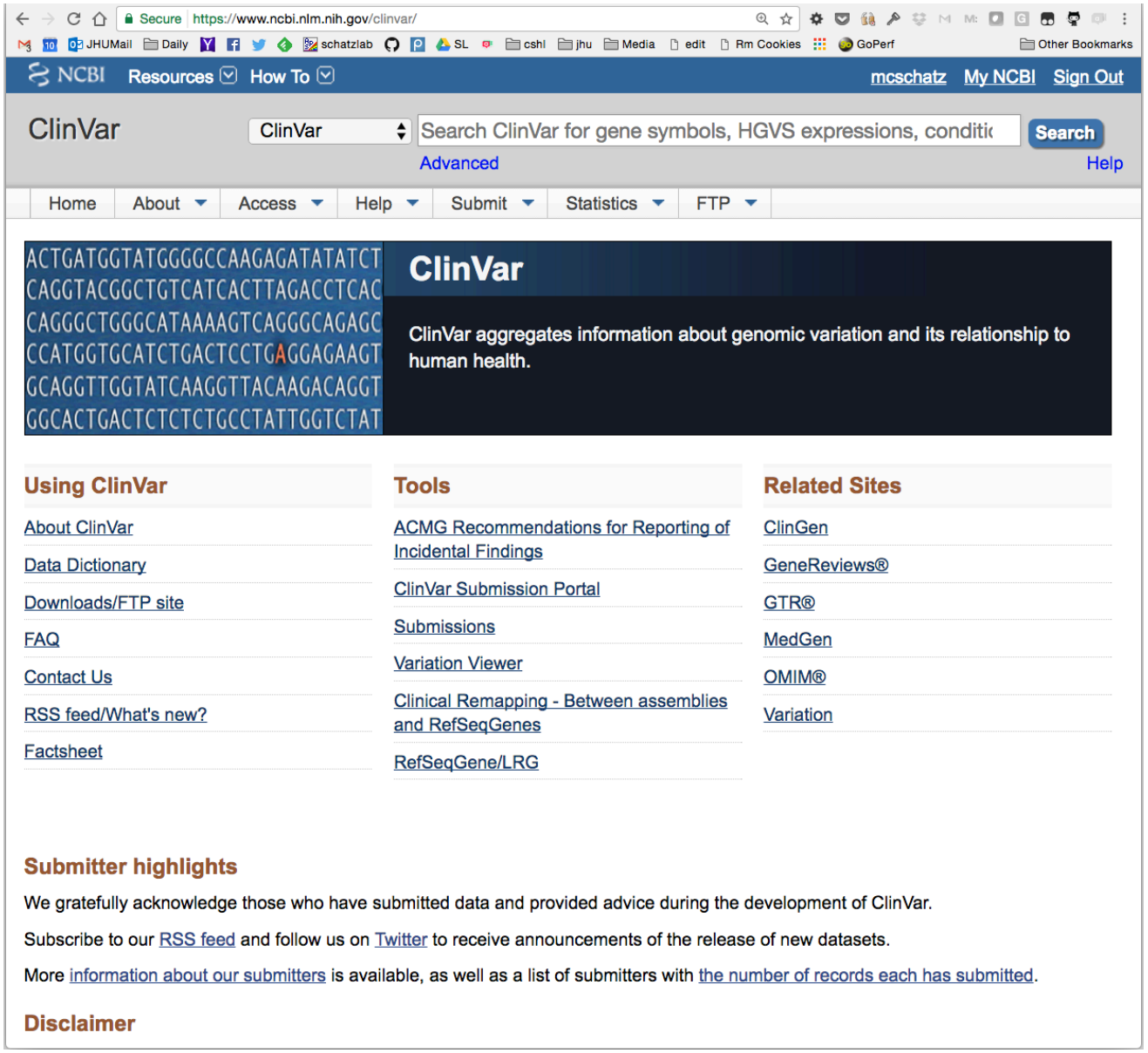
**Submissions received after reclustering of current build will appear as new rs# clusters in the next build.*

BUILD STATISTICS:

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

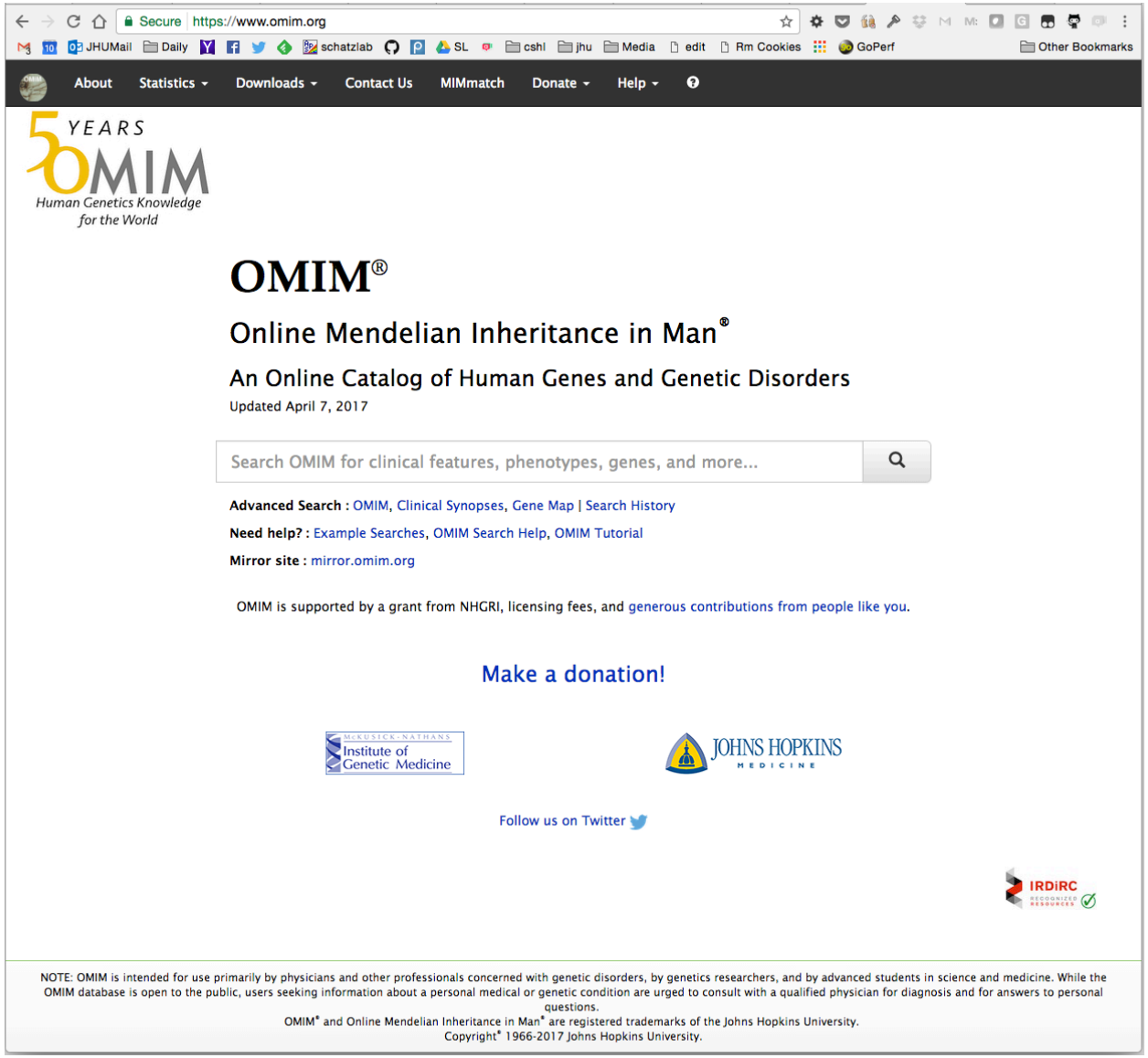
- Periodic release of databases of known variants and their population frequencies
- Generally assumed to be non-disease related
- However, as catalog grows, almost certainly to contain some medically relevant SNPs.

ClinVar

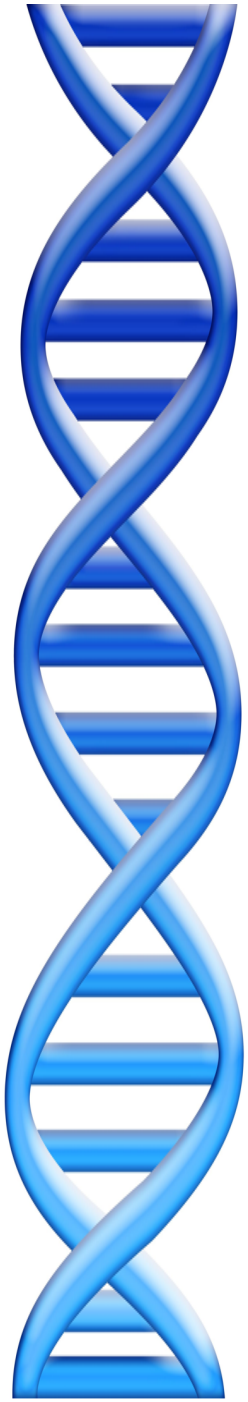


- ClinVar is a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence
- Currently has 295k mutations
- Most (179k) variants have uncertain affect, only 23 have “4 stars” of signifance

OMIM

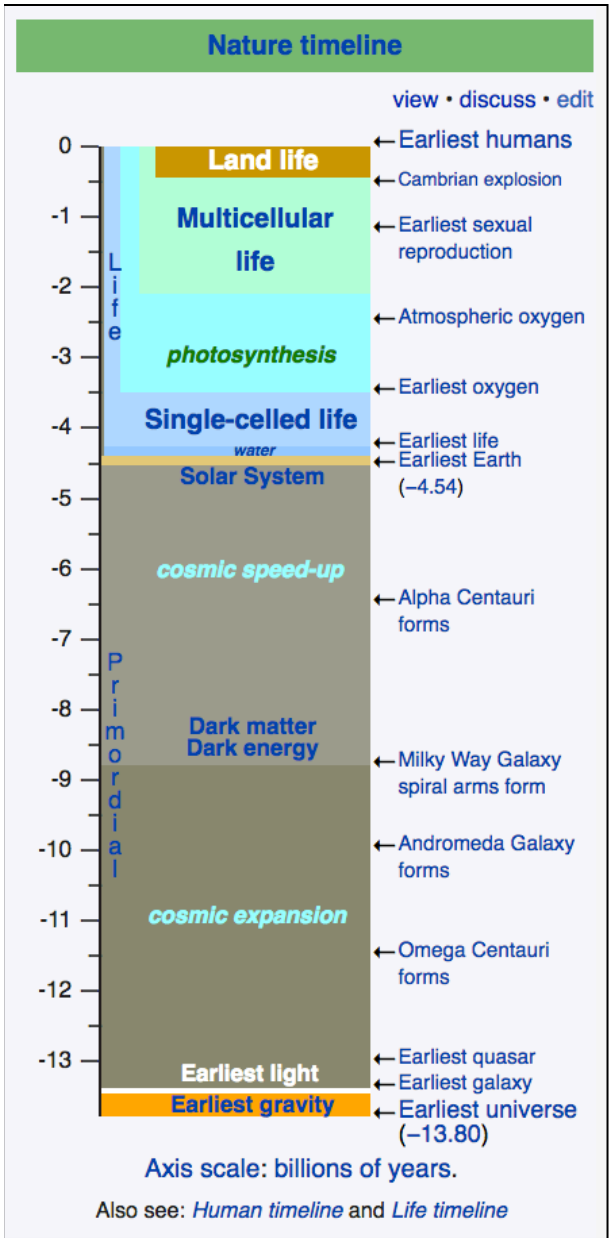


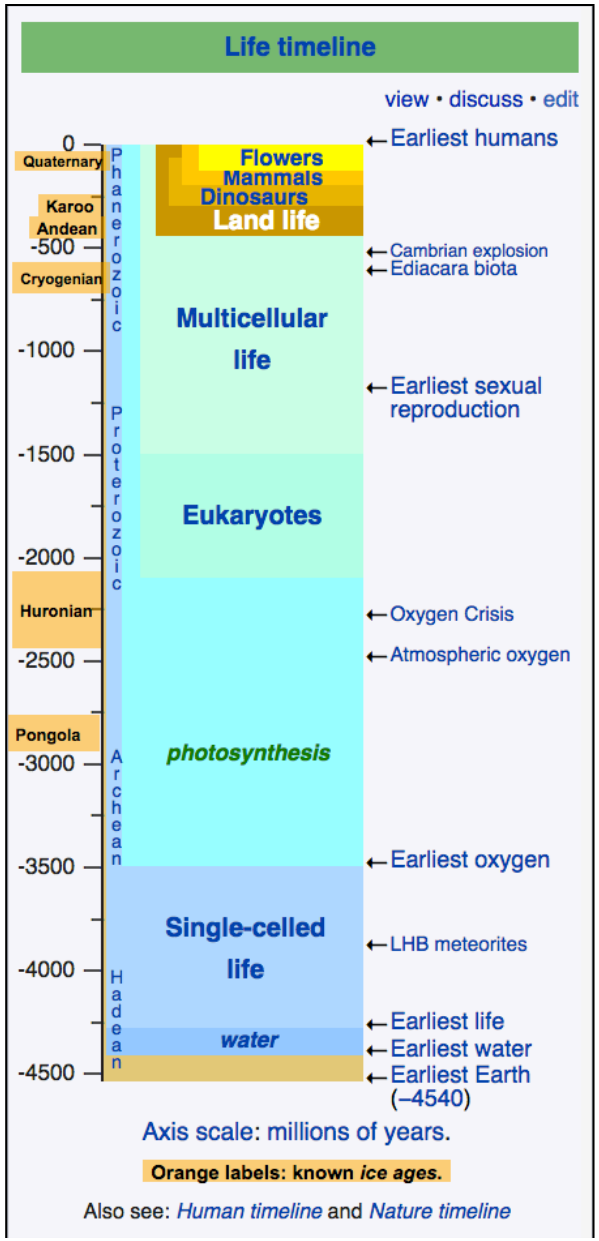
- For many different diseases and phenotypes, lists what are all of the known genetic associations
- Has records for nearly all genes, ~5k different conditions with known molecular basis, ~1k with unknown basis, ~1k with questionable basis
- Started at JHU 50 years ago ☺

