Lecture 19. Ancient and Modern Humans

Michael Schatz

March 31, 2021

JHU 600.749: Applied Comparative Genomics

Preliminary Project Report

Assignment Date: March 24, 2021
Due Date: Monday, April 7, 2021 @ 11:59pm

Each team should submit a PDF of your preliminary project proposal (2 to 3 pages) to GradeScope by 11:59pm on Wednesday April 7.

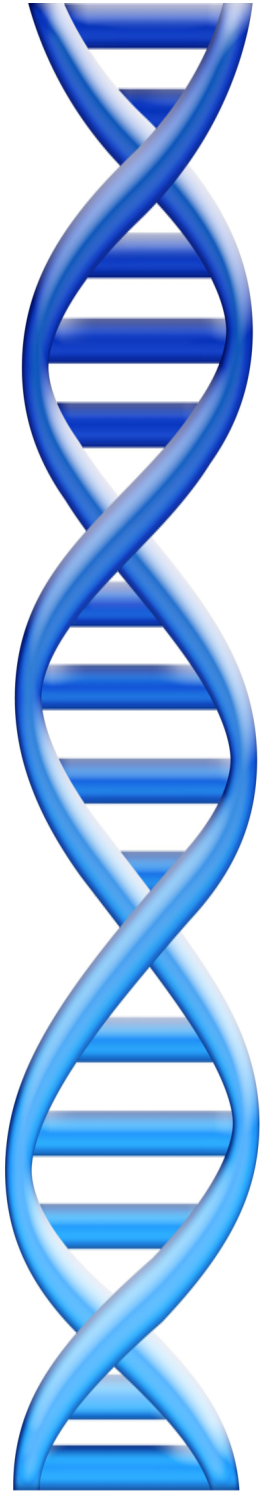
The preliminary report should have at least:

- Title of your project
- List of team members and email addresses
- 1 paragraph abstract summarizing the project
- 1+ paragraph of Introduction
- 1+ paragraph of Methods that you are using
- 1+ paragraph of Results, describing the data evaluated and any any preliminary results
- 1+ paragraph of Discussion (what you have seen or expect to see)
- 1+ figure showing a preliminary result
- 5+ References to relevant papers and data

The preliminary report should use the Bioinformatics style template. Word and LaTeX templates are available at https://academic.oup.com/bioinformatics/pages/submission_online. Overleaf is recommended for LaTex submissions. Google Docs is recommended for non-latex submissions, especially group projects. Paperpile is recommended for citation management.

Later, you will present your project in class starting the week of April 21. You will also submit your final written report (5-7 pages) of your project by May 13

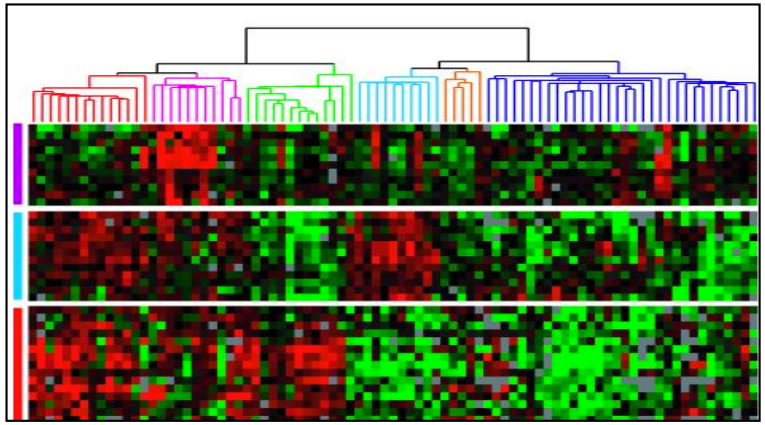
Please use Piazza if you have any general questions!



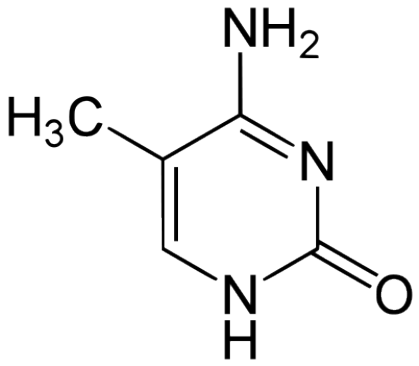
Part 0:
Gene Regulation

*-seq in 4 short vignettes

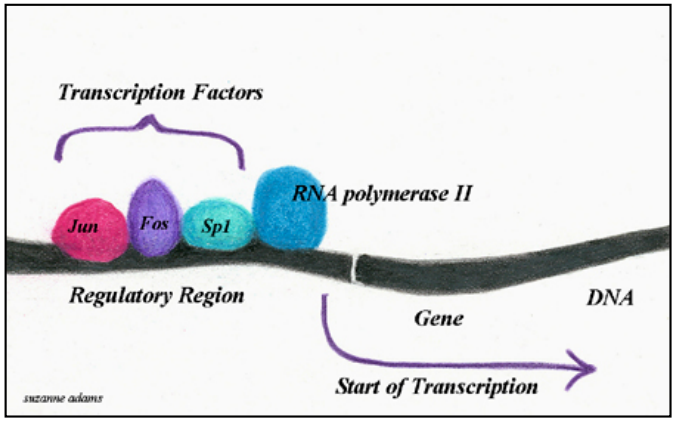
RNA-seq



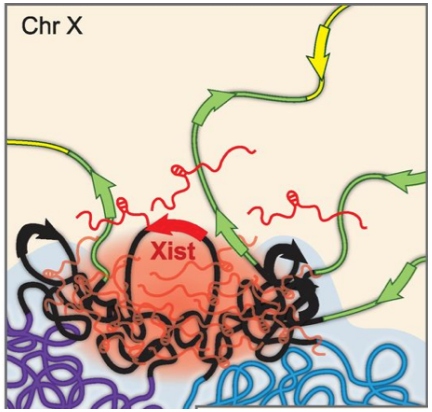
Methyl-seq



ChIP-seq



Hi-C



ARTICLE

doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.

Major Findings

1. *The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type.*

2. *Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus, some of them are expected to be functional.*

3. *Classifying the genome into seven chromatin states indicates an initial set of 399,124 regions with enhancer-like features and 70,292 regions with promoter-like features, as well as hundreds of thousands of quiescent regions. High-resolution analyses further subdivide the genome into thousands of narrow states with distinct functional properties.*

4. *It is possible to correlate quantitatively RNA sequence production and processing with both chromatin marks and transcription factor binding at promoters, indicating that promoter functionality can explain most of the variation in RNA expression.*

5. *Many non-coding variants in individual genome sequences lie in ENCODE-annotated functional regions; this number is at least as large as those that lie in protein-coding genes.*

6. *Single nucleotide polymorphisms (SNPs) associated with disease by GWAS are enriched within non-coding functional elements, with a majority residing in or near ENCODE-defined regions that are outside of protein-coding genes. In many cases, the disease phenotypes can be associated with a specific cell type or transcription factor.*

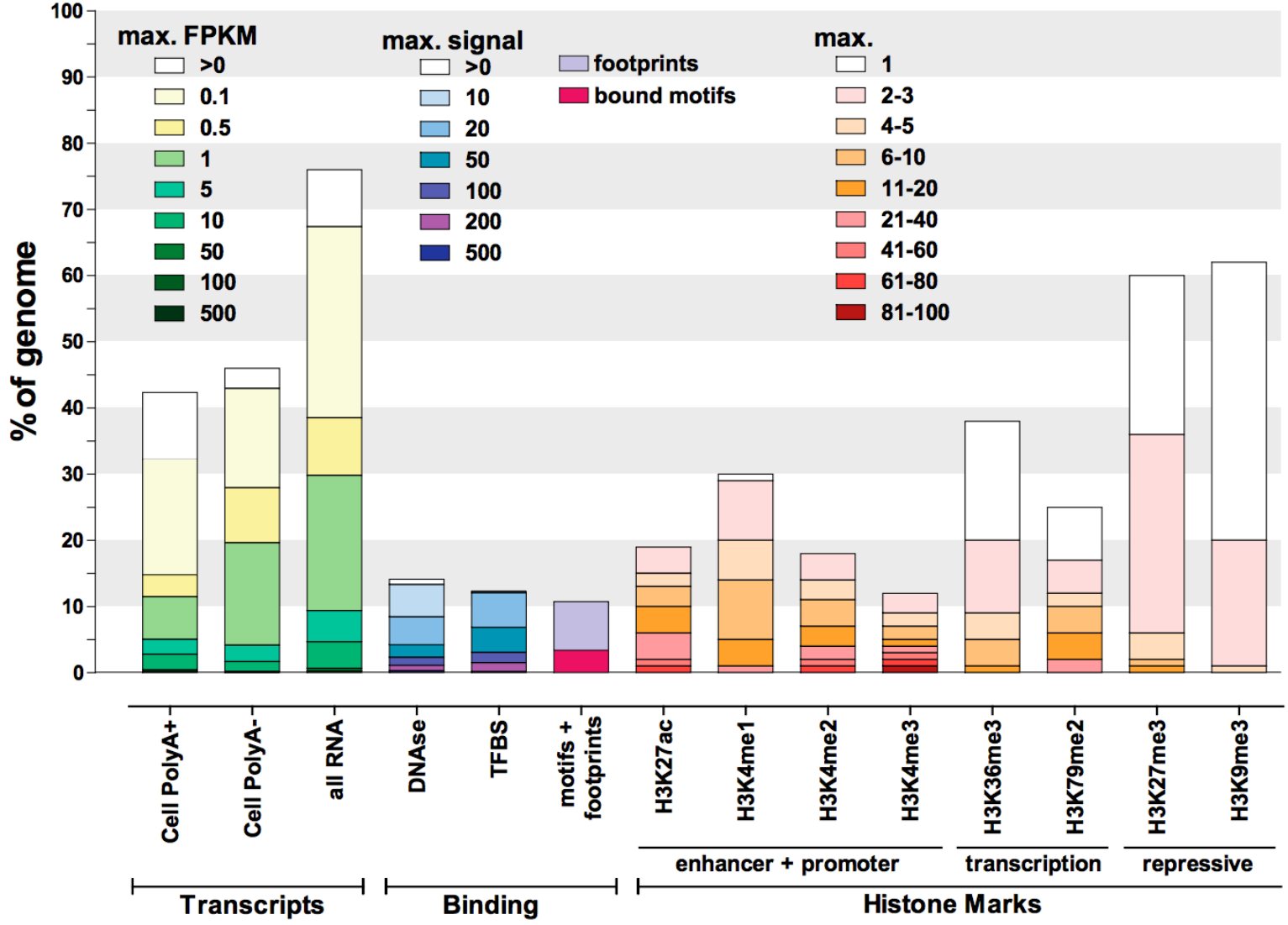
Summary of ENCODE elements

"Accounting for all these elements, a surprisingly large amount of the human genome, 80.4%, is covered by at least one ENCODE-identified element"

- 62% transcribed
- 56% enriched for histone marks
- 15% open chromatin
- 8% TF binding
- 19% At least one DHS or TF Chip-seq peak
- 4% TF binding site motif
- (Note protein coding genes comprise ~2.94% of the genome)

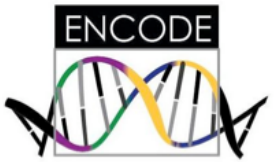
"Given that the ENCODE project did not assay all cell types, or all transcription factors, and in particular has sampled few specialized or developmentally restricted cell lineages, ***these proportions must be underestimates of the total amount of functional bases."***

Pervasive Transcription and Regulation



Defining functional DNA elements in the human genome
Kellis et al (2014). *PNAS* 6131–6138, doi: 10.1073/pnas.1318948111

Major Findings



1. *The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type.*

2. *Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus, some of them are expected to be functional.*

3. ***Classifying the genome into seven chromatin states indicates an initial set of 399,124 regions with enhancer-like features and 70,292 regions with promoter-like features, as well as hundreds of thousands of quiescent regions. High-resolution analyses further subdivide the genome into thousands of narrow states with distinct functional properties.***

4. *It is possible to correlate quantitatively RNA sequence production and processing with both chromatin marks and transcription factor binding at promoters, indicating that promoter functionality can explain most of the variation in RNA expression.*

5. *Many non-coding variants in individual genome sequences lie in ENCODE-annotated functional regions; this number is at least as large as those that lie in protein-coding genes.*

6. *Single nucleotide polymorphisms (SNPs) associated with disease by GWAS are enriched within non-coding functional elements, with a majority residing in or near ENCODE-defined regions that are outside of protein-coding genes. In many cases, the disease phenotypes can be associated with a specific cell type or transcription factor.*