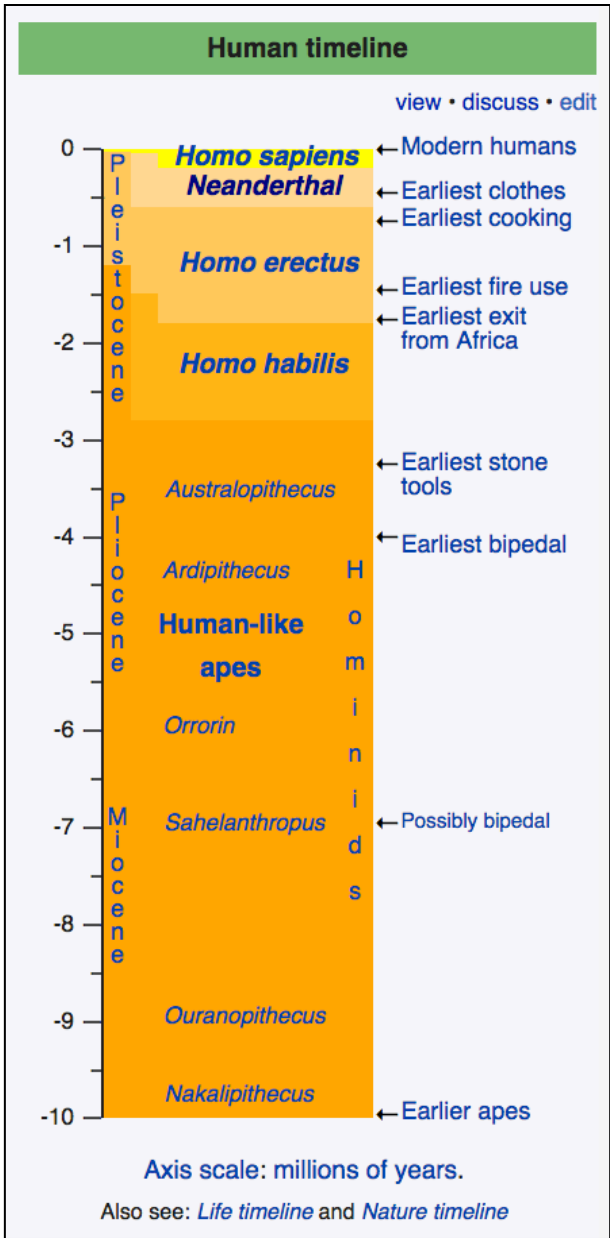


Part 3:
Ancient Hominds

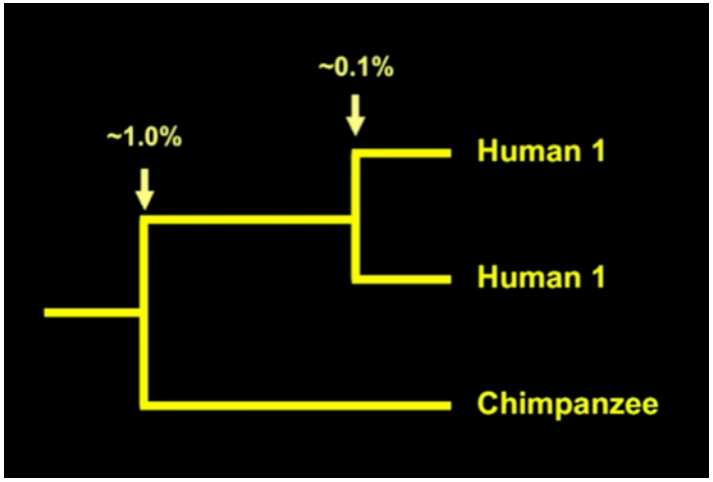
Our Origins

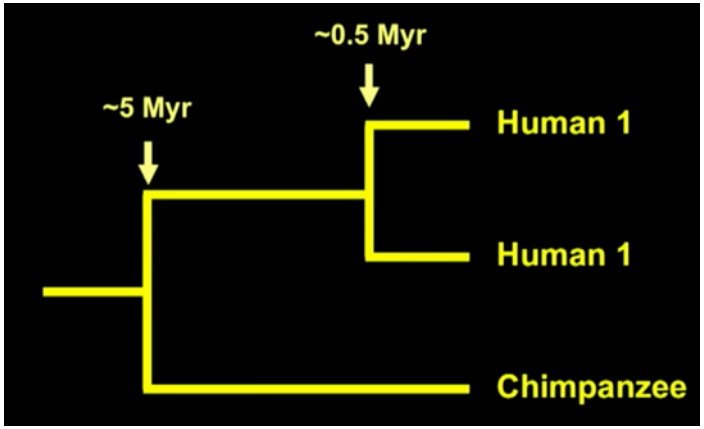

Nature timeline
view • discuss • edit
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
-10
-11
-12
-13
Life
Primordial
Land life
Multicellular life
photosynthesis
Single-celled life
water
Solar System
cosmic speed-up
Dark matter
Dark energy
cosmic expansion
Earliest light
Earliest gravity
Earliest humans
Cambrian explosion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


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


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

Mutation Rates and Evolutionary Time





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

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

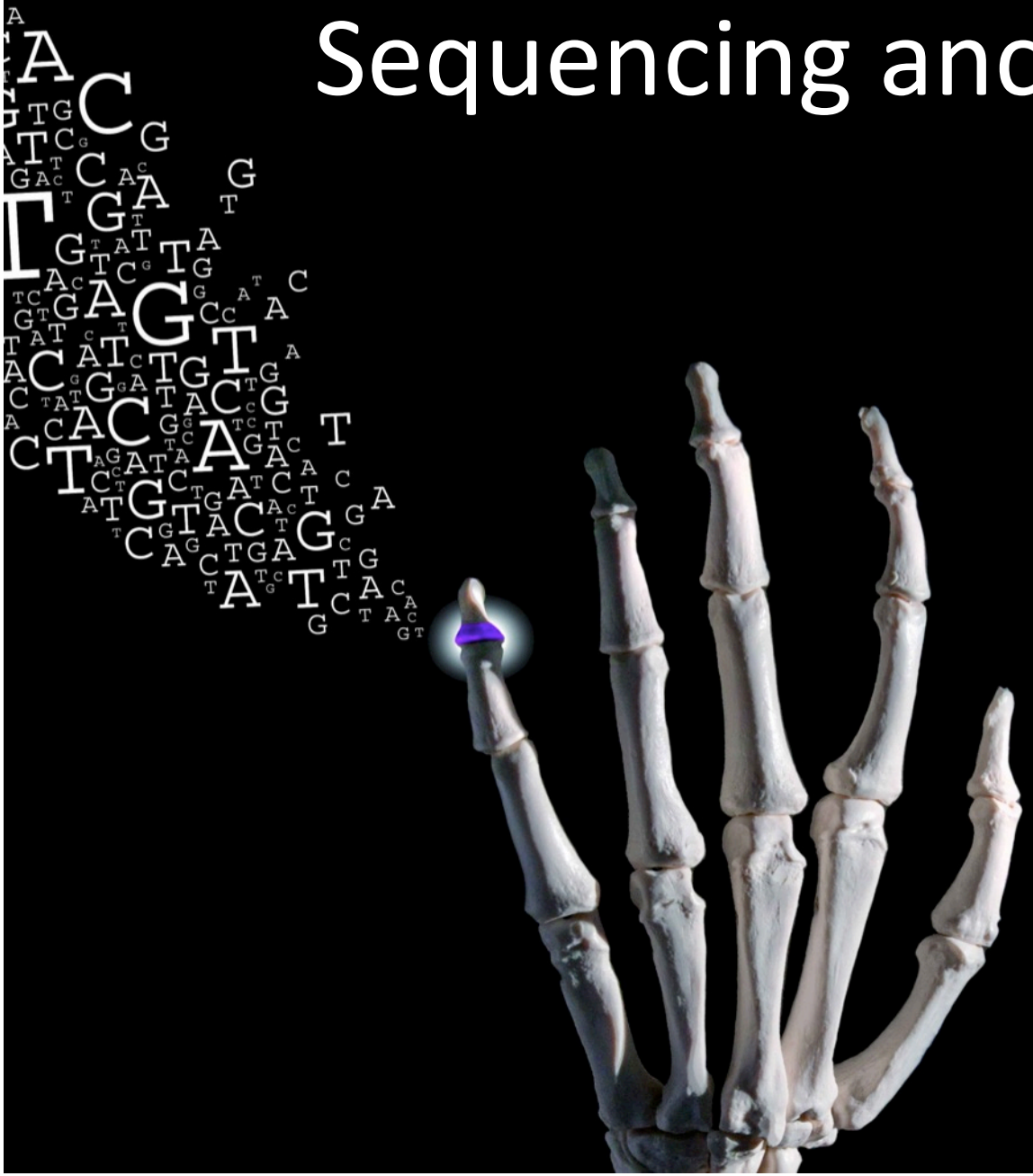
We see "African" alleles all around the world
- Zero SNPs occur exclusively in Africa
- Only 12 SNPs across the entire genome 'unique' to Africa (allowing 95% tolerance)
- We are all African (either currently living in Africa or recent exiles)!

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

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

Sequencing ancient genomes

Janet Kelso
Max-Planck Institute



Homo neanderthalensis

•Proto-Neanderthals emerge around 600k years ago

•“True” Neanderthals emerge around 200k years ago

•Died out approximately 40,000 years ago

•Known for their robust physique

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

Homo sapiens sapiens

•Apparently emerged from earlier hominids in Africa around 50k years ago

•Capable of amazing intellectual and social behaviors

•Mostly Harmless ☺


Engis
Neandertal
Spy
Arcy-sur-Cure
La Quina
Saint-Césaire
Le Moustier
La Ferrassie
Steinheim
Châtelperron
La Chapelle-aux-Saints
Moula
Külna
Sipka
Tata
Erd
Krapina
Vindija
Sukhaya Mechetka
Molodovo
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)