Signal Integration

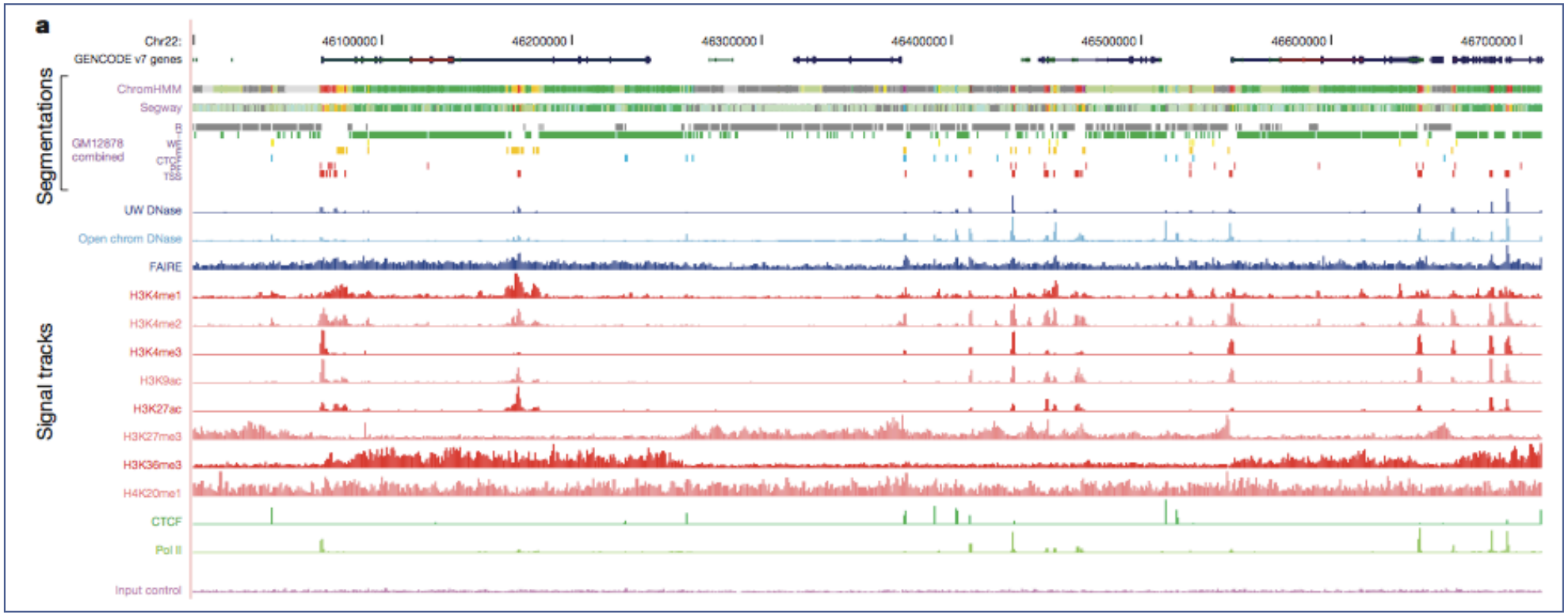
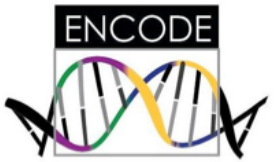


Table 3 | Summary of the combined state types

Label	Description	Details*	Colour
CTCF	CTCF-enriched element	Sites of CTCF signal lacking histone modifications, often associated with open chromatin. Many probably have a function in insulator assays, but because of the multifunctional nature of CTCF, we are conservative in our description. Also enriched for the cohesin components RAD21 and SMC3; CTCF is known to recruit the cohesin complex.	Turquoise
E	Predicted enhancer	Regions of open chromatin associated with H3K4me1 signal. Enriched for other enhancer-associated marks, including transcription factors known to act at enhancers. In enhancer assays, many of these (>50%) function as enhancers. A more conservative alternative would be *cis*-regulatory regions. Enriched for sites for the proteins encoded by *EP300*, *FOS*, *FOSL1*, *GATA2*, *HDAC8*, *JUNB*, *JUND*, *NFE2*, *SMARCA4*, *SMARCB1*, *SIRT6* and *TAL1* genes in K562 cells. Have nuclear and whole-cell RNA signal, particularly poly(A)− fraction.	Orange
PF	Predicted promoter flanking region	Regions that generally surround TSS segments (see below).	Light red
R	Predicted repressed or low-activity region	This is a merged state that includes H3K27me3 polycomb-enriched regions, along with regions that are silent in terms of observed signal for the input assays to the segmentations (low or no signal). They may have other signals (for example, RNA, not in the segmentation input data). Enriched for sites for the proteins encoded by *REST* and some other factors (for example, proteins encoded by *BRF2*, *CEBPB*, *MAFK*, *TRIM28*, *ZNF274* and *SETDB1* genes in K562 cells).	Grey
TSS	Predicted promoter region including TSS	Found close to or overlapping GENCODE TSS sites. High precision/recall for TSSs. Enriched for H3K4me3. Sites of open chromatin. Enriched for transcription factors known to act close to promoters and polymerases Pol II and Pol III. Short RNAs are most enriched in these segments.	Bright red
T	Predicted transcribed region	Overlap gene bodies with H3K36me3 transcriptional elongation signal. Enriched for phosphorylated form of Pol II signal (elongating polymerase) and poly(A)$^+$ RNA, especially cytoplasmic.	Dark green
WE	Predicted weak enhancer or open chromatin *cis*-regulatory element	Similar to the E state, but weaker signals and weaker enrichments.	Yellow

- Use ChromHMM and Segway to Summarize the individual assays into 7 functional/regulatory states

Major Findings



1. *The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type.*

2. *Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus, some of them are expected to be functional.*

3. *Classifying the genome into seven chromatin states indicates an initial set of 399,124 regions with enhancer-like features and 70,292 regions with promoter-like features, as well as hundreds of thousands of quiescent regions. High-resolution analyses further subdivide the genome into thousands of narrow states with distinct functional properties.*

4. *It is possible to correlate quantitatively RNA sequence production and processing with both chromatin marks and transcription factor binding at promoters, indicating that promoter functionality can explain most of the variation in RNA expression.*

5. *Many non-coding variants in individual genome sequences lie in ENCODE-annotated functional regions; this number is at least as large as those that lie in protein-coding genes.*

6. ***Single nucleotide polymorphisms (SNPs) associated with disease by GWAS are enriched within non-coding functional elements, with a majority residing in or near ENCODE-defined regions that are outside of protein-coding genes. In many cases, the disease phenotypes can be associated with a specific cell type or transcription factor.***

ENCODE and Disease

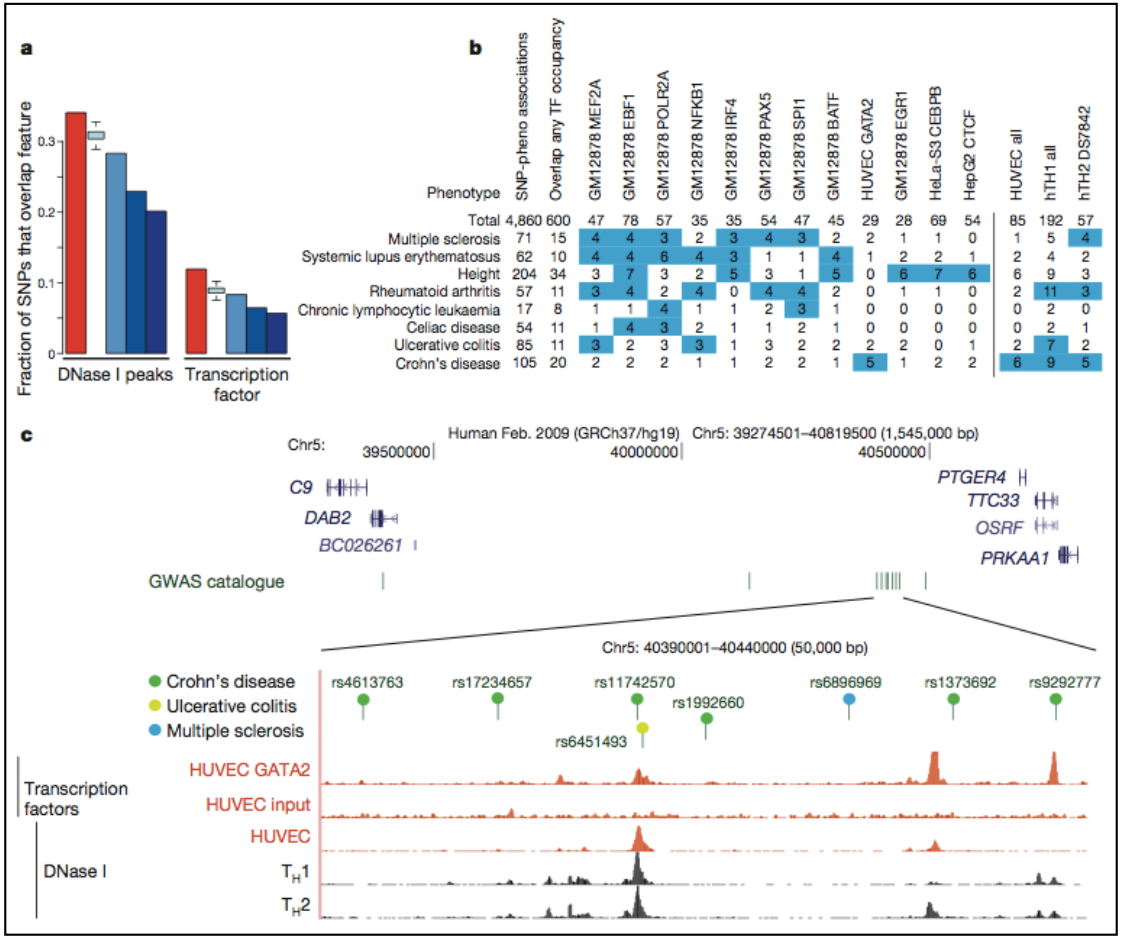


Figure 10 | Comparison of genome-wide-association-study-identified loci with ENCODE data. **a**, Overlap of lead SNPs in the NHGRI GWAS SNP catalogue (June 2011) with DHSs (left) or transcription-factor-binding sites (right) as red bars compared with various control SNP sets in blue. The control SNP sets are (from left to right): SNPs on the Illumina 2.5M chip as an example of a widely used GWAS SNP typing panel; SNPs from the 1000 Genomes project; SNPs extracted from 24 personal genomes (see personal genome variants track at http://main.genome-browser.bx.psu.edu (ref. 80)), all shown as blue bars. In addition, a further control used 1,000 randomizations from the genotyping SNP panel, matching the SNPs with each NHGRI catalogue SNP for allele frequency and distance to the nearest TSS (light blue bars with bounds at 1.5 times the interquartile range). For both DHSs and transcription-factor-binding regions, a larger proportion of overlaps with GWAS-implicated SNPs is found compared to any of the controls sets. **b**, Aggregate overlap of phenotypes to selected transcription-factor-binding sites (left matrix) or DHSs in selected cell lines (right matrix), with a count of overlaps between the phenotype and the cell line/factor. Values in blue squares pass an empirical *P*-value threshold ≤0.01 (based on the same analysis of overlaps between randomly chosen, GWAS-matched SNPs and these epigenetic features) and have at least a count of three overlaps. The *P* value for the total number of phenotype–transcription factor associations is <0.001. **c**, Several SNPs associated with Crohn's disease and other inflammatory diseases that reside in a large gene desert on chromosome 5, along with some epigenetic features indicative of function. The SNP (rs11742570) strongly associated to Crohn's disease overlaps a GATA2 transcription-factor-binding signal determined in HUVECs. This region is also DNase I hypersensitive in HUVECs and T-helper T_H1 and T_H2 cells. An interactive version of this figure is available in the online version of the paper.

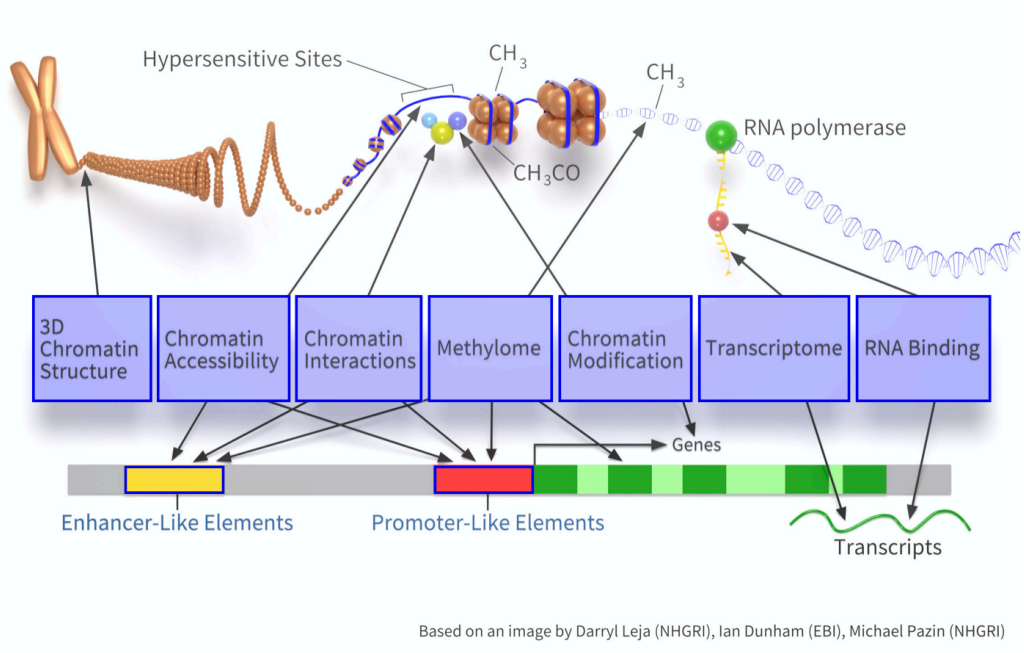
- 88% of GWAS SNPs are intronic or intergenic of unknown function
- We found that 12% of these GWAS-SNPs overlap transcription-factor-occupied regions whereas 34% overlap DHSs
- GWAS SNPs are particularly enriched in the segmentation classes associated with enhancers and TSSs across several cell types