Science
AAAS

A Draft Sequence of the Neandertal Genome

Richard E. Green, *et al.*

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

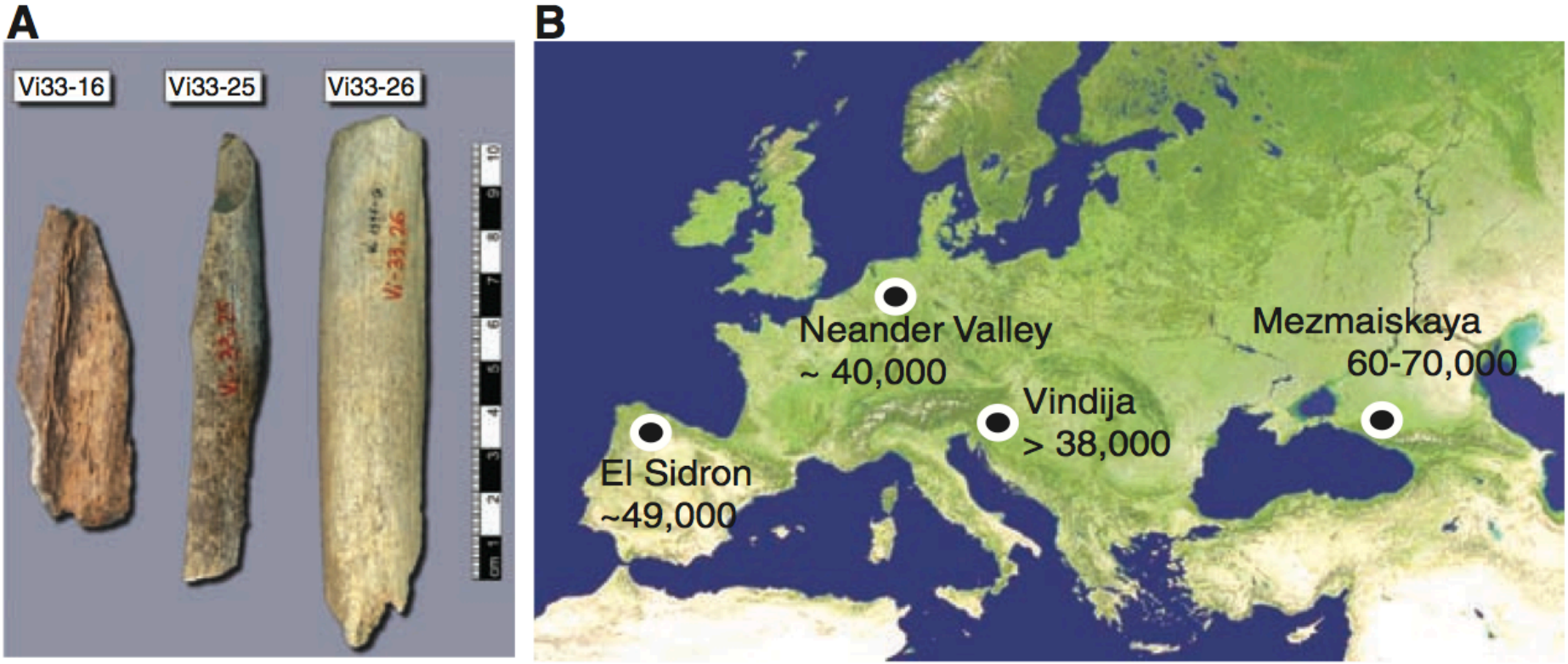
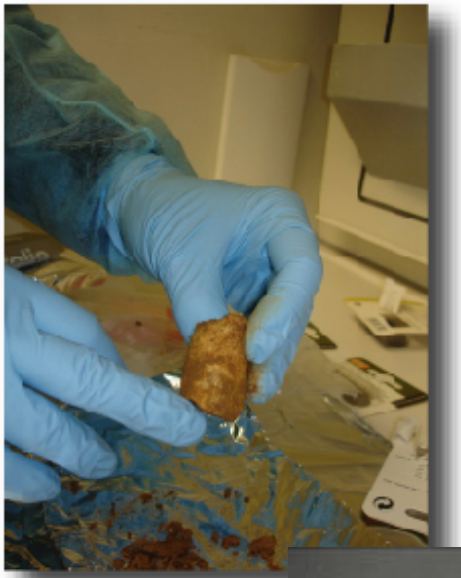
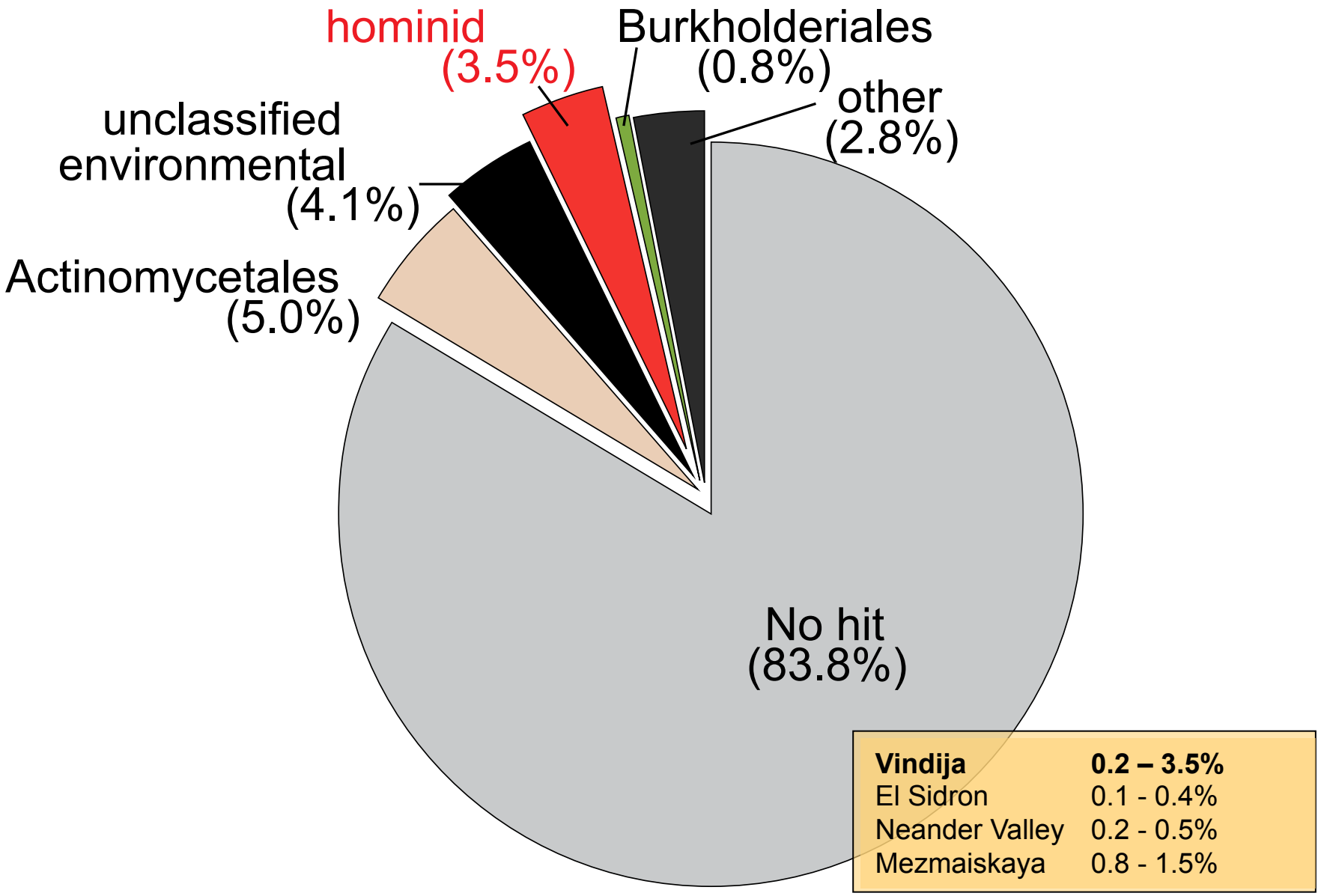


Fig. 1. Samples and sites from which DNA was retrieved. (**A**) The three bones from Vindija from which Neandertal DNA was sequenced. (**B**) Map showing the four archaeological sites from which bones were used and their approximate dates (years B.P.).

Extracting Ancient DNA

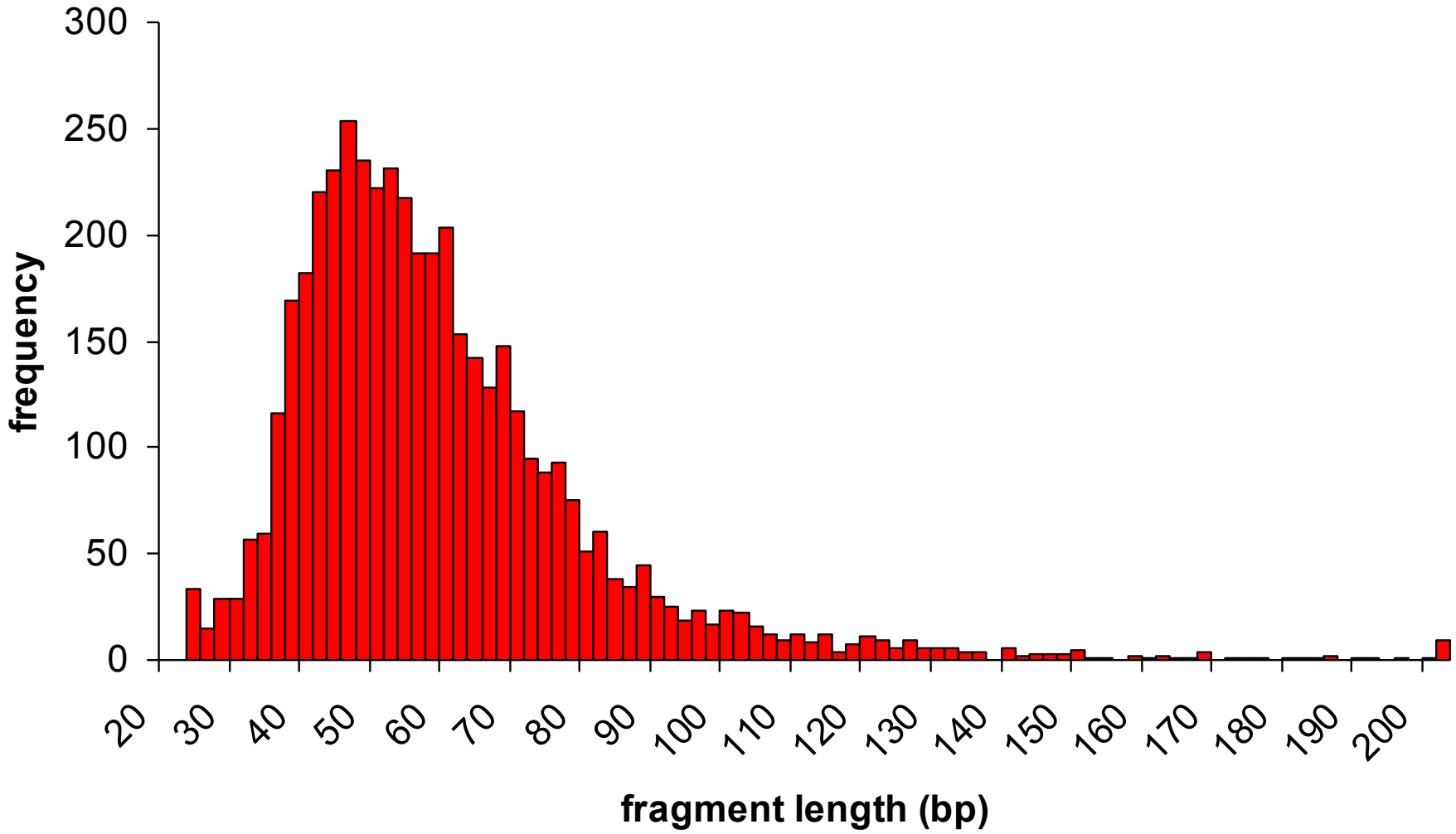


DNA is from mixed sources

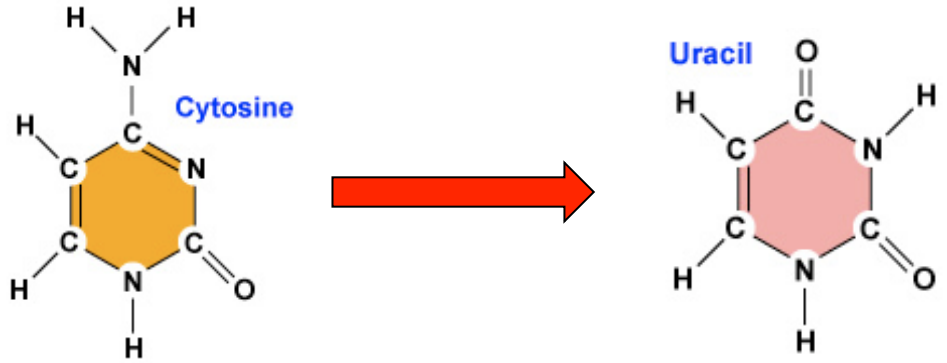


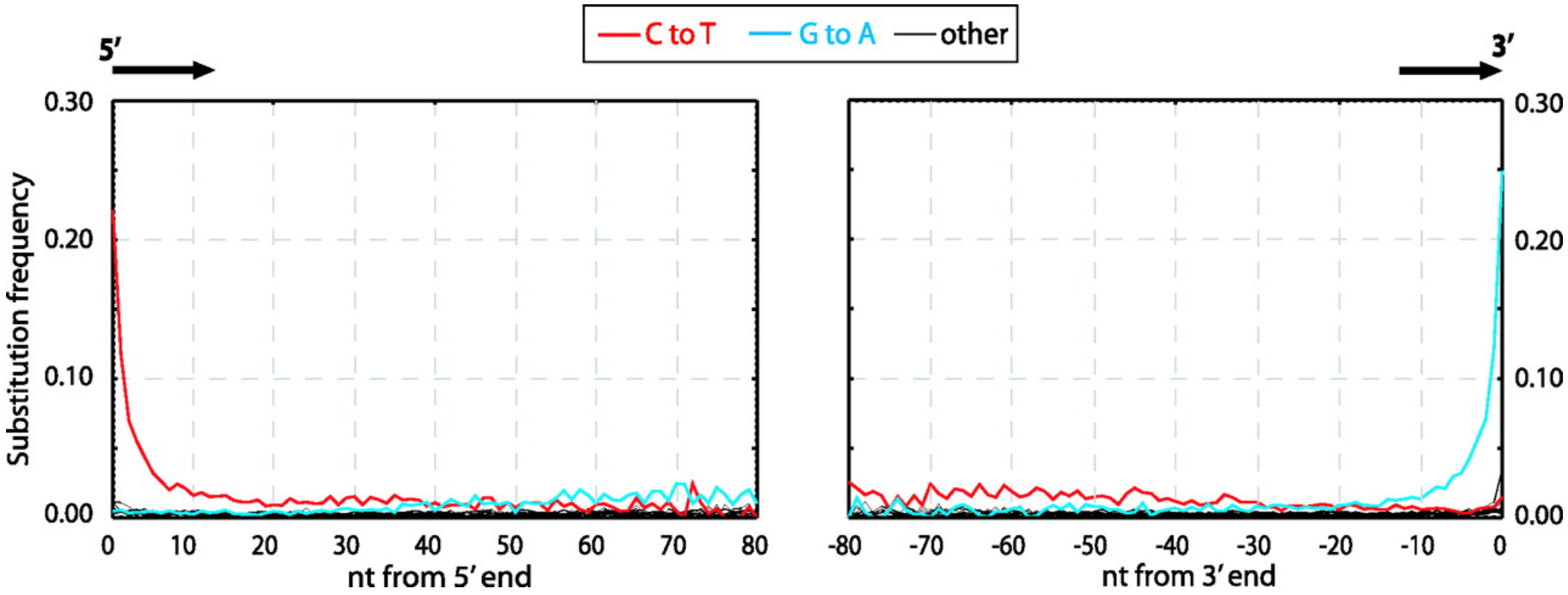
Vindija	**0.2 – 3.5%**
El Sidron	0.1 - 0.4%
Neander Valley	0.2 - 0.5%
Mezmaiskaya	0.8 - 1.5%