ENCODE Studies

encodeproject.org

ENCODE Data Encyclopedia Materials & Methods Help New » Search... Sign in / Create account

ENCODE: Encyclopedia of DNA Elements



About ENCODE Project | Getting Started | Experiments

Search ENCODE portal

ENCODE | Functional Characterization Experiments

About ENCODE Encyclopedia

candidate Cis-Regulatory Elements

Search for candidate Cis-Regulatory Elements

Hosted by SCREEN

Human GRCh38 | Mouse mm10

Visit hg19 site

HUMAN | MOUSE | WORM | FLY

Data Matrix

Project Biosample Type Assay ies

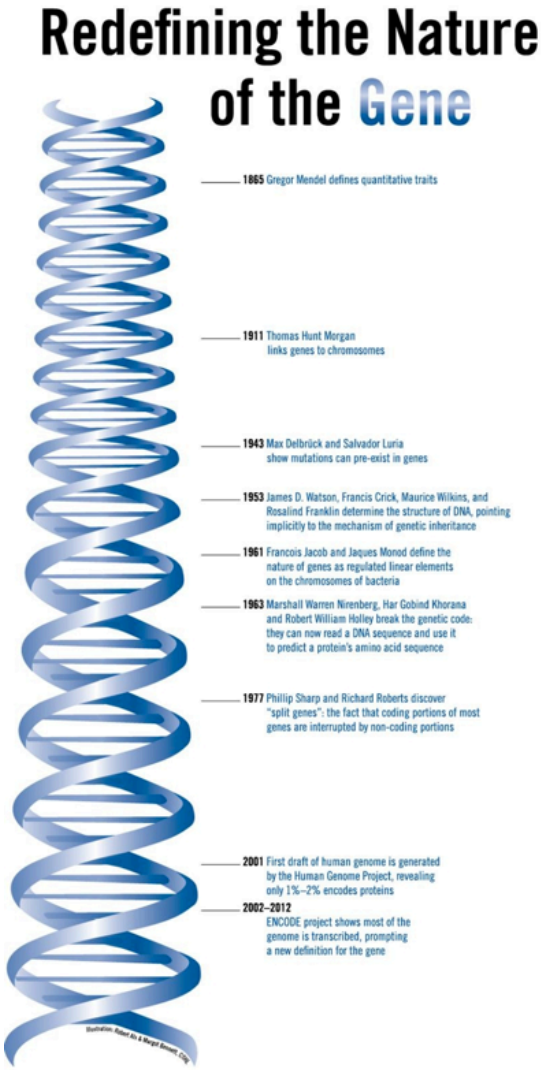
? Help

>5000 Citations for main paper; >>10k for all papers

Summary & Critique

- ***Summary***
  - *The unprecedented number of functional elements identified in this study provides a valuable resource to the scientific community as well as significantly enhances our understanding of the human genome.*

- ***Critique***
  - Was it correct?
  - What is functional?
  - What is conservation?
  - What was the control?
  - What are the tradeoffs of organizing so much funding ($288M!) around a single project; will other groups successfully use these data?



Essay

Current Biology Vol 23 No 7
R260

The ENCODE project: Missteps overshadowing a success

Two clichés of science journalism have now played out around the ENCODE project. ENCODE's publicity first presented a misleading "all the textbooks are wrong" narrative about noncoding human DNA. Now several critiques of ENCODE's narrative have been published, and one was so vitriolic that it fueled "undignified academic squabble" stories that focused on tone more than substance. Neither story line does justice to our actual understanding of genomes, to ENCODE's results, or to the role of big science in biology.

Sean R. Eddy

Essay

Current Biology Vol 23 No 7
R260

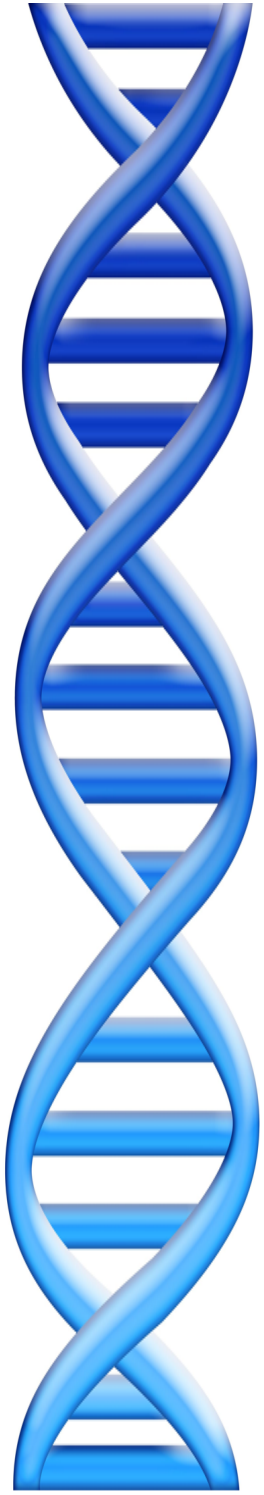
The ENCODE project: Missteps overshadowing a success

Two clichés of science journalism have now played out around the ENCODE project. ENCODE's publicity first presented a misleading "all the textbooks

*"To clarify what noise means, I propose the **Random Genome Project**. Suppose we put a few million bases of entirely random synthetic DNA into a human cell, and do an ENCODE project on it. Will it be reproducibly transcribed into mRNA-like transcripts, reproducibly bound by DNA-binding proteins, and reproducibly wrapped around histones marked by specific chromatin modifications? I think yes.*

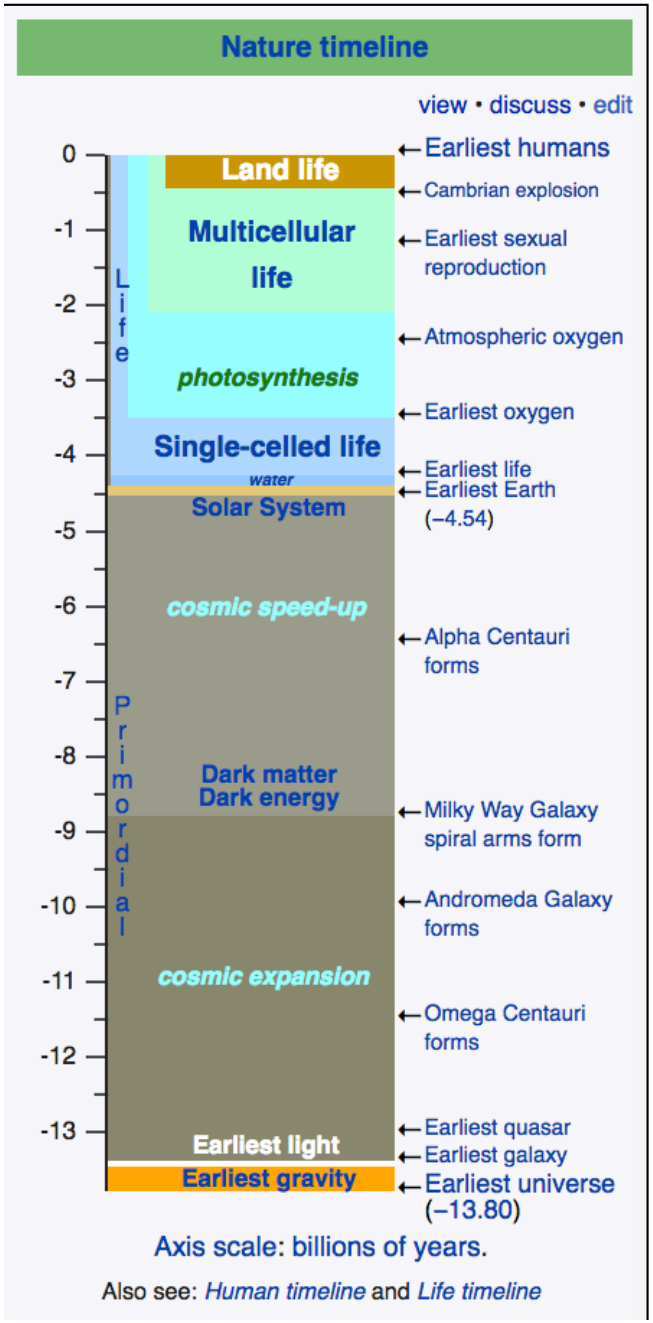
A striking feature of genetic regulation is that regulatory factors (proteins or RNAs) generally recognize and bind to small sites, small enough that any given factor will find specific binding sites even in random DNA. Promoters, enhancers, splice sites, poly-A addition sites, and other functional features in the genome all have substantial random occurrence frequencies. These sites are not nonspecific in a random genome. They are specific sequences, albeit randomly occurring and not under selection for any function.

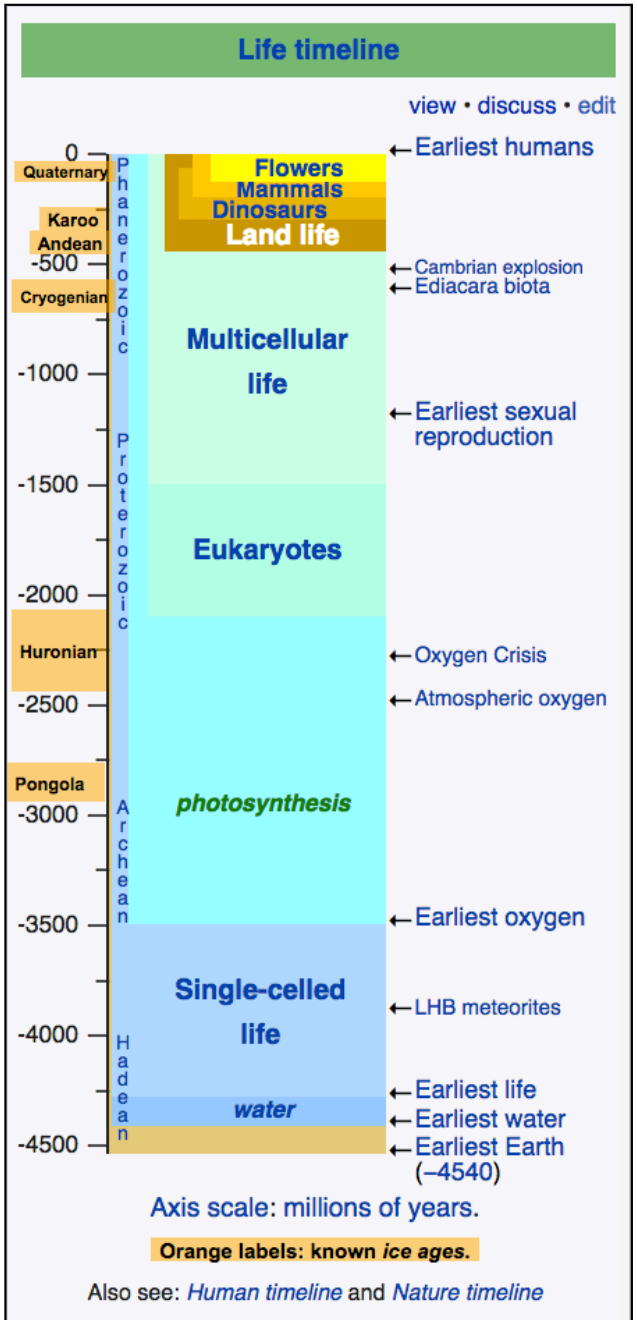
Would biochemical activities in the random genome be regulated under different conditions? For example, would they be cell type-specific? Surely yes, because the regulatory factors themselves (such as transcription factors) are regulated and expressed in specific cell types and conditions."

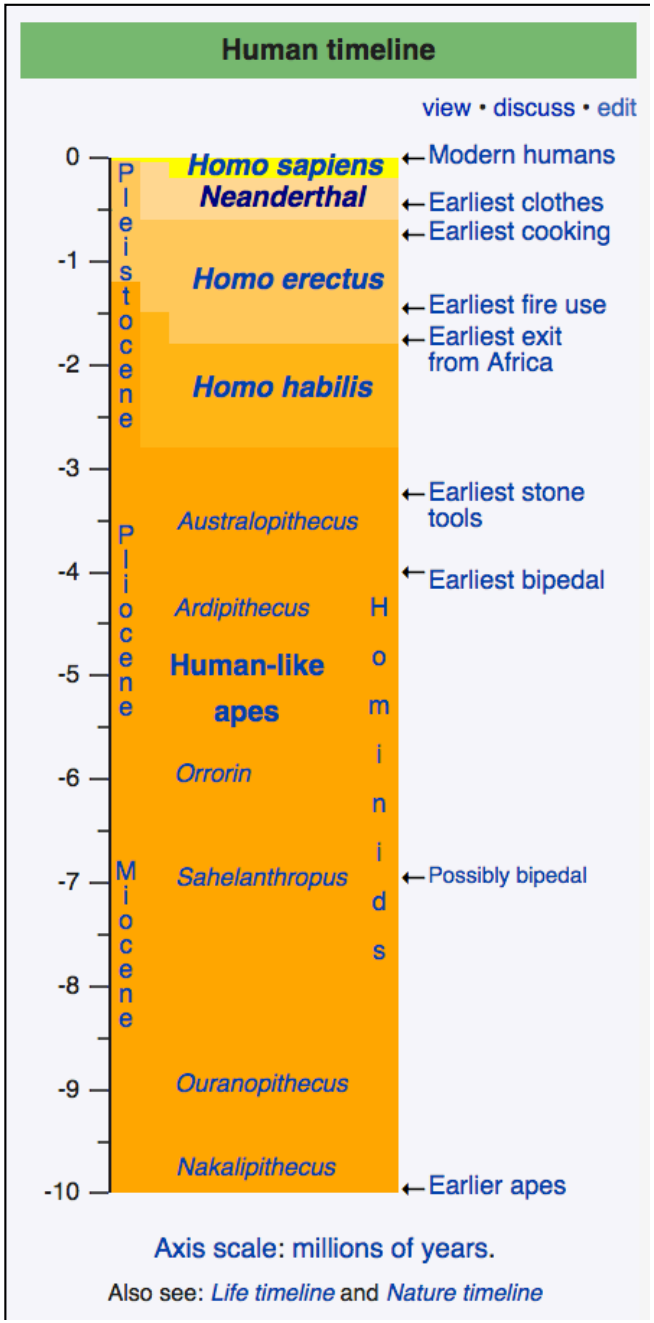


Part 1:
Ancient Hominds

Our Origins


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


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


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

Sequencing ancient genomes

Janet Kelso
Max-Planck Institute


Engis
Neandertal
Spy
Arcy-sur-Cure
La Quina
Saint-Césaire
Le Moustier
La Ferrassie
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)

Homo neanderthalensis

•Proto-Neanderthals emerge around 600k years ago

•“True” Neanderthals emerge around 200k years ago

•Died out approximately 40,000 years ago

•Known for their robust physique

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

Homo sapiens sapiens

- Apparently emerged from earlier hominids in Africa around 50k years ago
- Capable of amazing intellectual and social behaviors
- Mostly Harmless ☺

Science
AAAS

A Draft Sequence of the Neandertal Genome

Richard E. Green, *et al.*

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

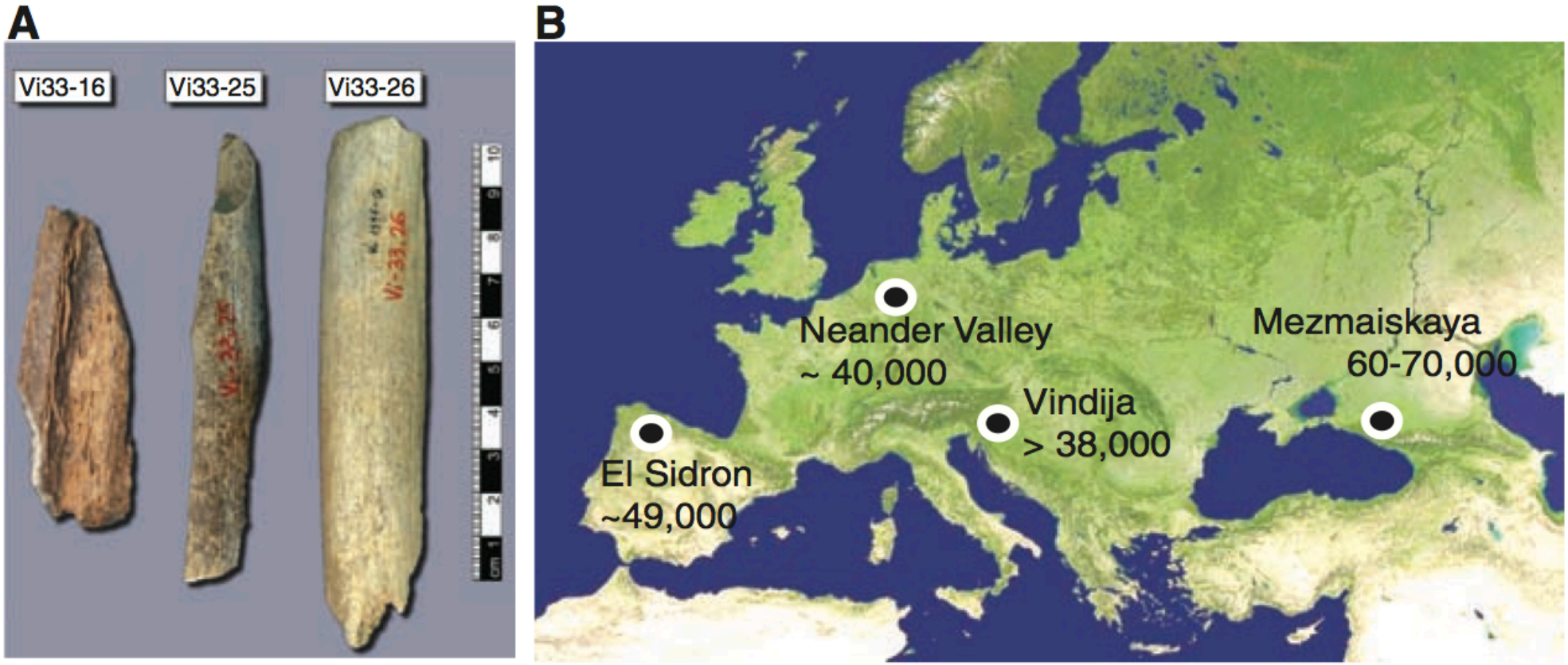
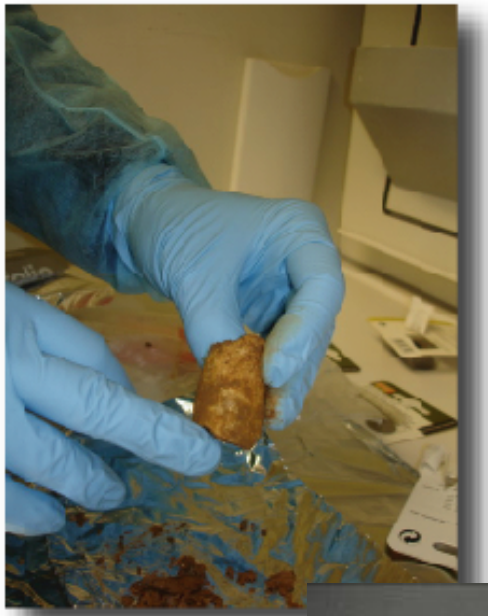
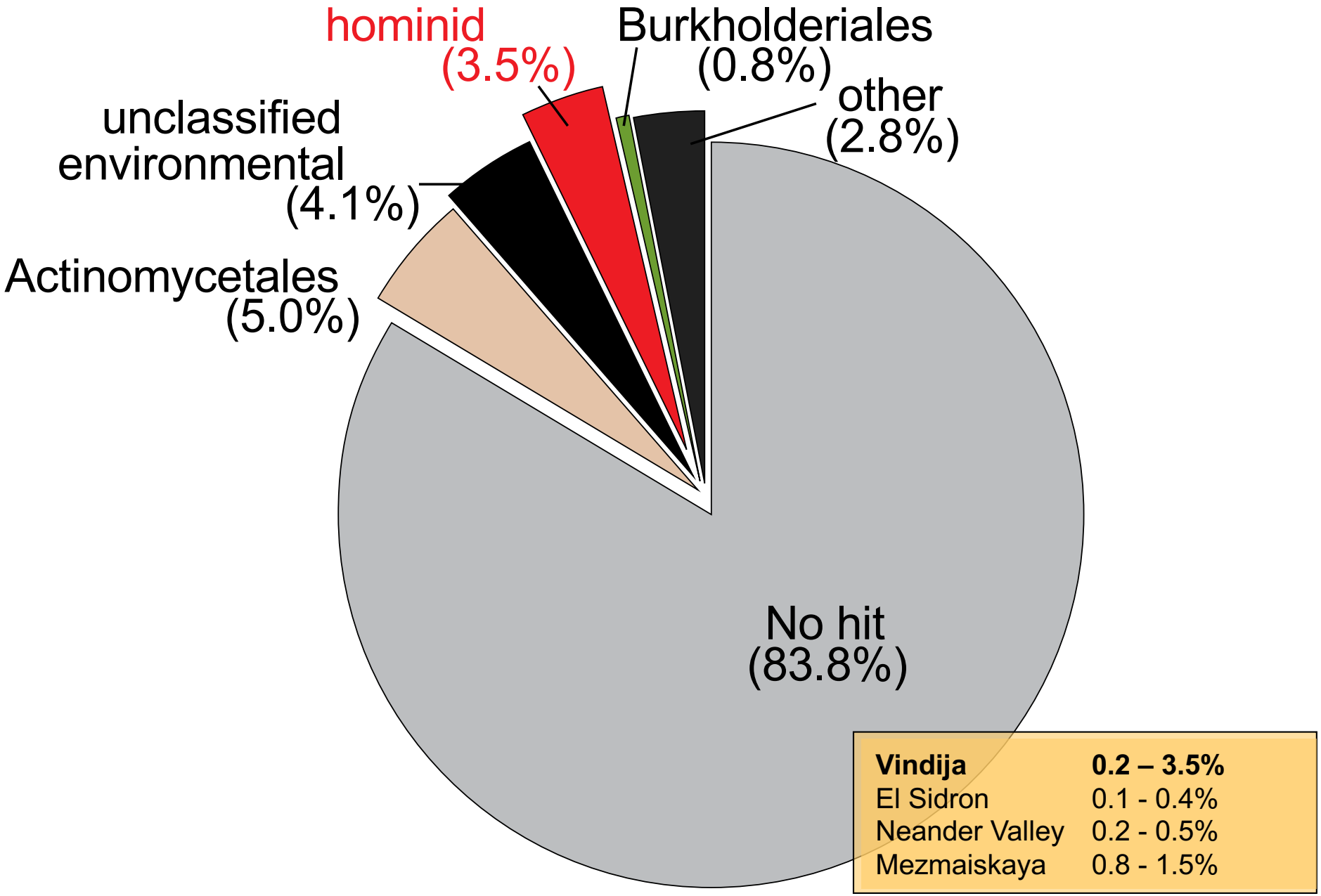


Fig. 1. Samples and sites from which DNA was retrieved. (**A**) The three bones from Vindija from which Neandertal DNA was sequenced. (**B**) Map showing the four archaeological sites from which bones were used and their approximate dates (years B.P.).