DNA is degraded



DNA is chemically damaged





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

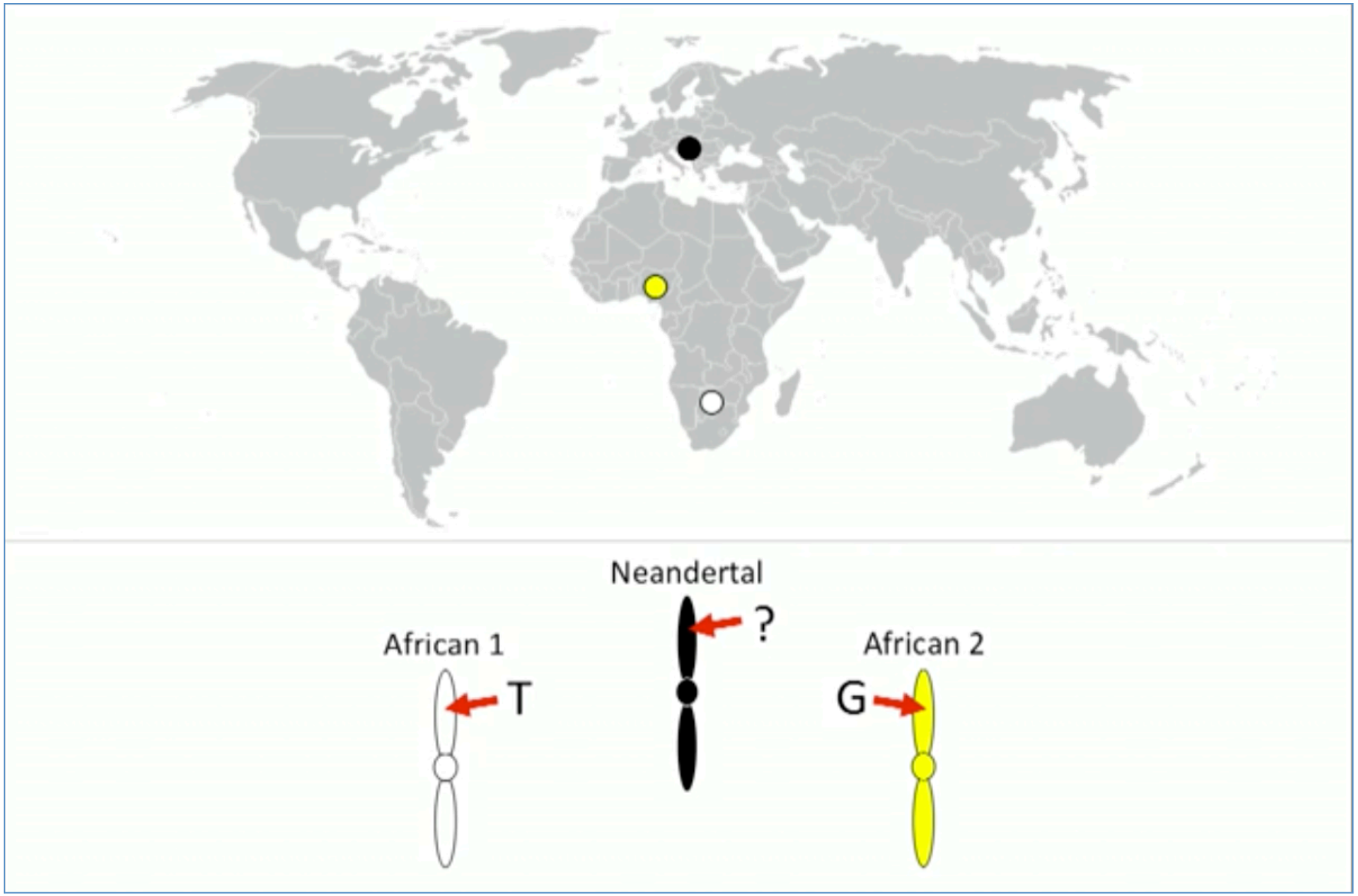
Green et al. 2010

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

~35 Illumina flow cells

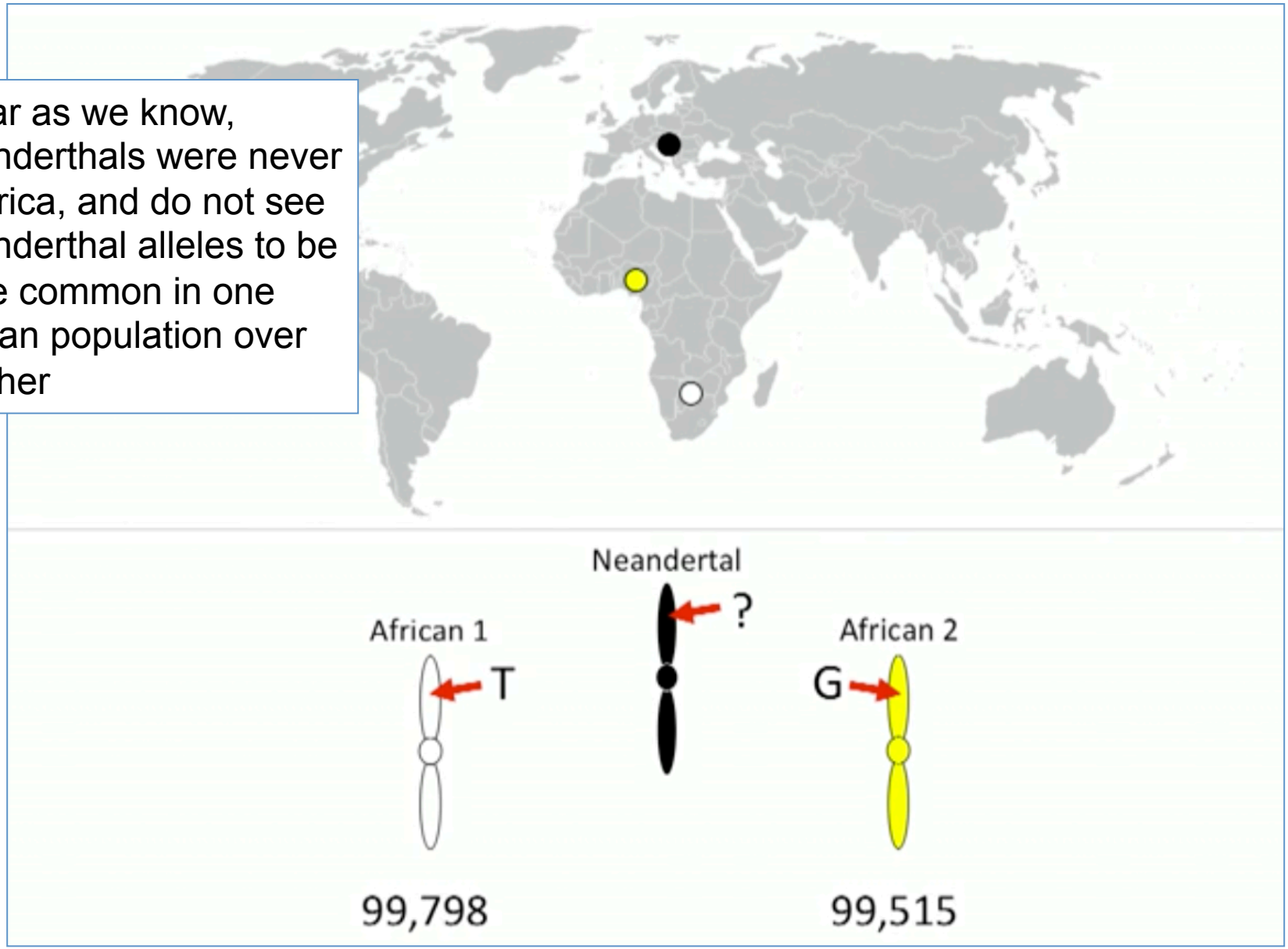
Genome coverage ~1.3 X

Did we mix?



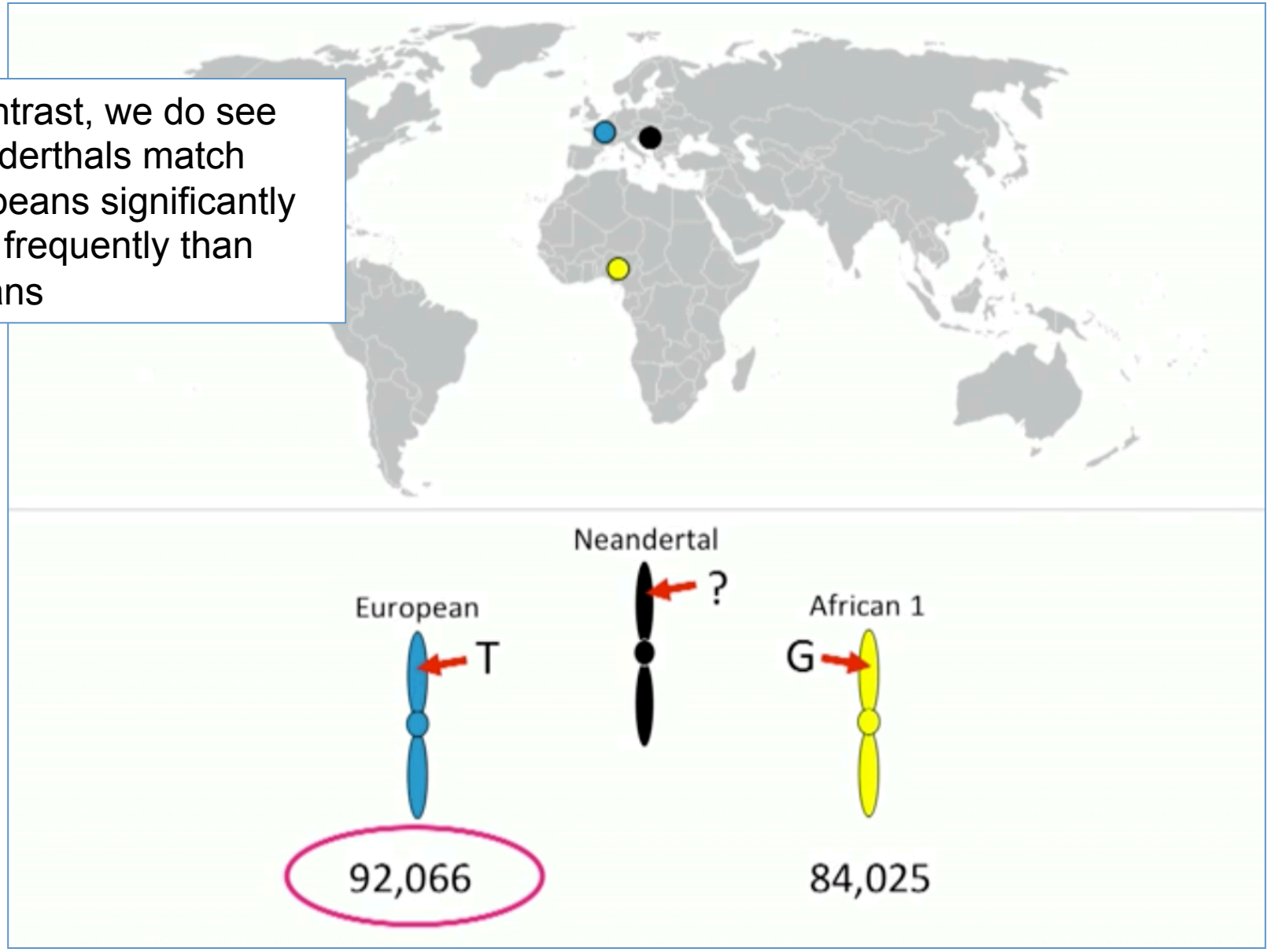
Did we mix?

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



Did we mix?