Extracting Ancient DNA

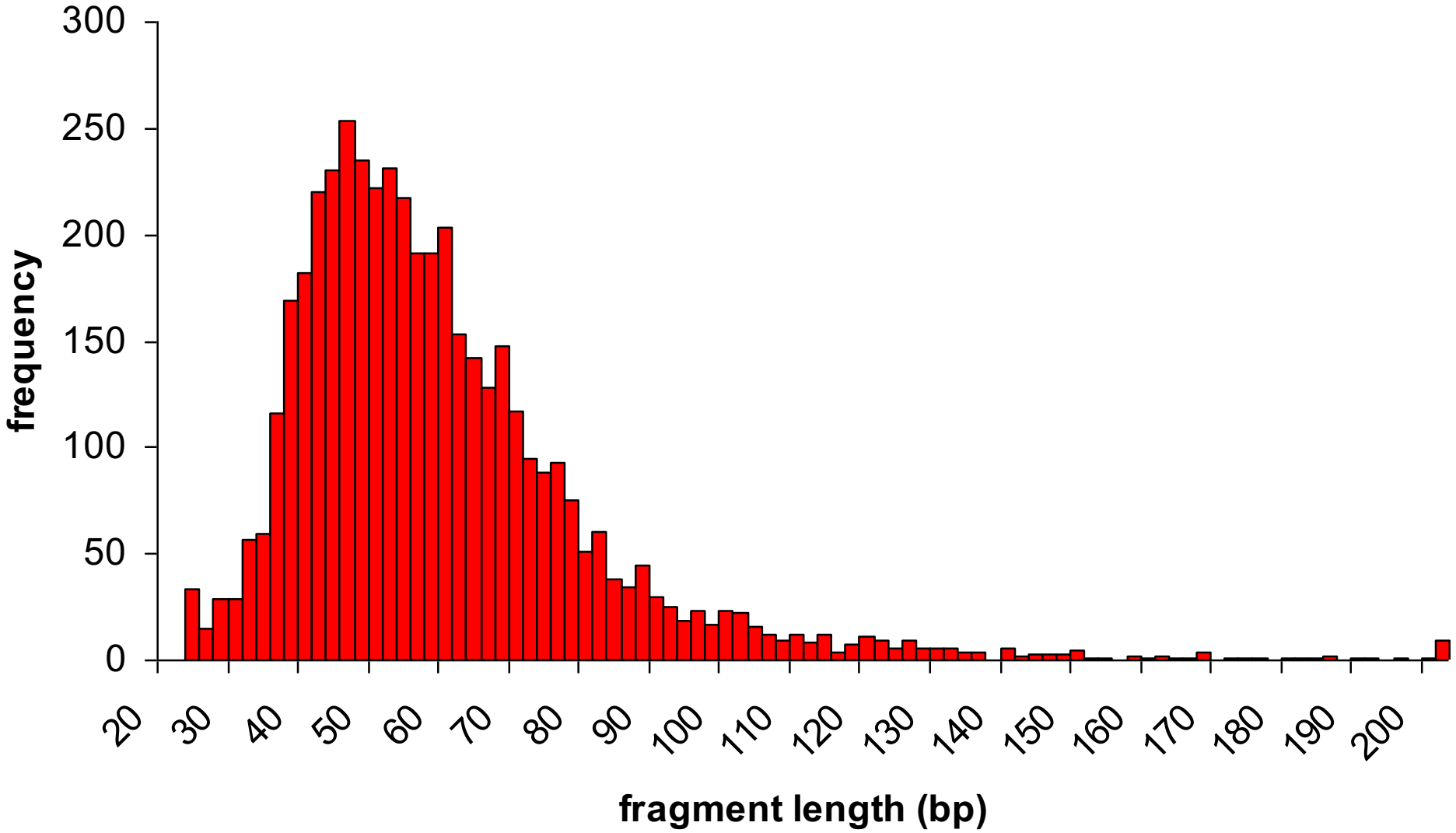


DNA is from mixed sources

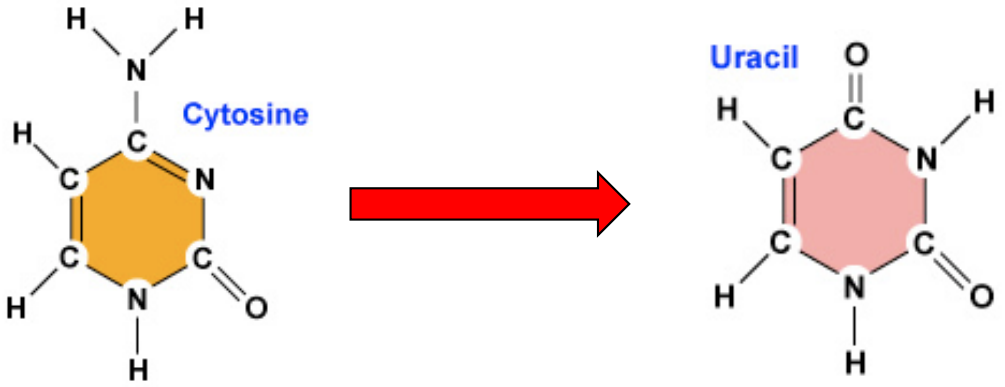


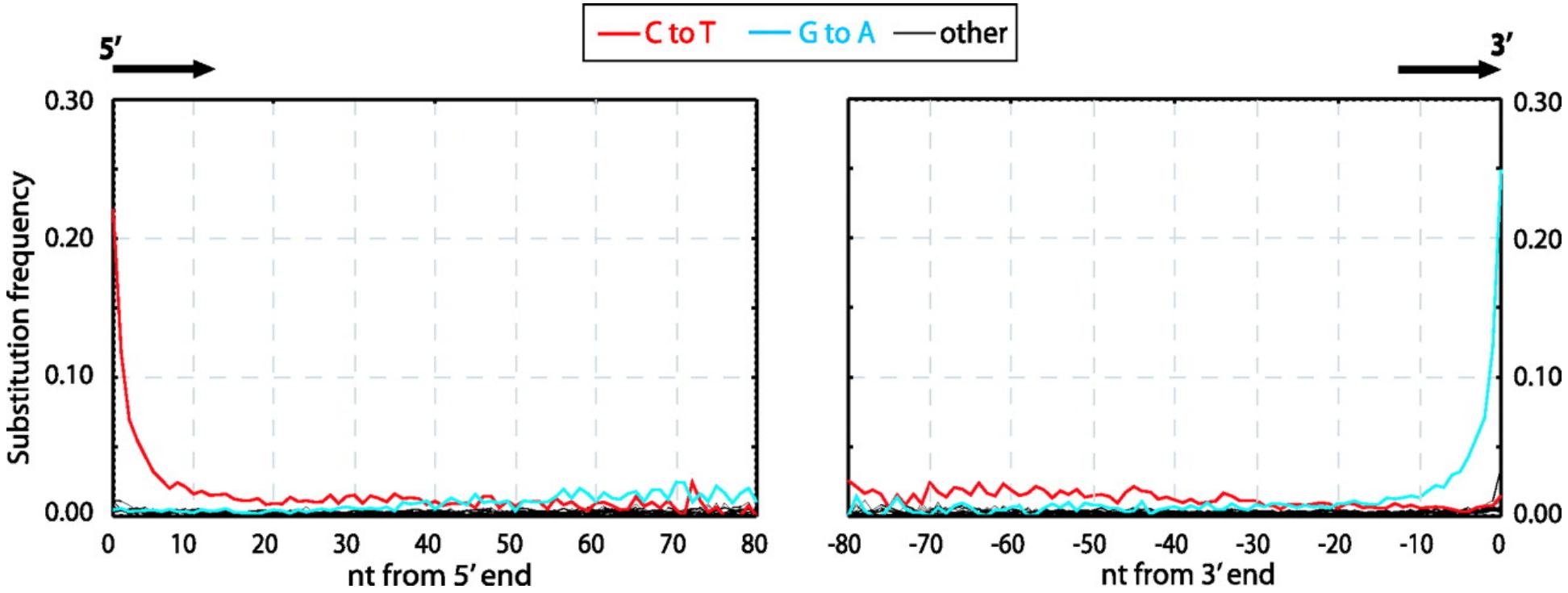
Vindija	**0.2 – 3.5%**
El Sidron	0.1 - 0.4%
Neander Valley	0.2 - 0.5%
Mezmaiskaya	0.8 - 1.5%

DNA is degraded



DNA is chemically damaged





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

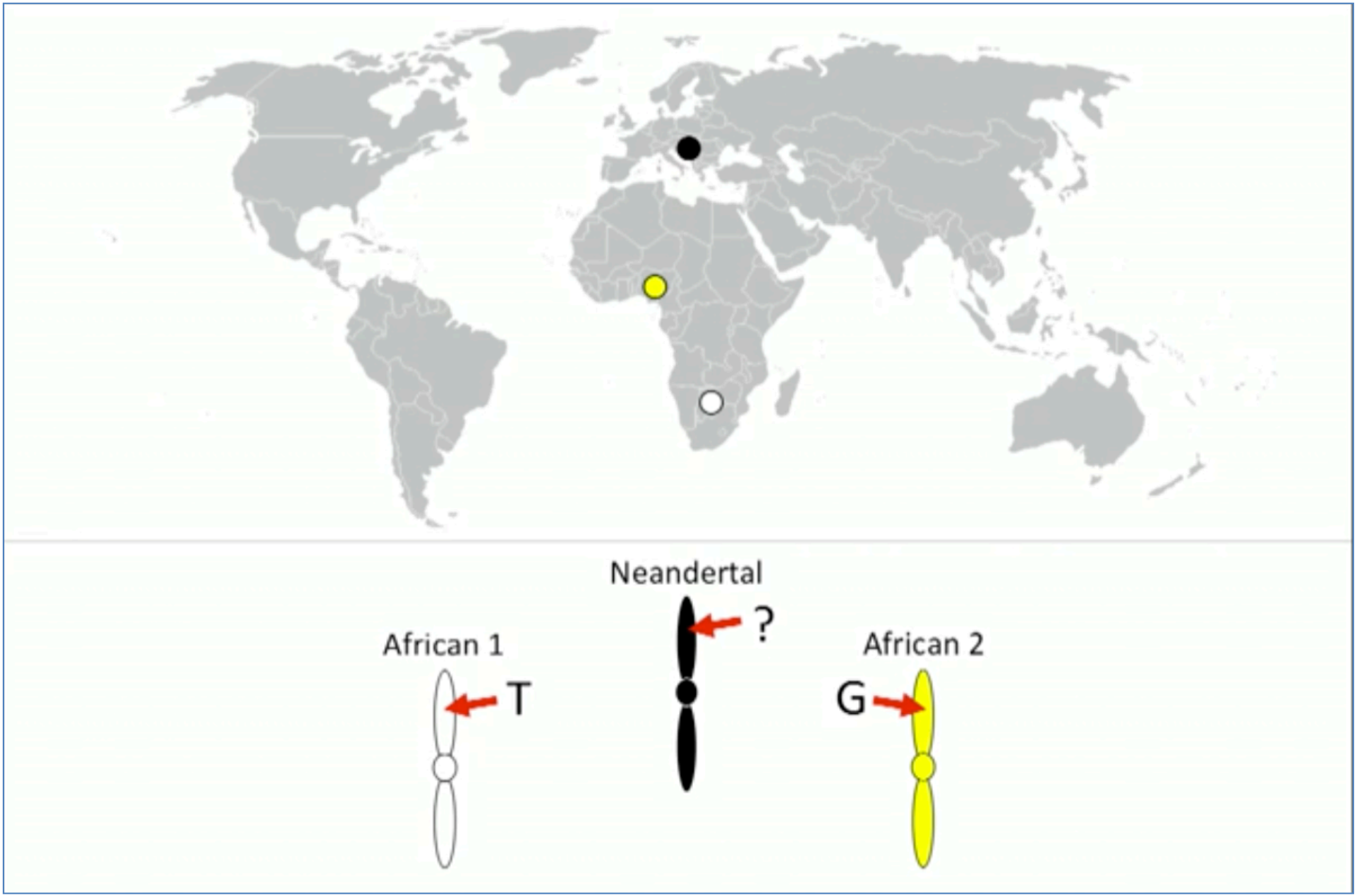
Green et al. 2010

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

~35 Illumina flow cells

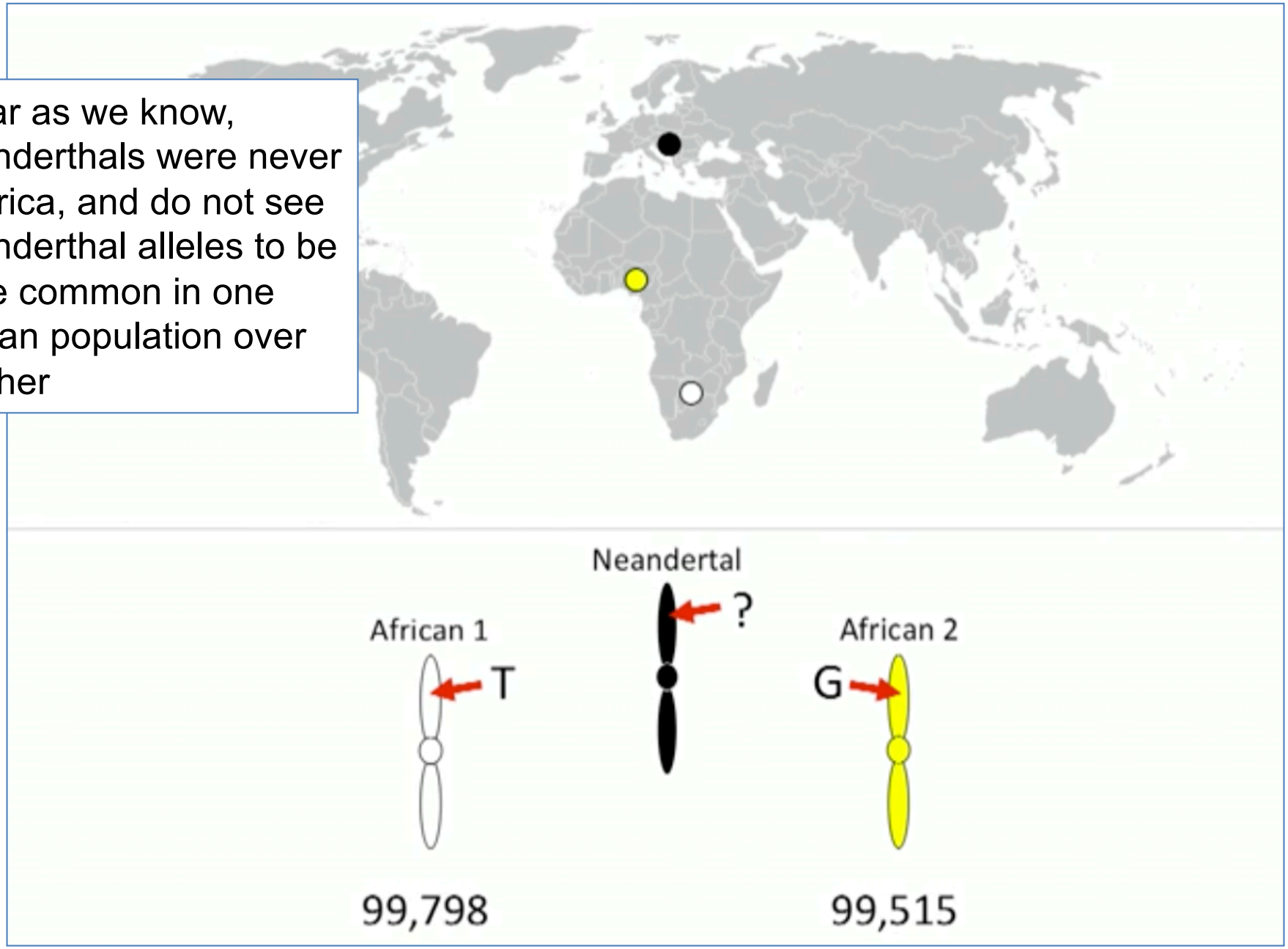
Genome coverage ~1.3 X

Did we mix?



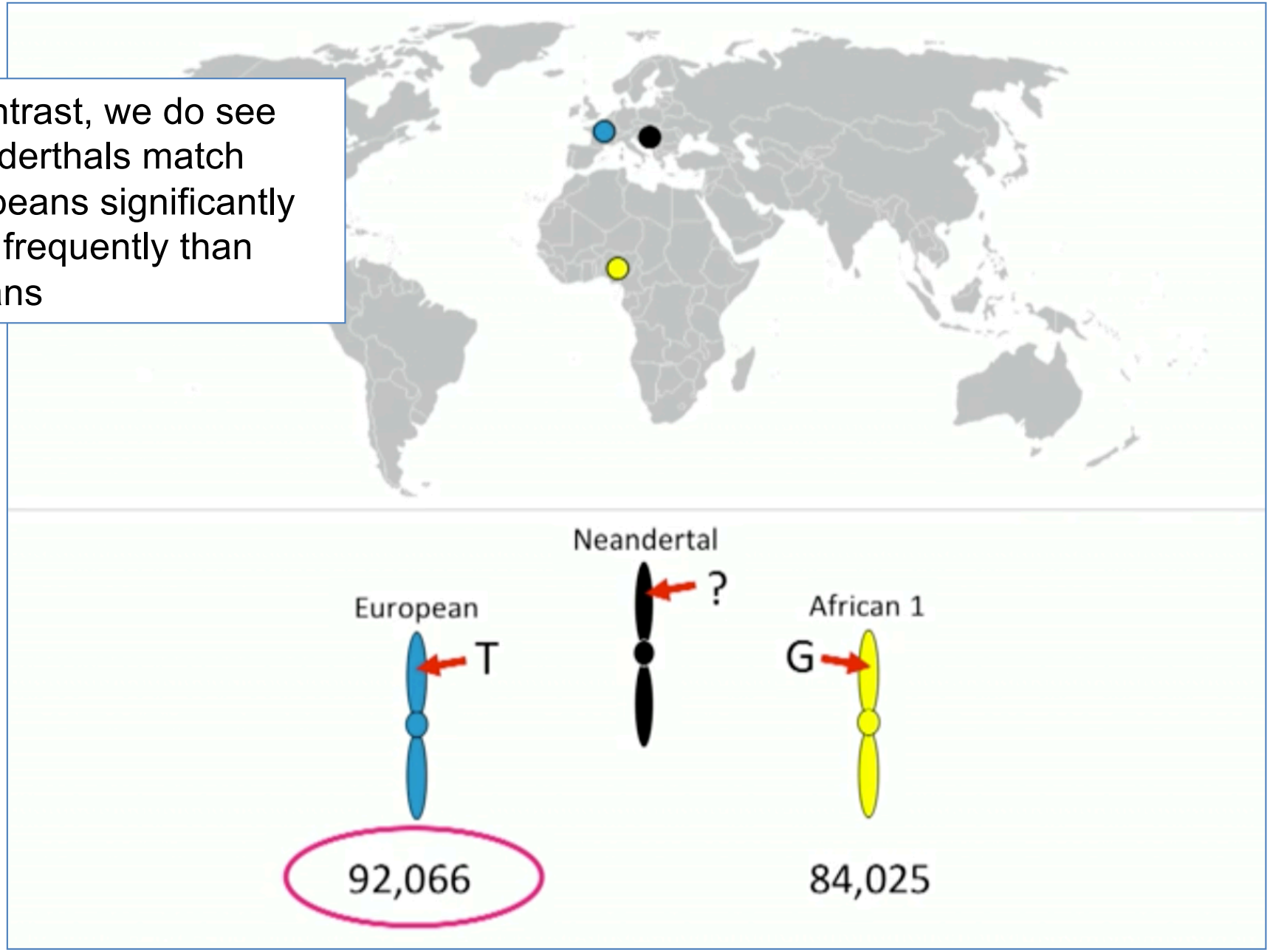
Did we mix?

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



Did we mix?

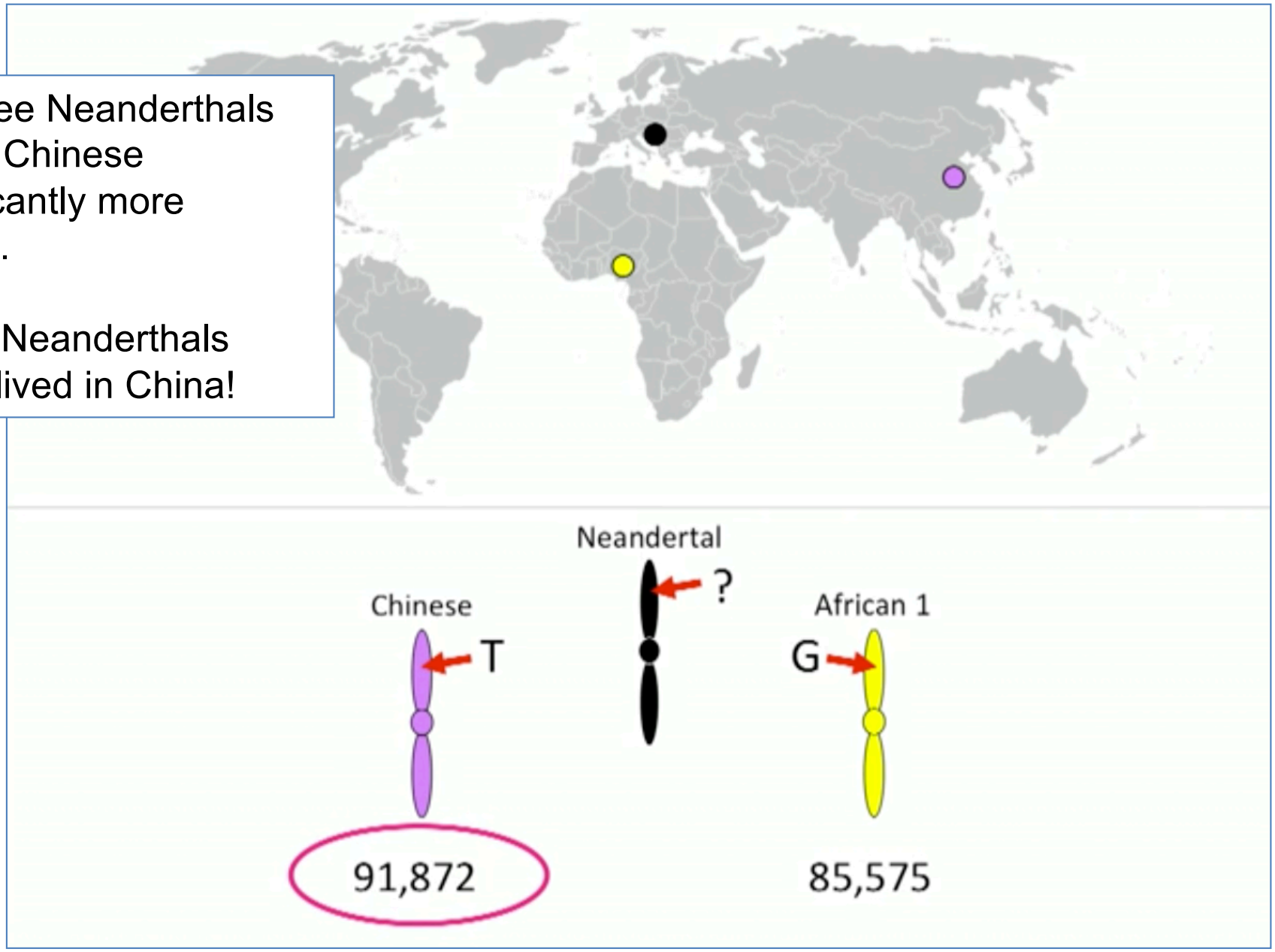
In contrast, we do see Neanderthals match Europeans significantly more frequently than Africans



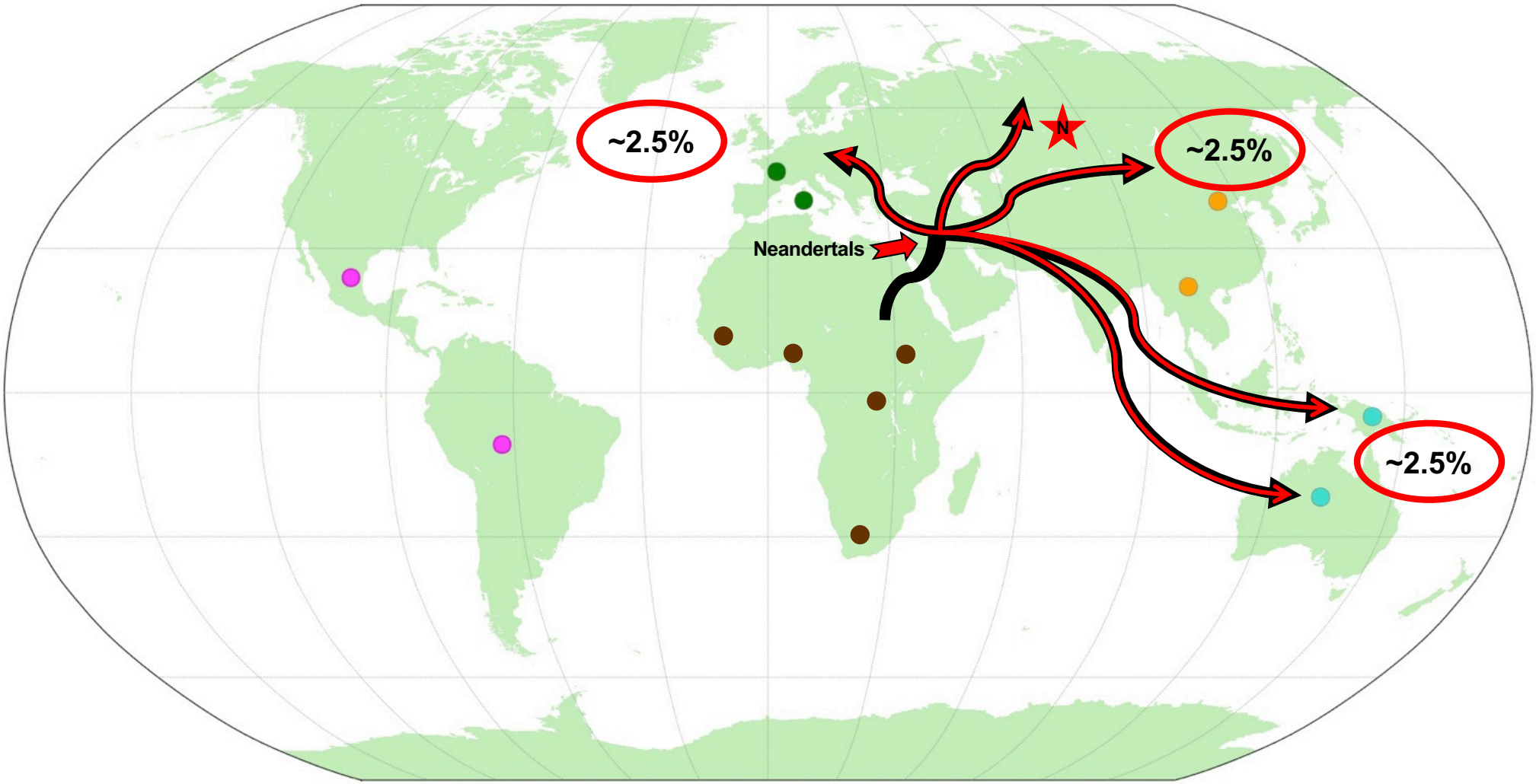
Did we mix?

Also see Neanderthals match Chinese significantly more often…

… but Neanderthals never lived in China!



Neanderthal Interbreeding

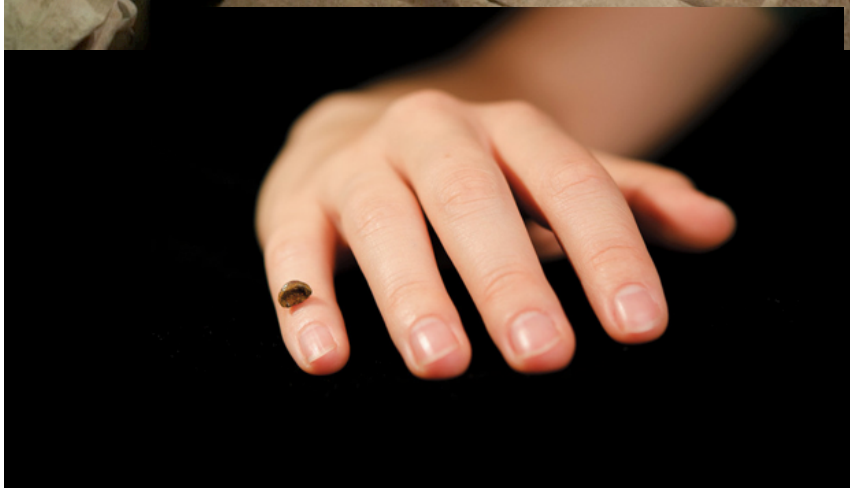


As modern humans migrated out of Africa, they apparently interbred with Neanderthal's so we see their alleles across the rest of the world and carry about 2.5% of their genome with us!

What about other ancient hominids?