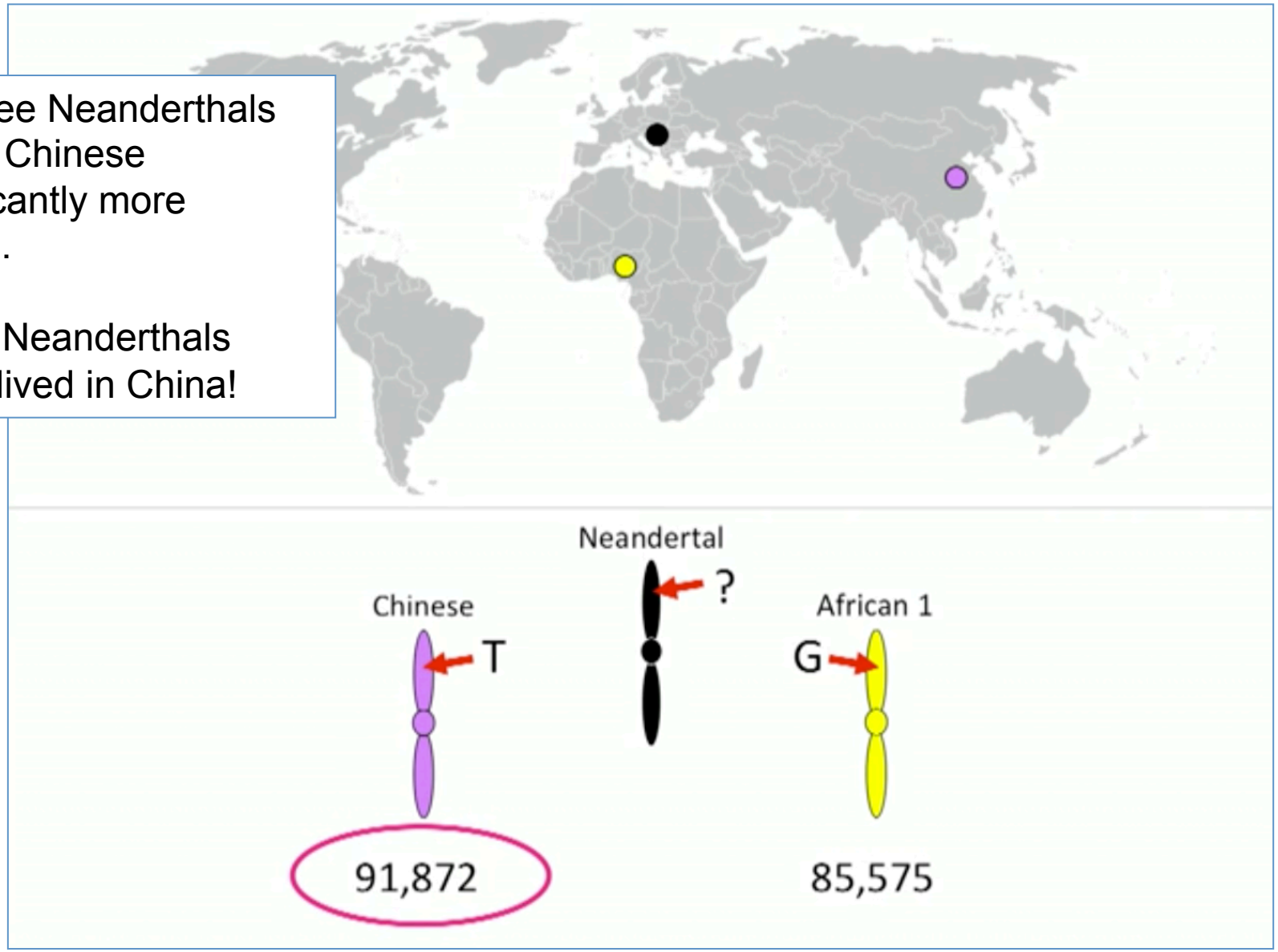
In contrast, we do see Neanderthals match Europeans significantly more frequently than Africans



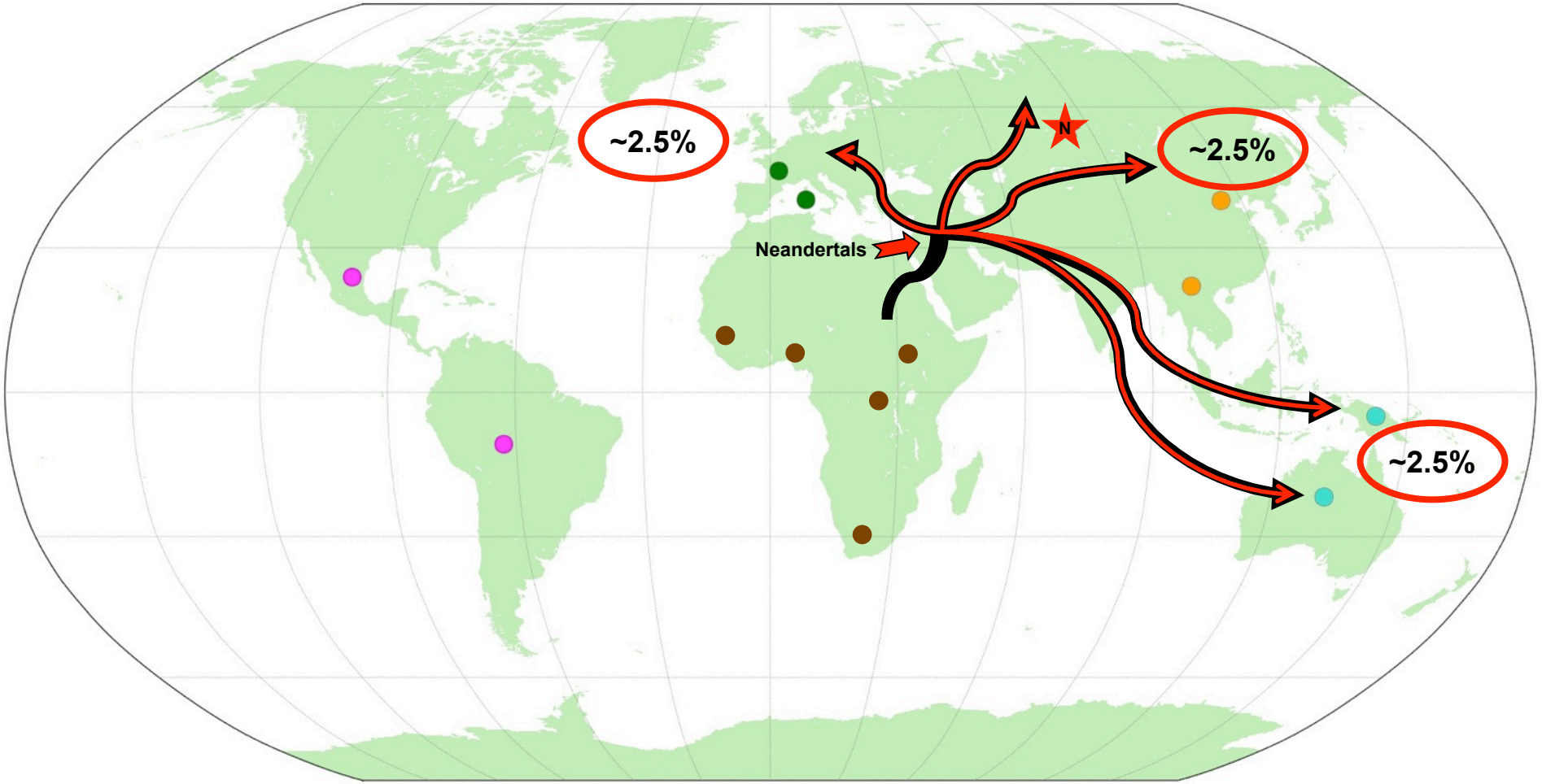
Did we mix?

Also see Neanderthals match Chinese significantly more often…

… but Neanderthals never lived in China!



Neanderthal Interbreeding

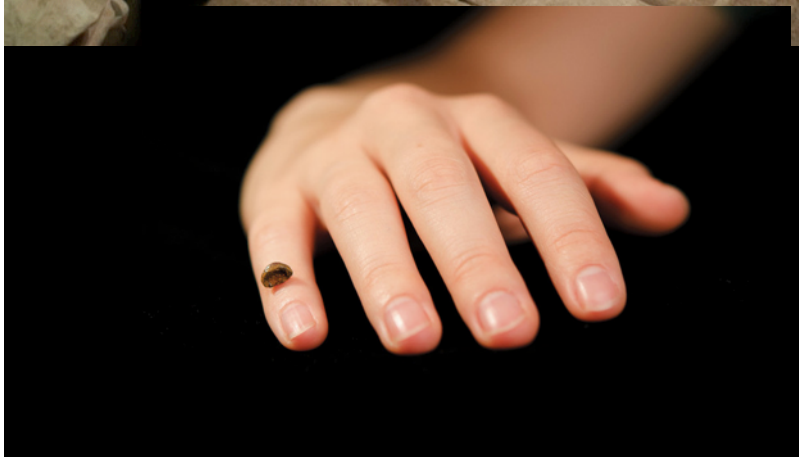


As modern humans migrated out of Africa, they apparently interbred with Neanderthal's so we see their alleles across the rest of the world and carry about 2.5% of their genome with us!

What about other ancient hominids?