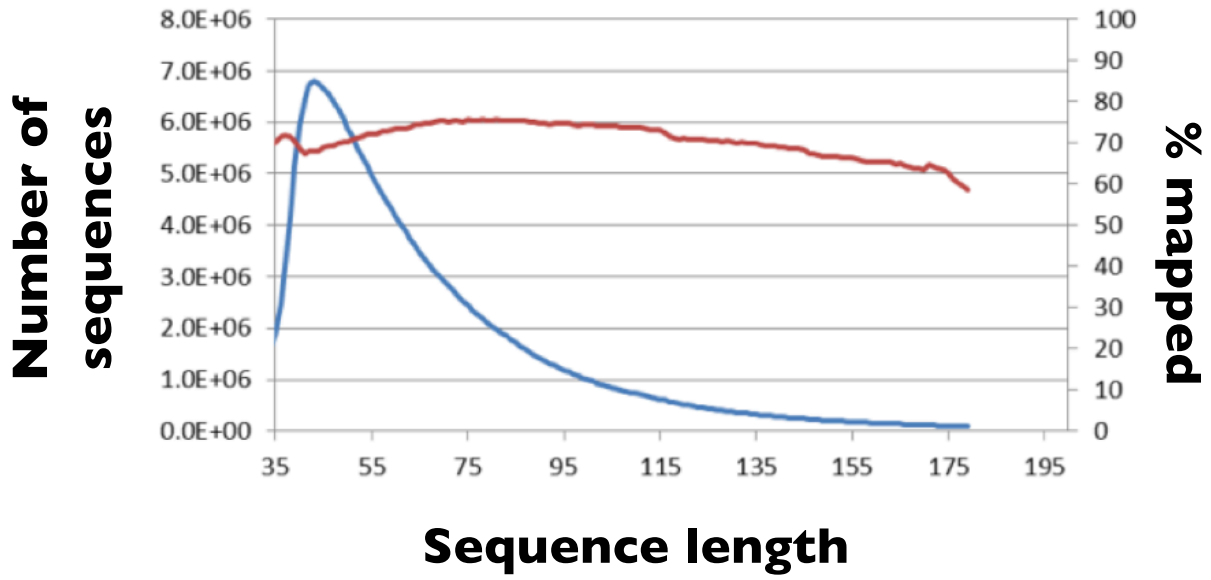
Denisova cave
Altai mountains
Russia

Academician A.P. Derevianko



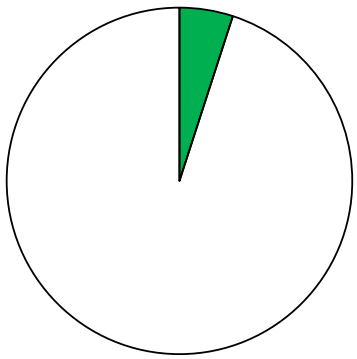

UNITED STATES OF AMERICA
E·PLURIBUS
·UNUM·
ONE CENT

Extraordinary preservation

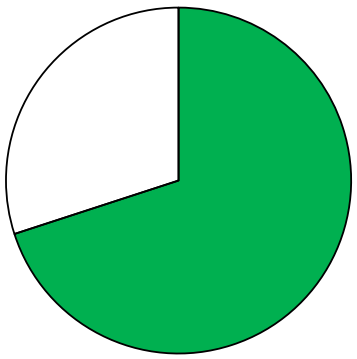


5% endogenous DNA

Best Neandertal bone

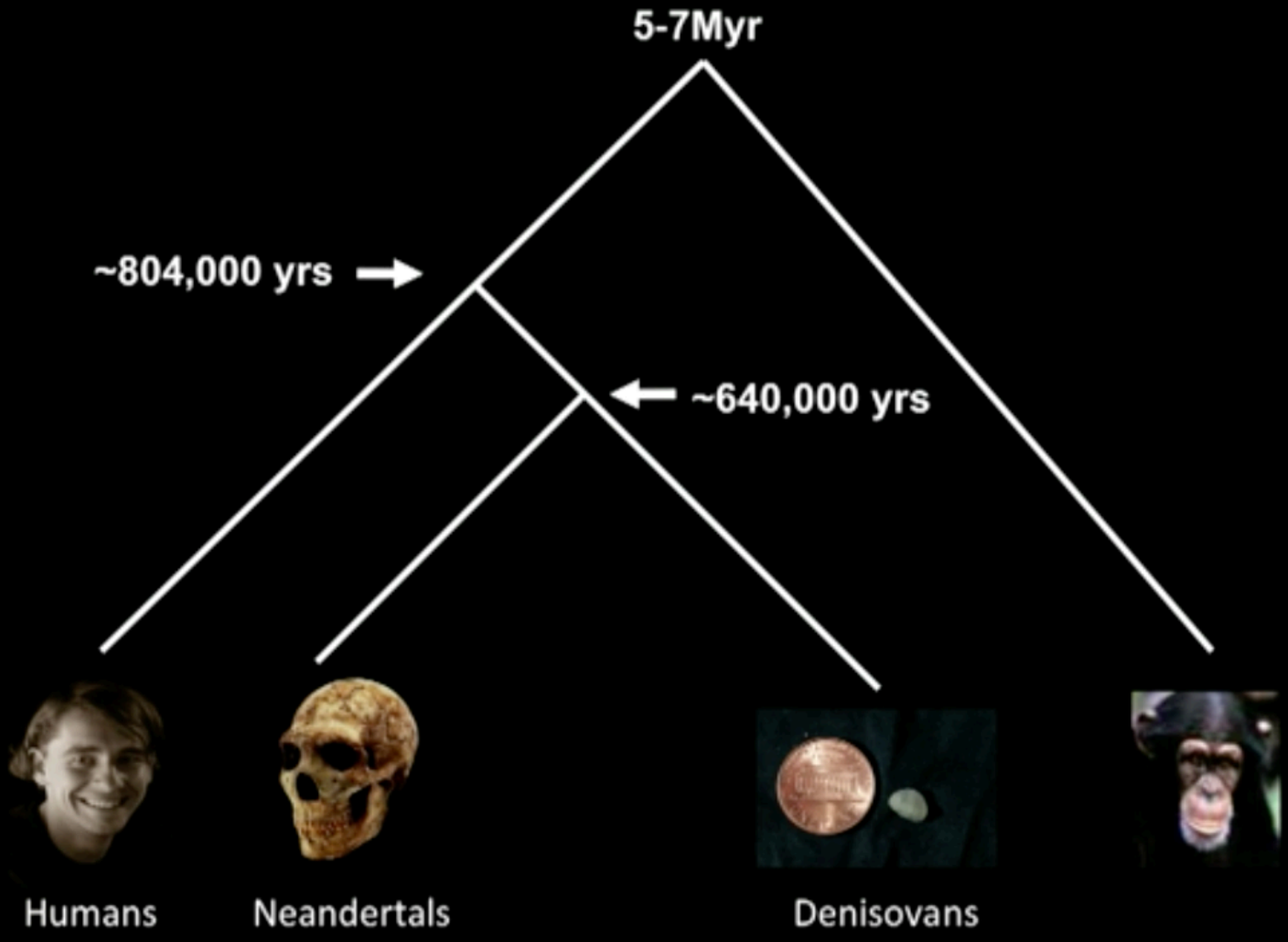


>70% endogenous DNA



Denisova bone

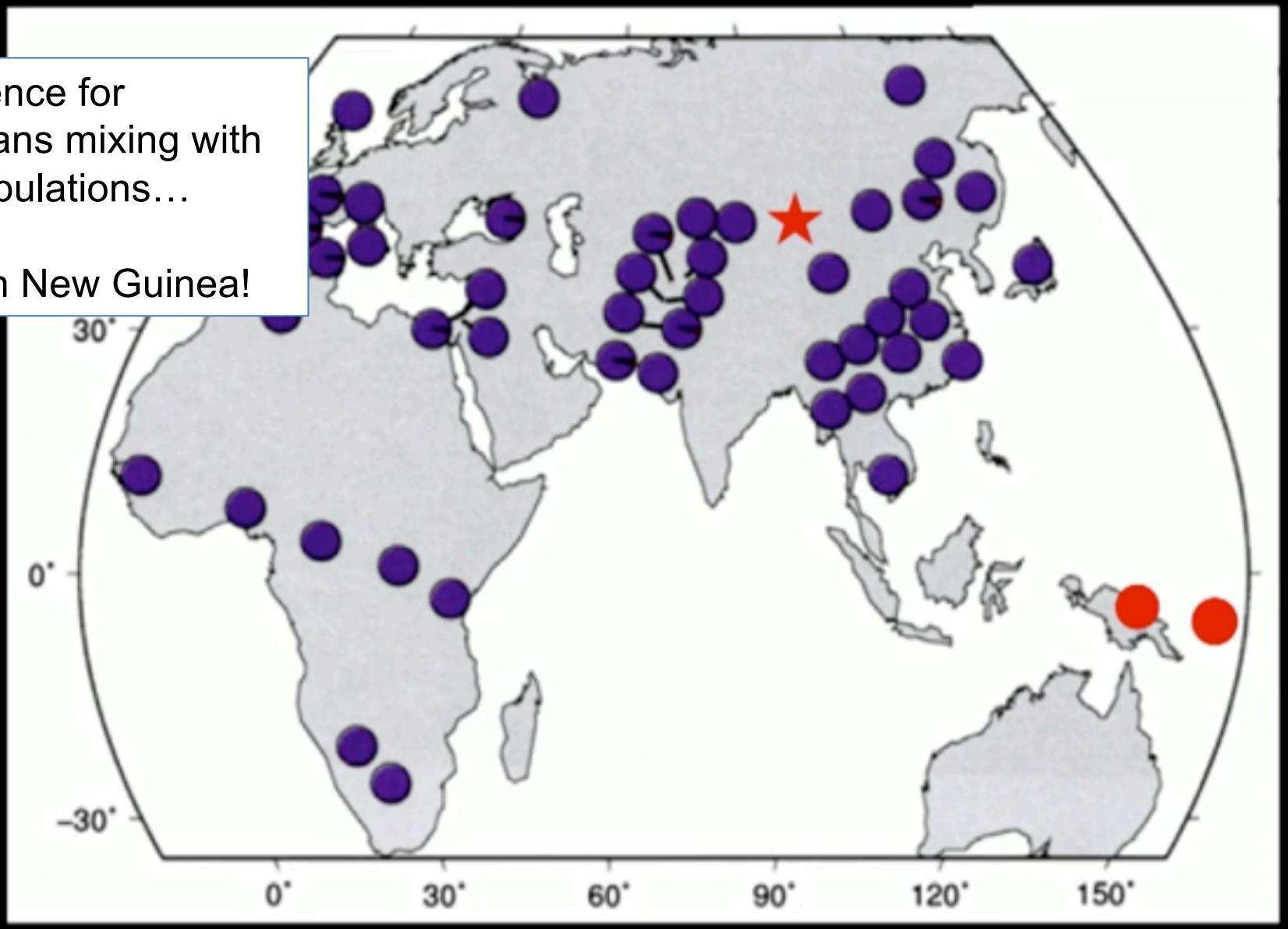
Denisovans & Neandertals



Did we mix?

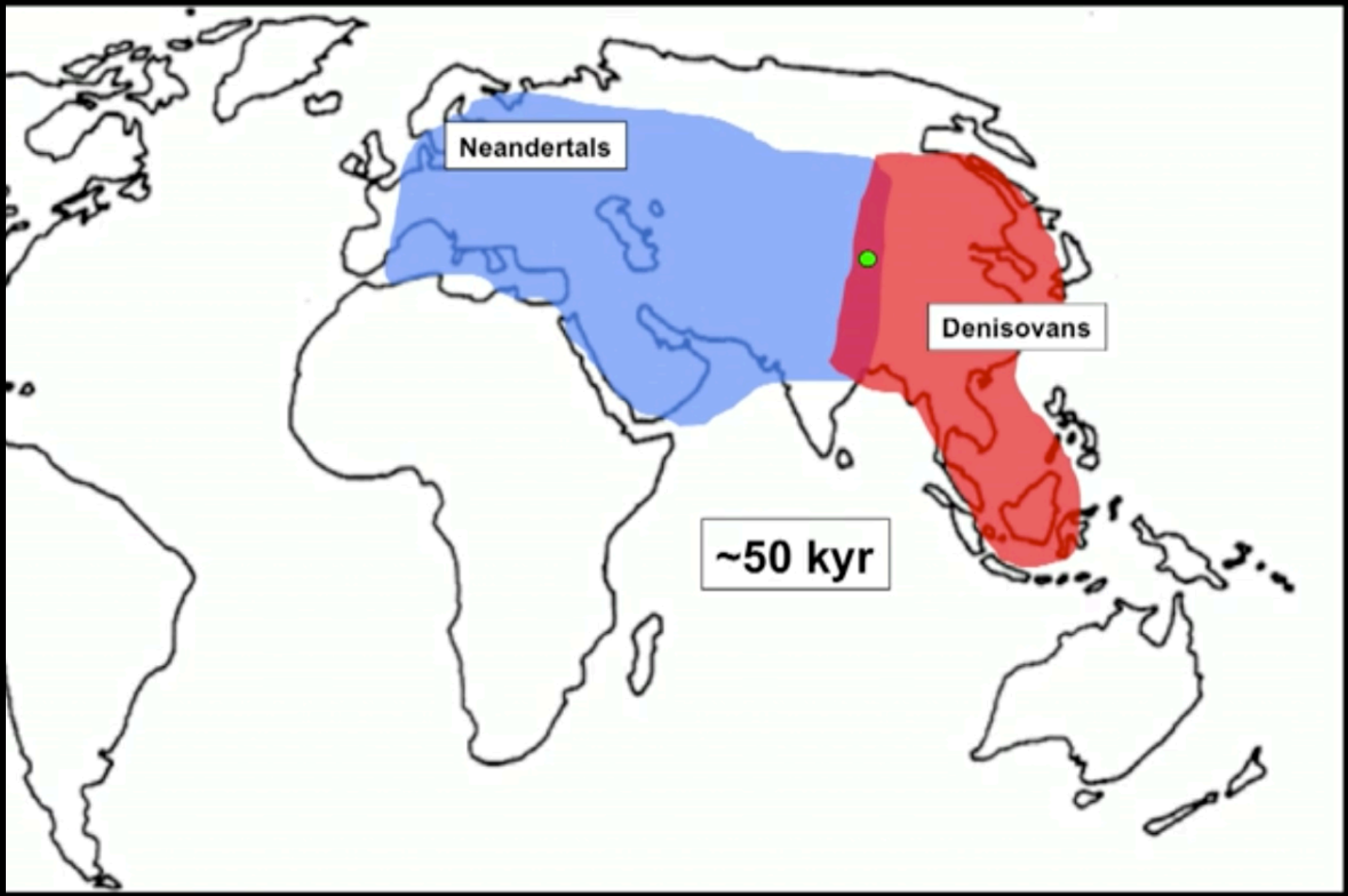
No evidence for Denisovans mixing with other populations…

Except in New Guinea!