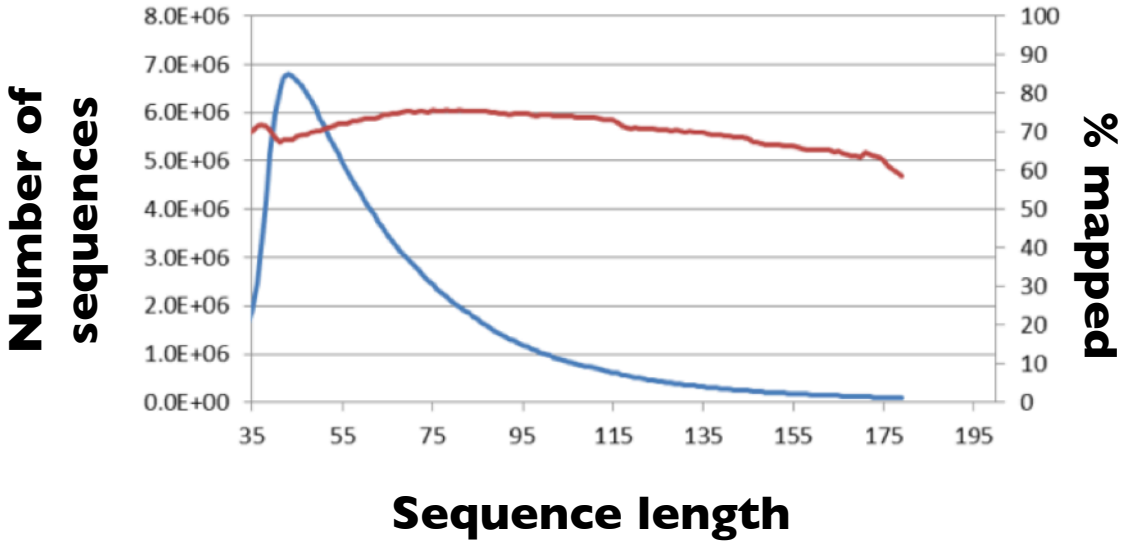
Denisova cave
Altai mountains
Russia

Academician A.P. Derevianko



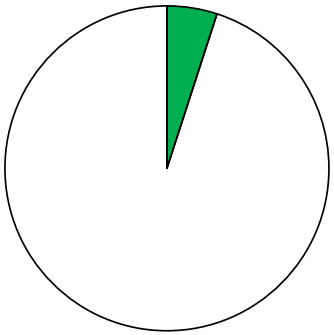

UNITED STATES OF AMERICA
E PLURIBUS UNUM
ONE CENT

Extraordinary preservation

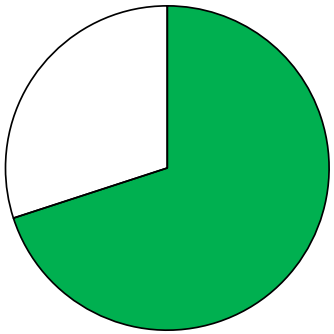


5% endogenous DNA

Best Neandertal bone

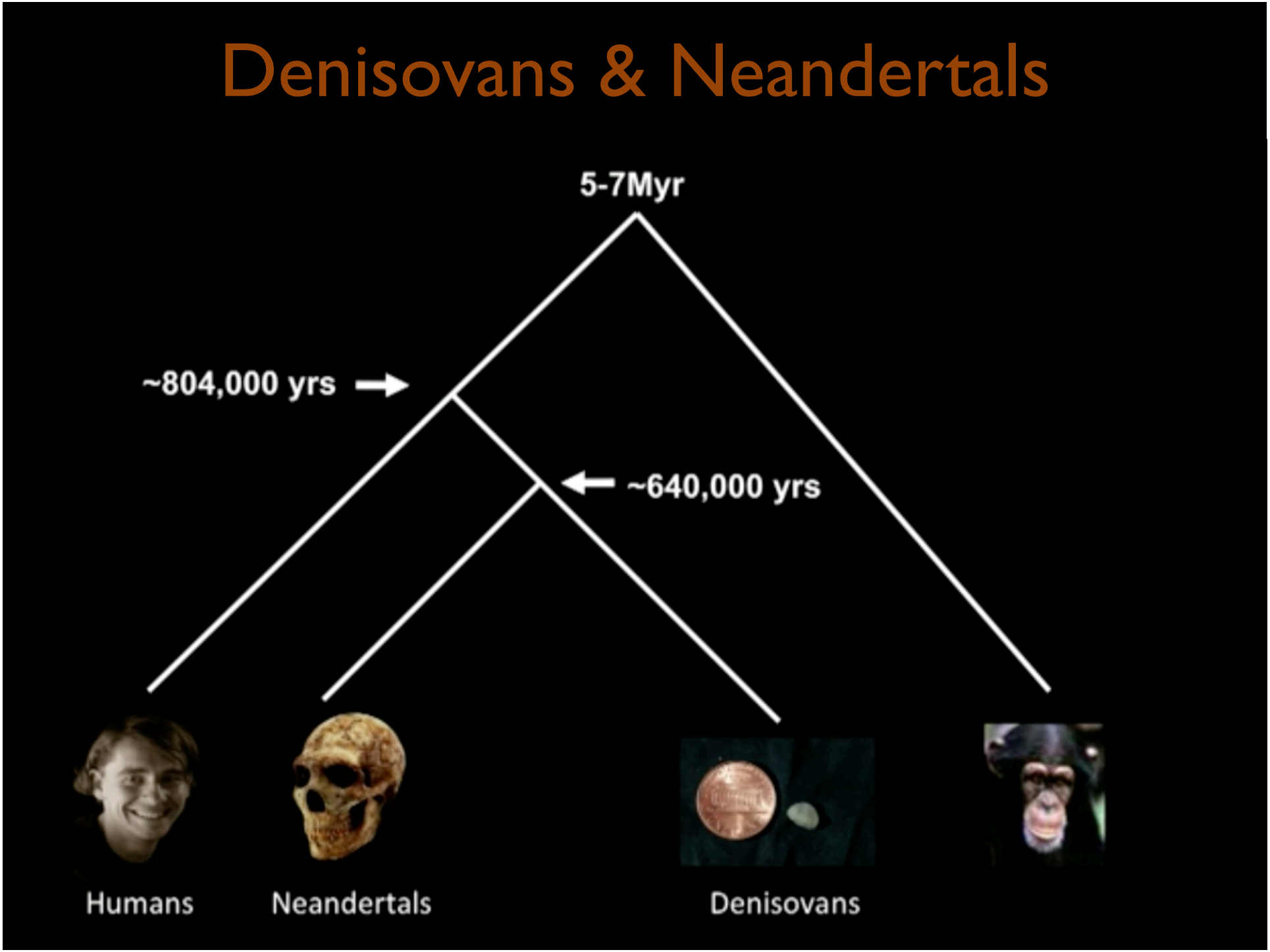


>70% endogenous DNA



Denisova bone

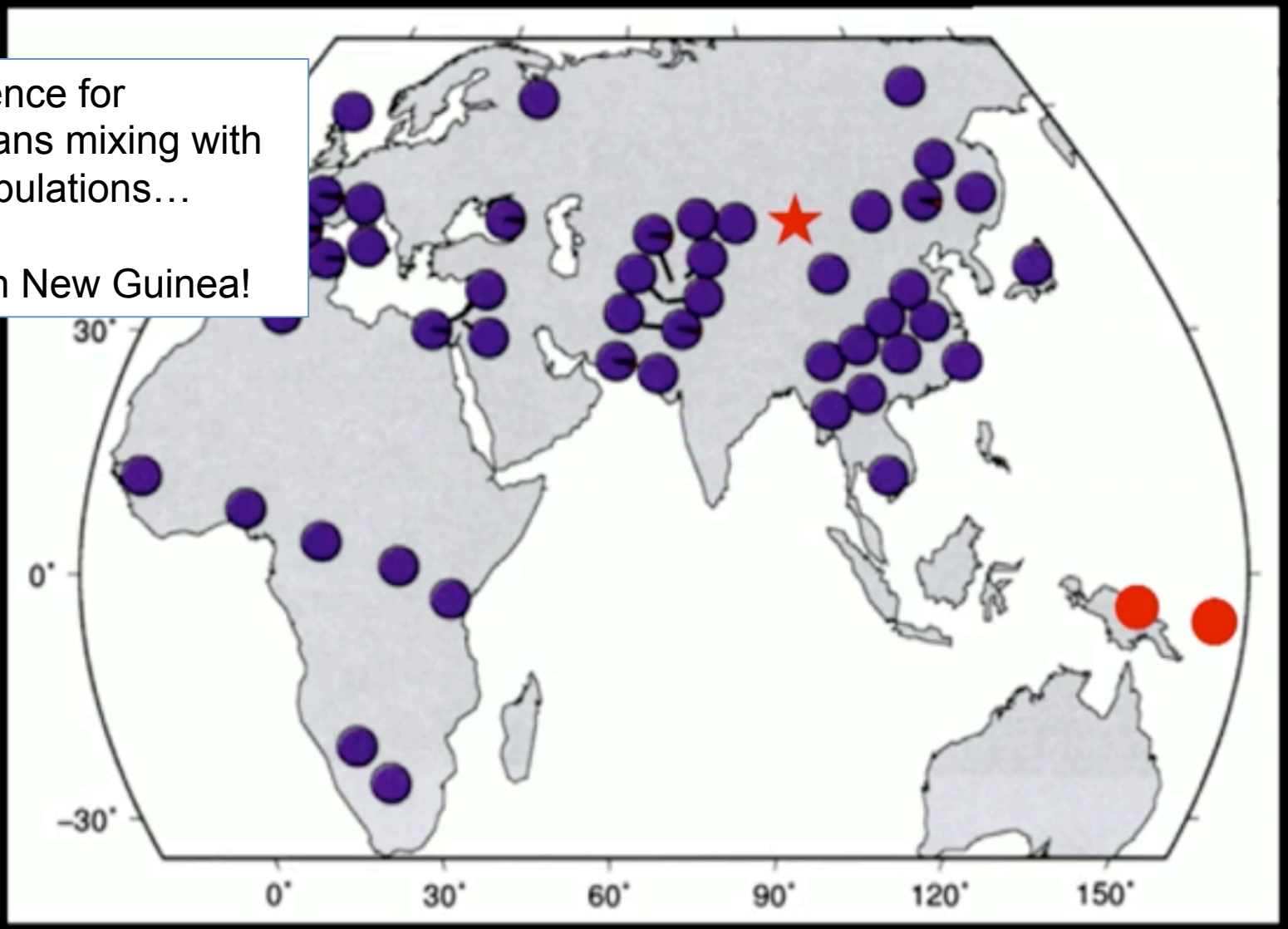
Denisovans & Neandertals



Did we mix?

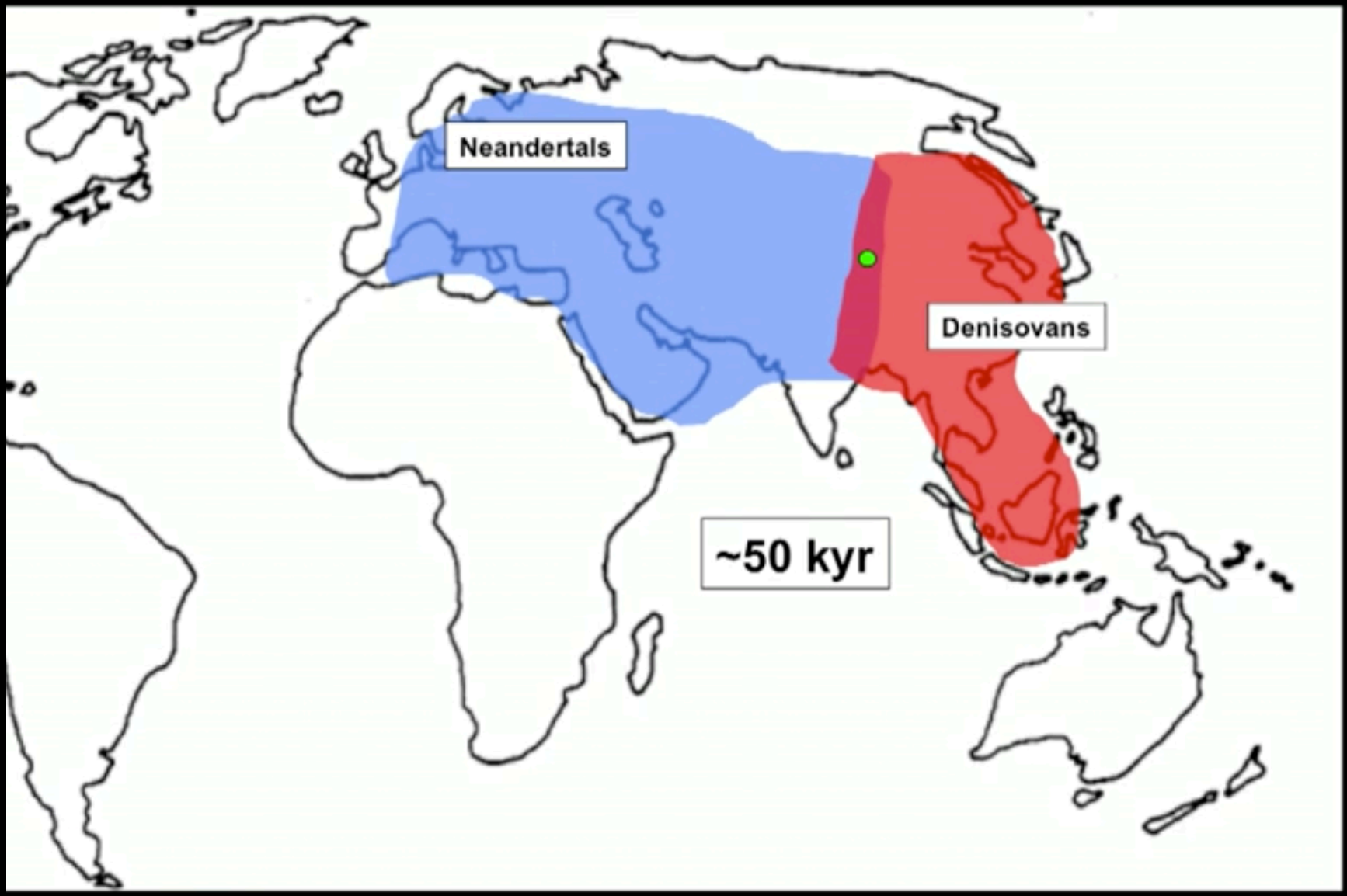
No evidence for Denisovans mixing with other populations…

Except in New Guinea!

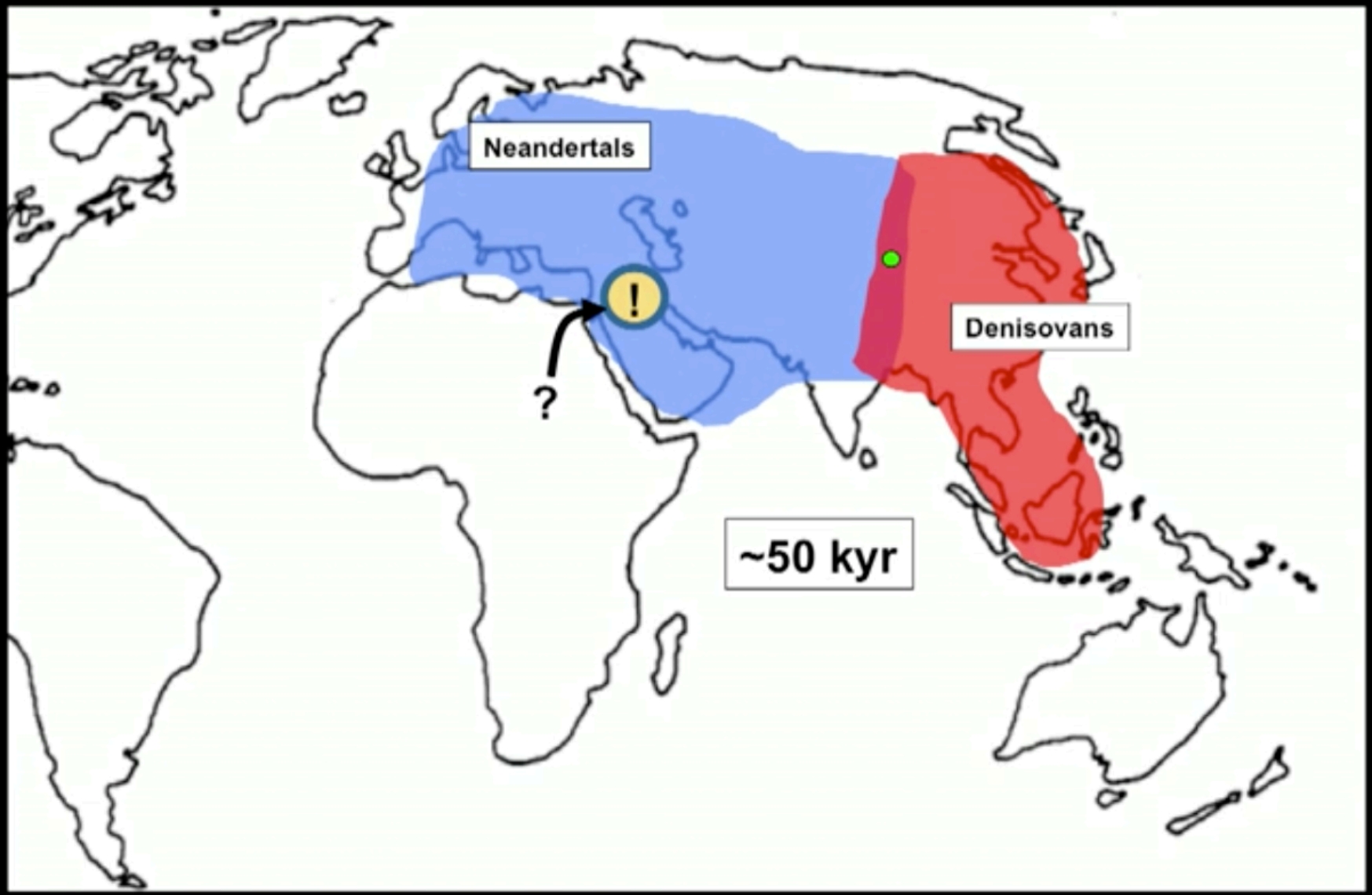


Map after Pickrell et al., 2009

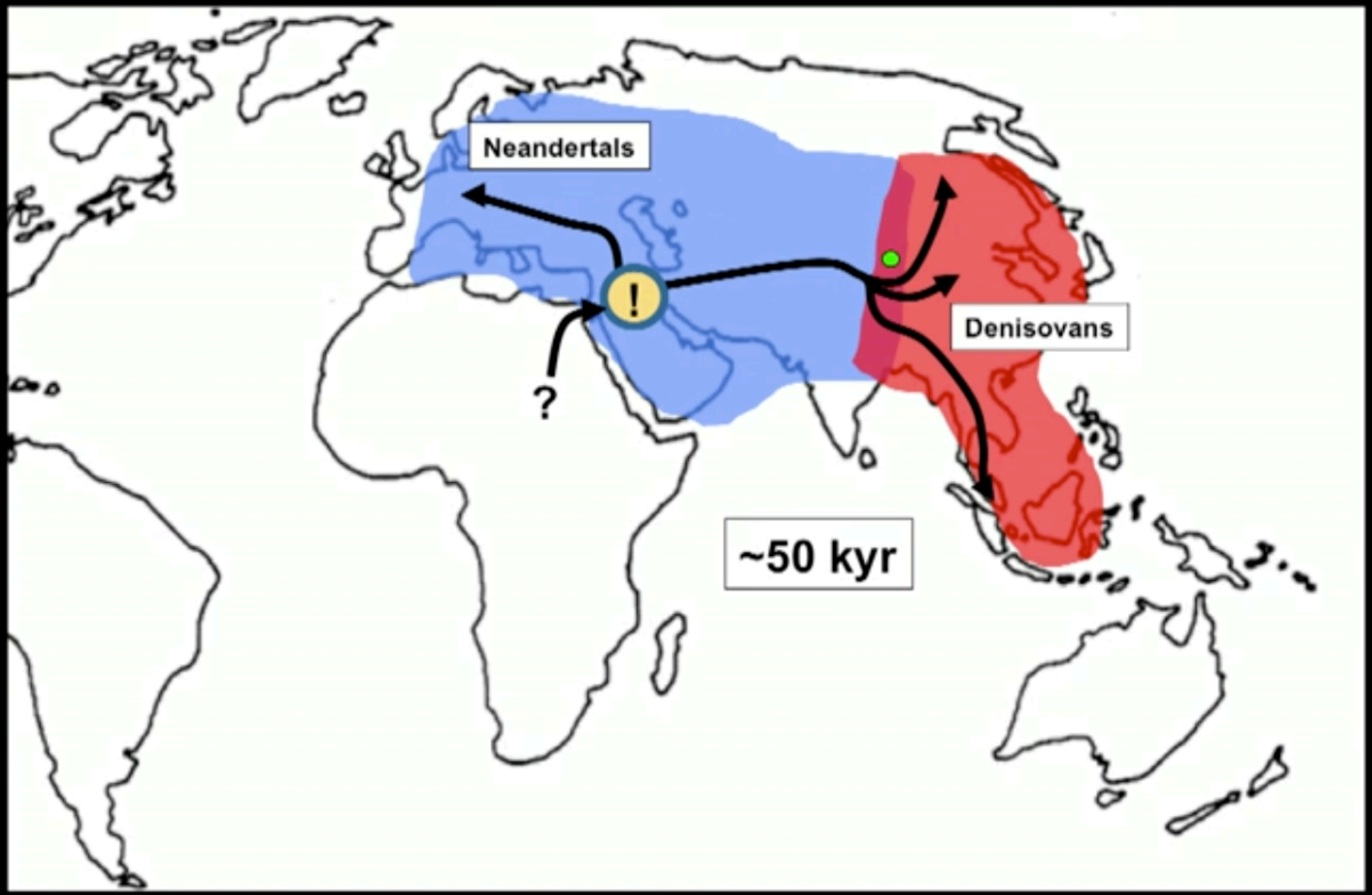
Timeline of ancient hominids



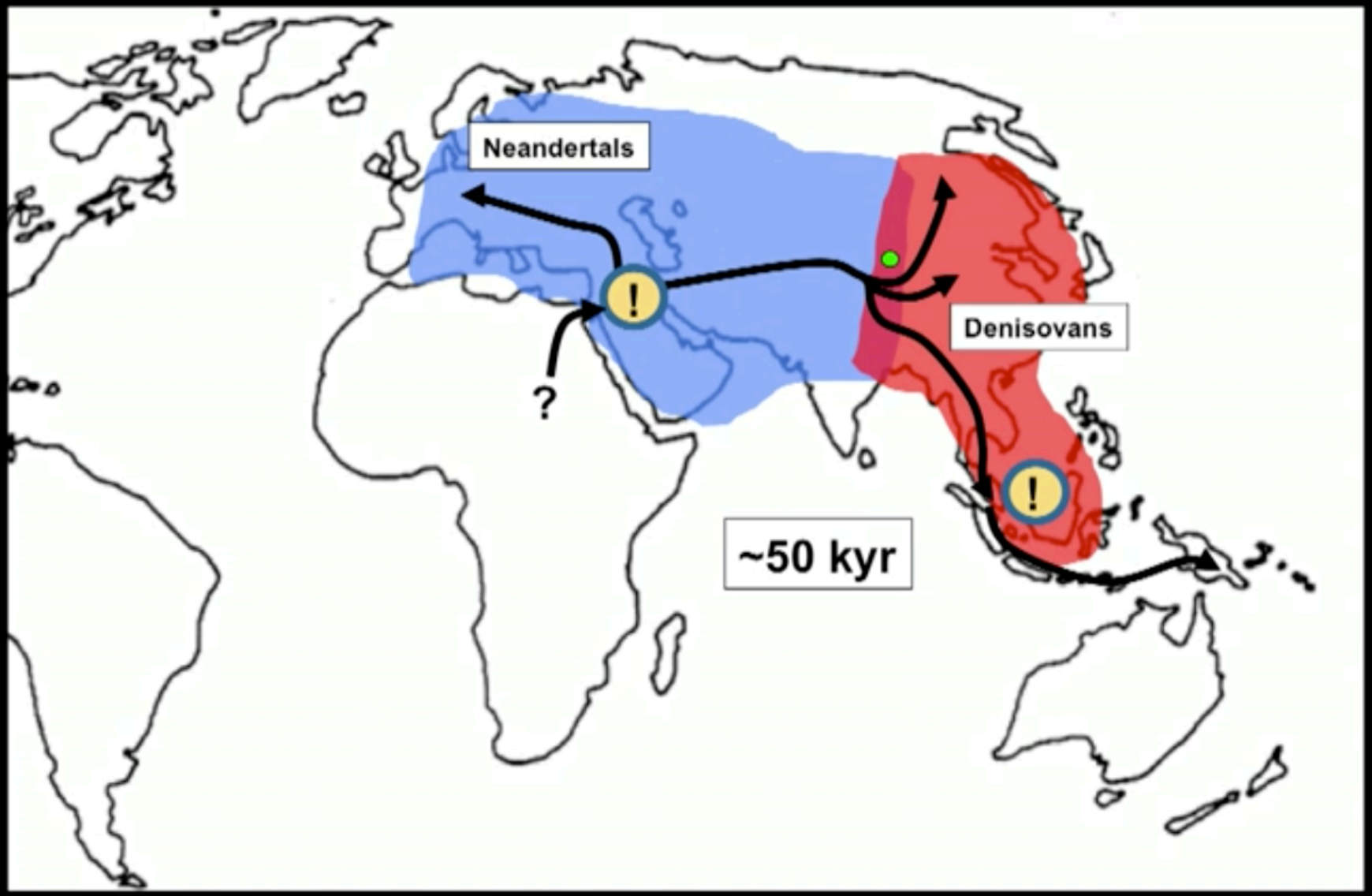
Timeline of ancient hominids



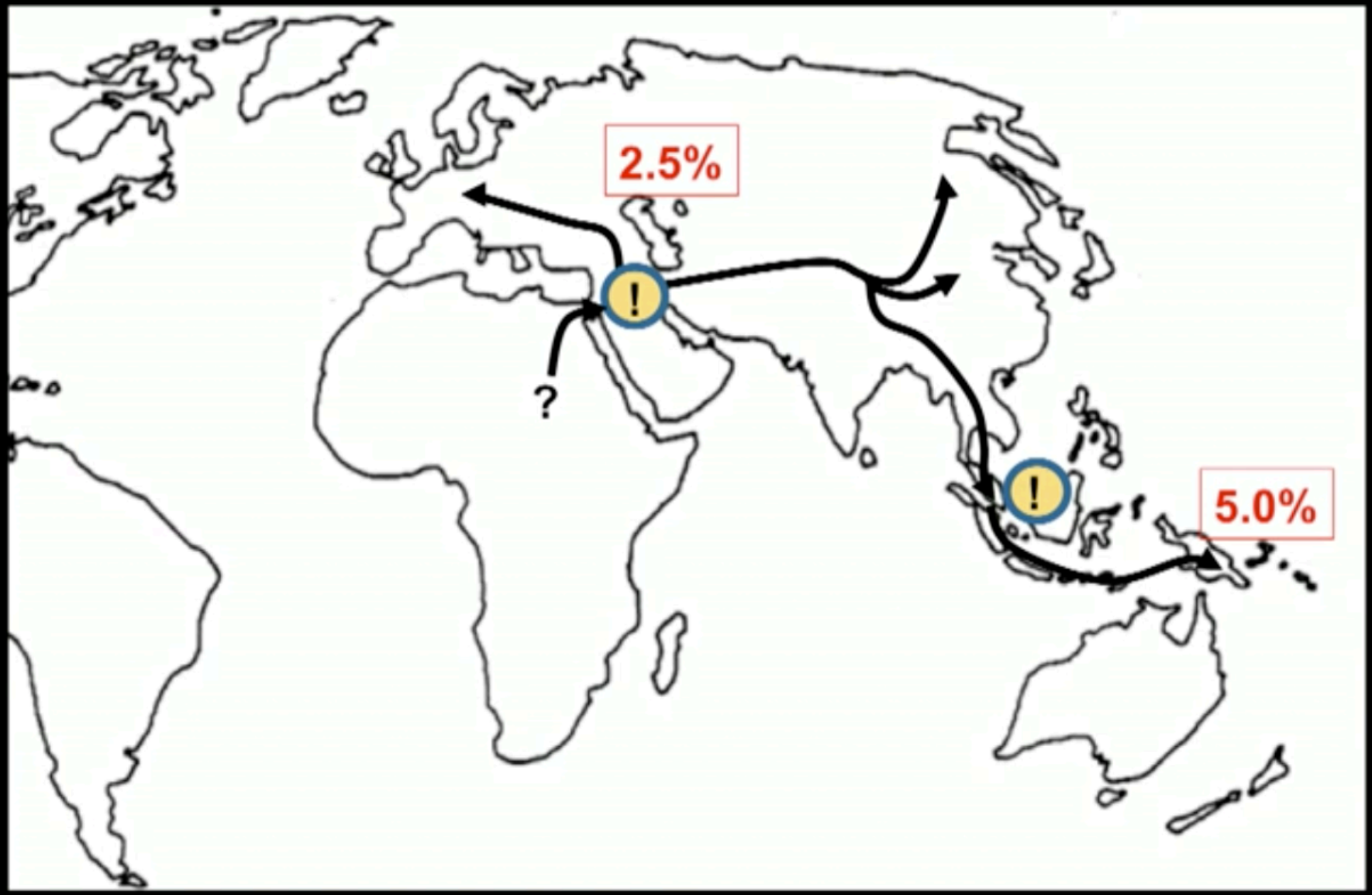
Timeline of ancient hominids



Timeline of ancient hominids



Timeline of ancient hominids



We have always mixed!

Science

REPORTS

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

Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals

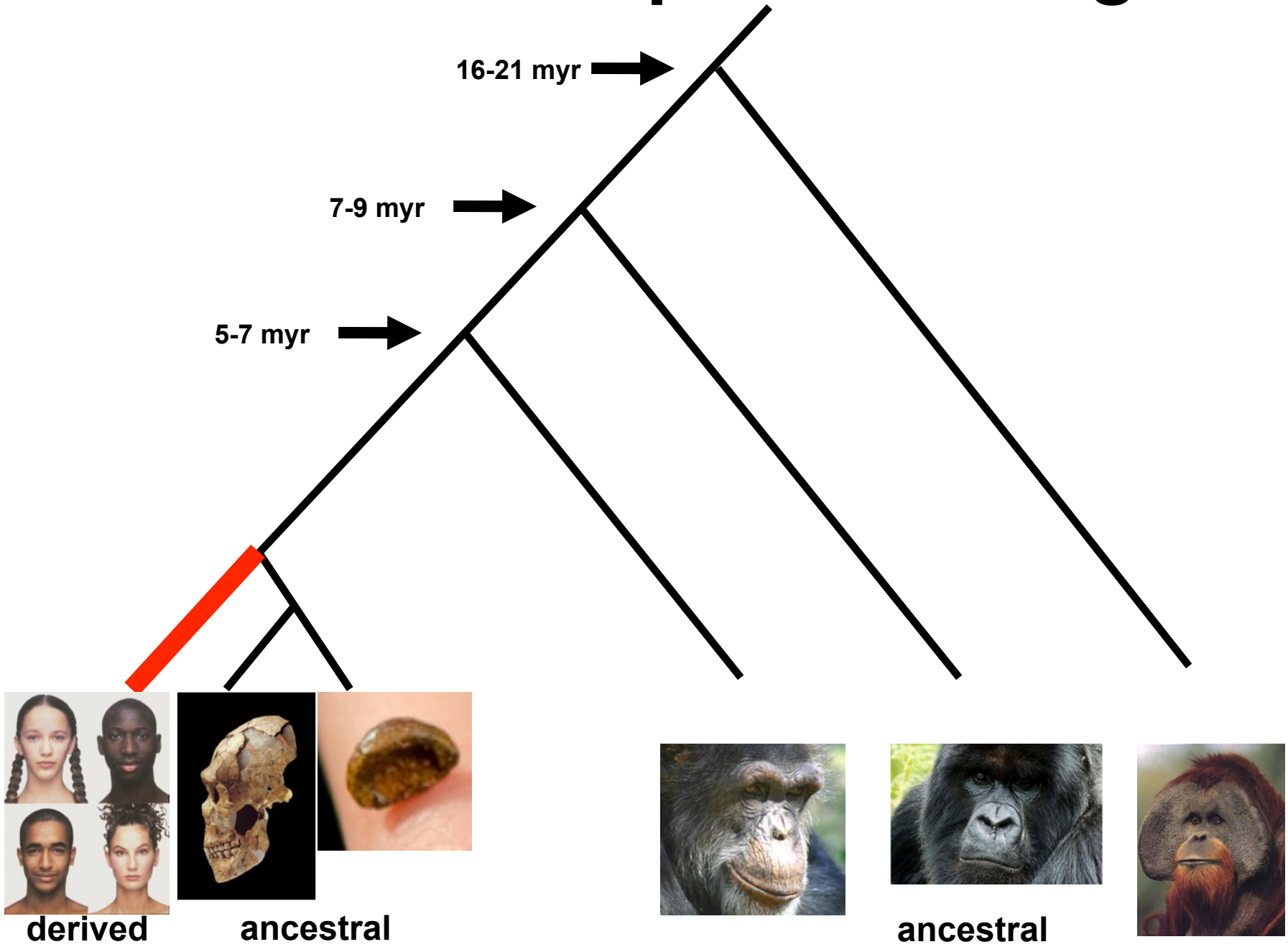
Benjamin Vernot,[1] Serena Tucci,[1,2] Janet Kelso,[3] Joshua G. Schraiber,[1] Aaron B. Wolf,[1] Rachel M. Gittelman,[1] Michael Dannemann,[3] Steffi Grote,[3] Rajiv C. McCoy,[1] Heather Norton,[4] Laura B. Scheinfeldt,[5] David A. Merriwether,[6] George Koki,[7] Jonathan S. Friedlaender,[8] Jon Wakefield,[9]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

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

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

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

Modern human-specific changes



Recipe for a modern human

109,295 single nucleotide changes (SNCs)

7,944 insertions and deletions

Changes in protein coding genes

277 cause fixed amino acid substitutions

87 affect splice sites

Changes in Non-coding & regulatory sequences

26 affect well-defined motifs inside regulatory regions

Enrichment analysis

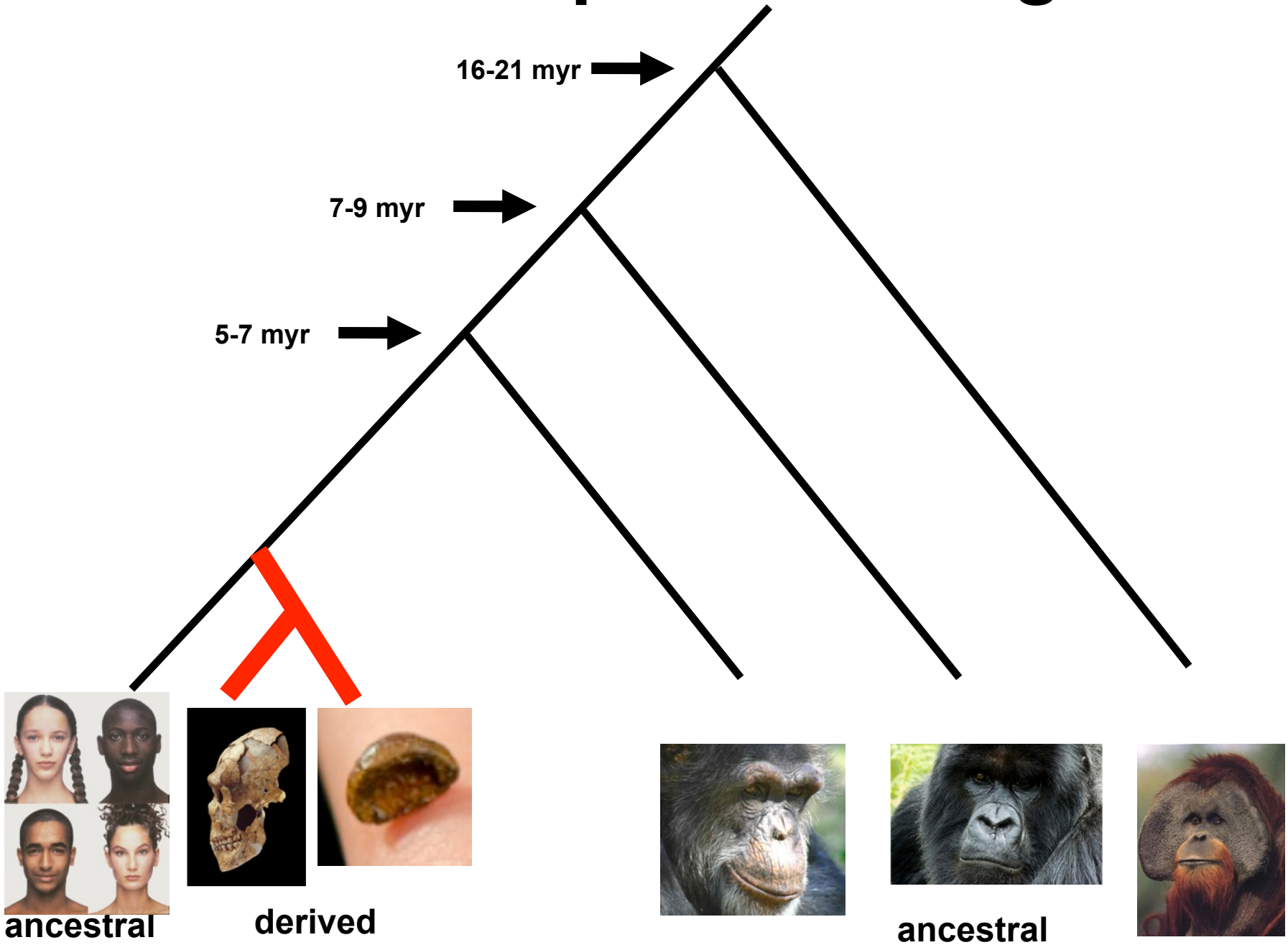
Nonsynonymous	None	- Giant melanosomes in melanocytes (p-6.77e-6; FWER=0.091;
Splice sites		
3' UTR	None	- 1-3 toe syndactyly (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - 1-5 toe syndactyly (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Aplasia/Hypoplasia of the distal phalanx of the thumb (p=1.34288e-05; 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) **Skeletal and hair morphology** - Abnormality of the finger (p=0.0005; FWER=0.44; FDR=0.08) - Brachydactyly syndrome (p=0.00062; FWER=0.48; FDR=0.088)

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

FOXP2 Analysis

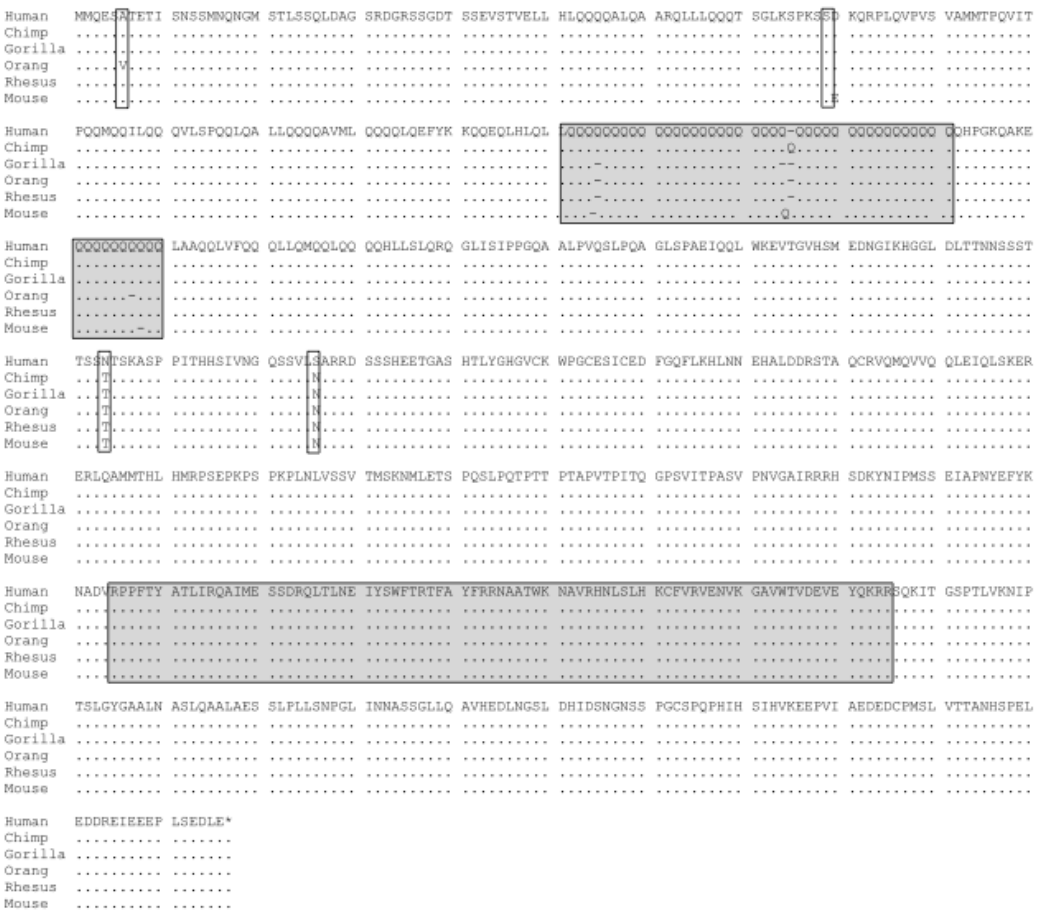


Figure 1 Alignment of the amino-acid sequences inferred from the *FOXP2* cDNA sequences. The polyglutamine stretches and the forkhead domain are shaded. Sites that differ human sequence are boxed.

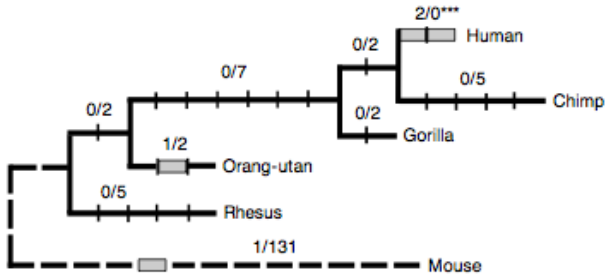


Figure 2 Silent and replacement nucleotide substitutions mapped on a phylogeny of primates. Bars represent nucleotide changes. Grey bars indicate amino-acid changes.

- 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

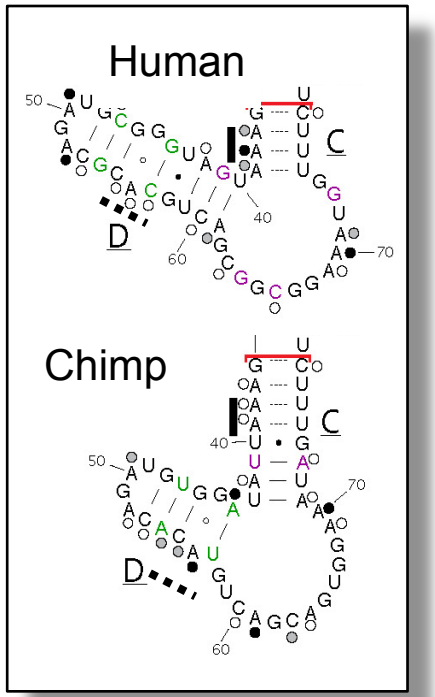
Enard *et al* (2002) *Nature*. doi:10.1038/nature01025

What makes us human?

"*Human Accelerated Regions*"

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.

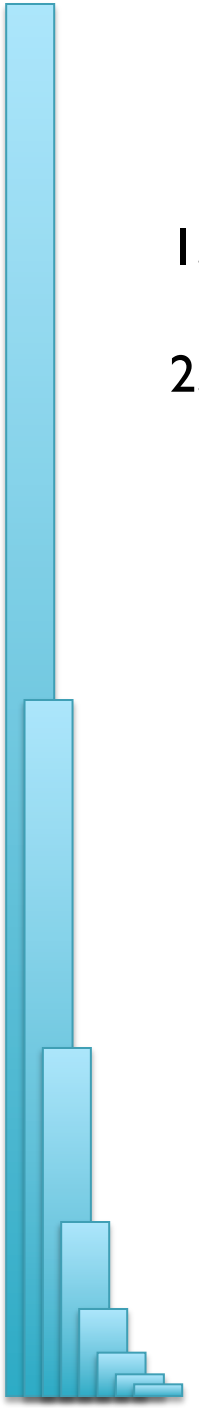


(Pollard et al., *Nature*, 2006)

Next Steps

1. Questions on project?

2. Check out the course webpage




Welcome to Applied Comparative Genomics
https://github.com/schatzlab/appliedgenomics

Questions?