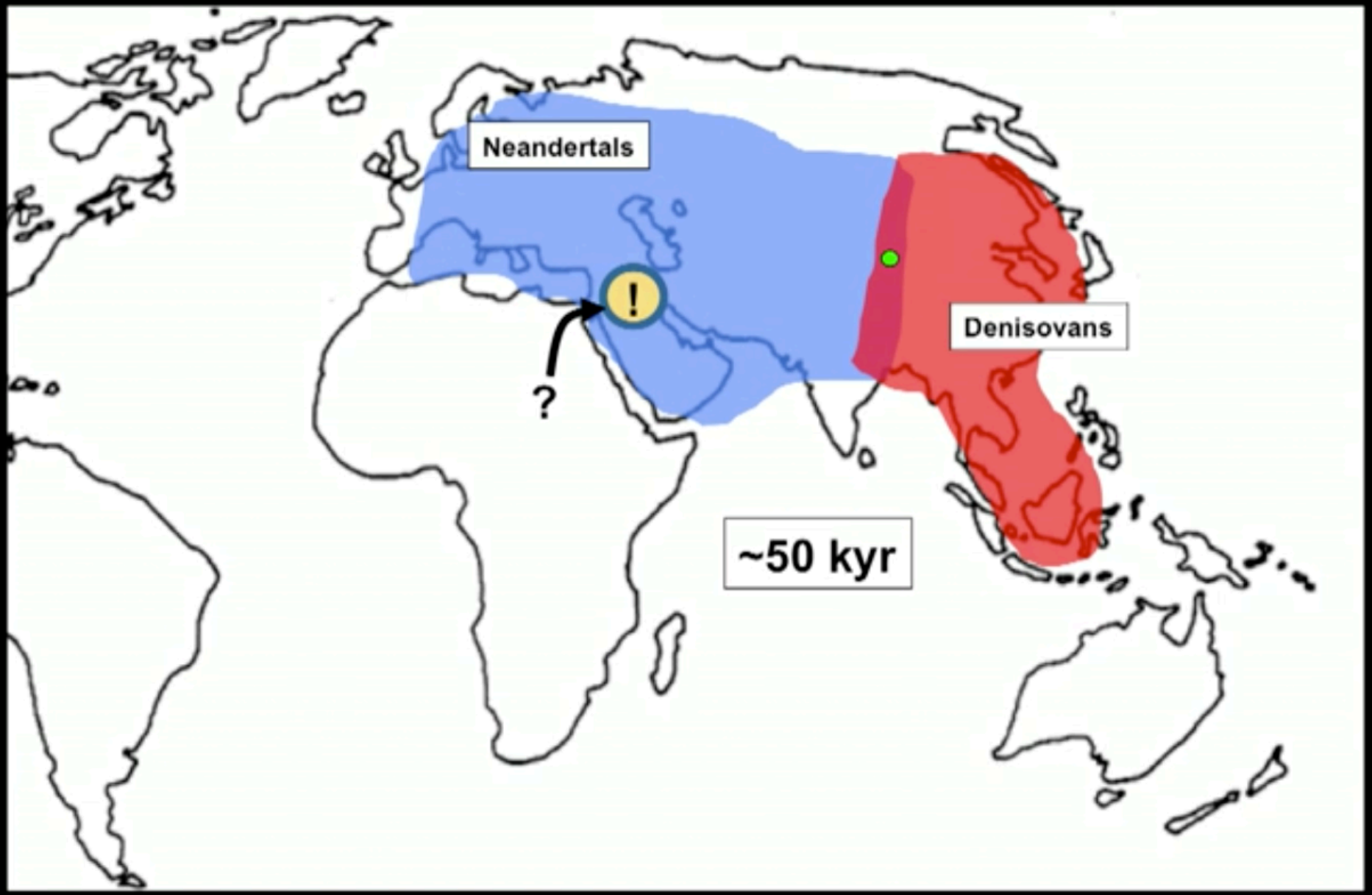


Map after Pickrell et al., 2009

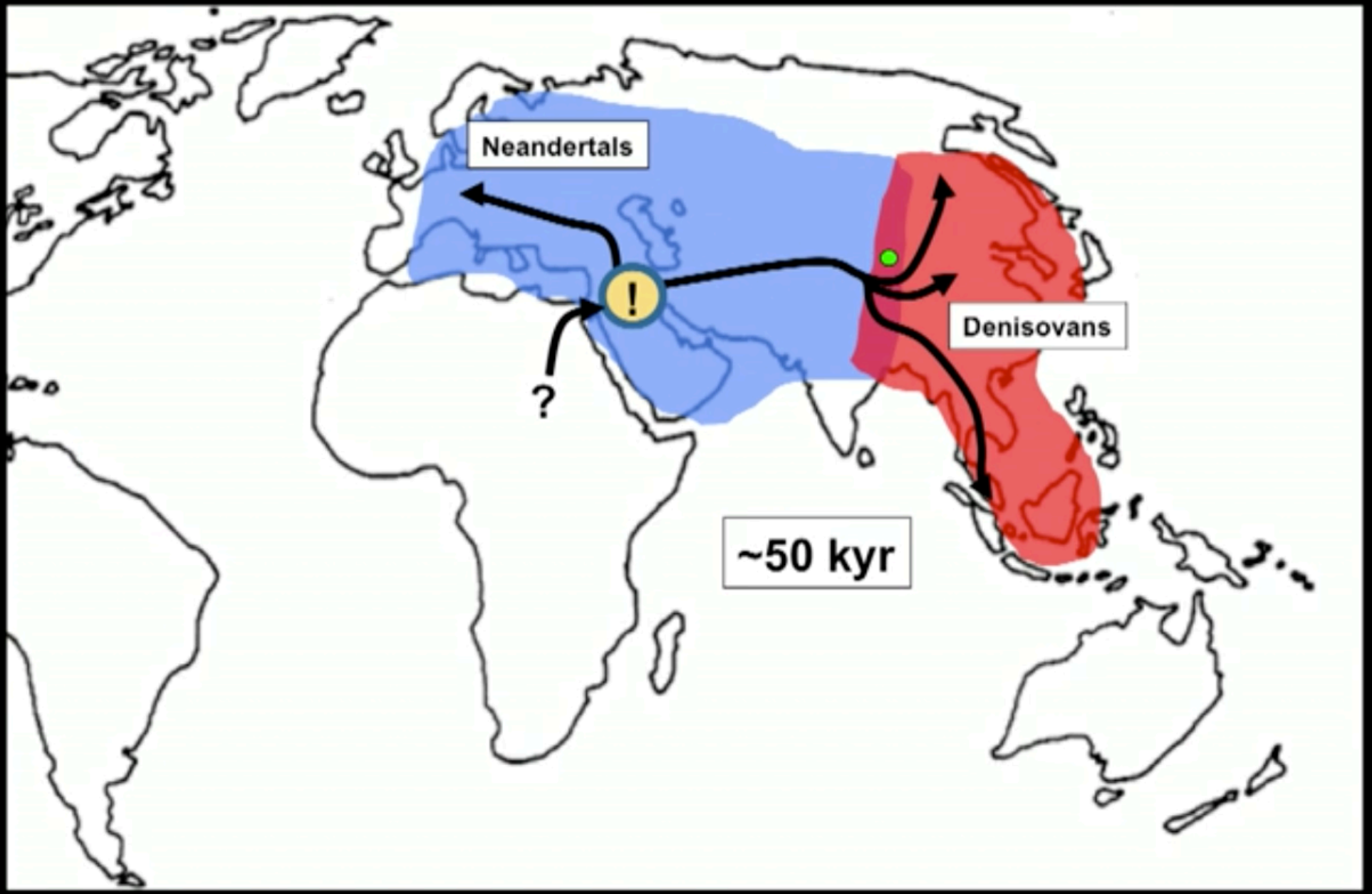
Timeline of ancient hominids



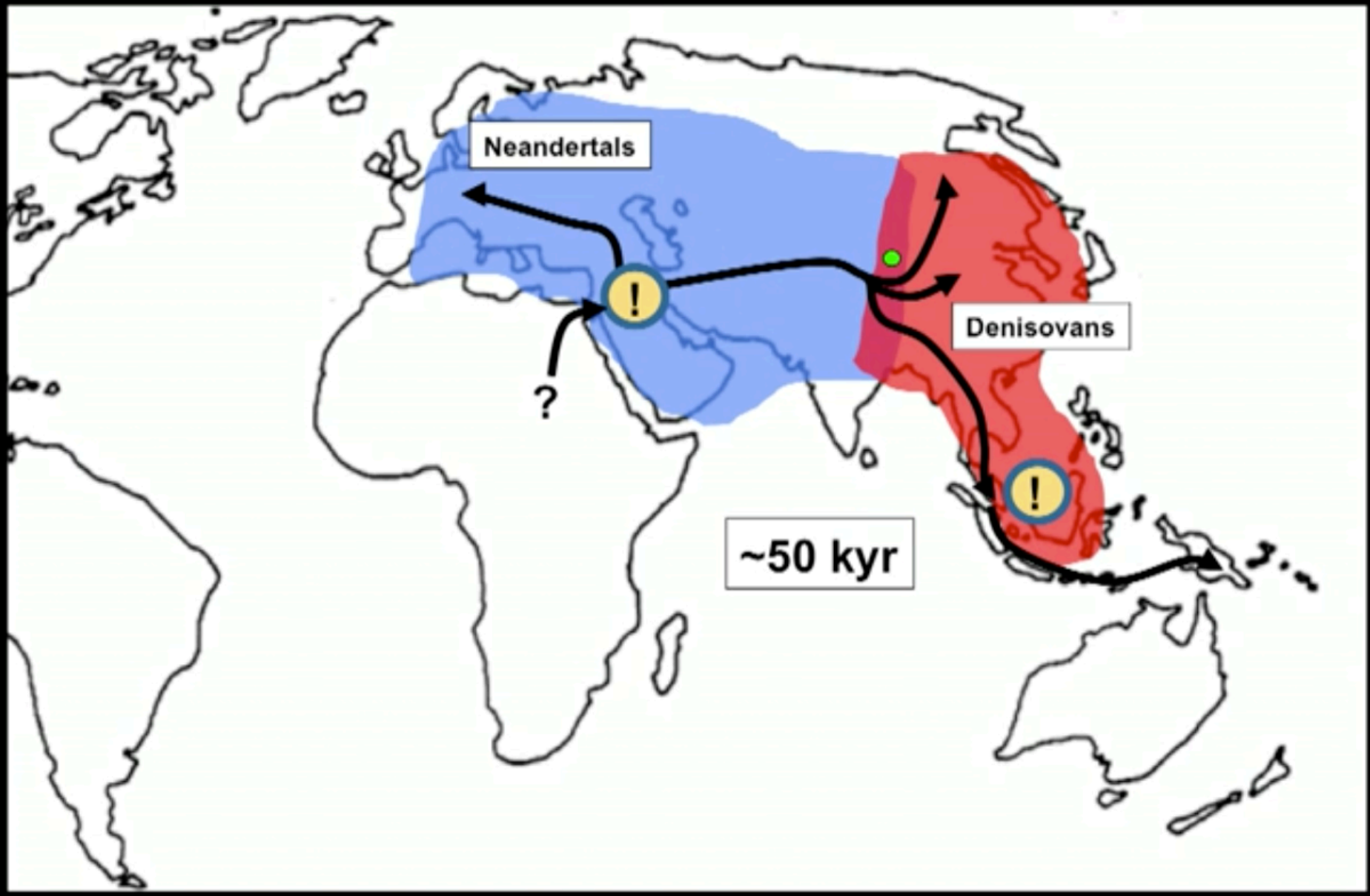
Timeline of ancient hominids



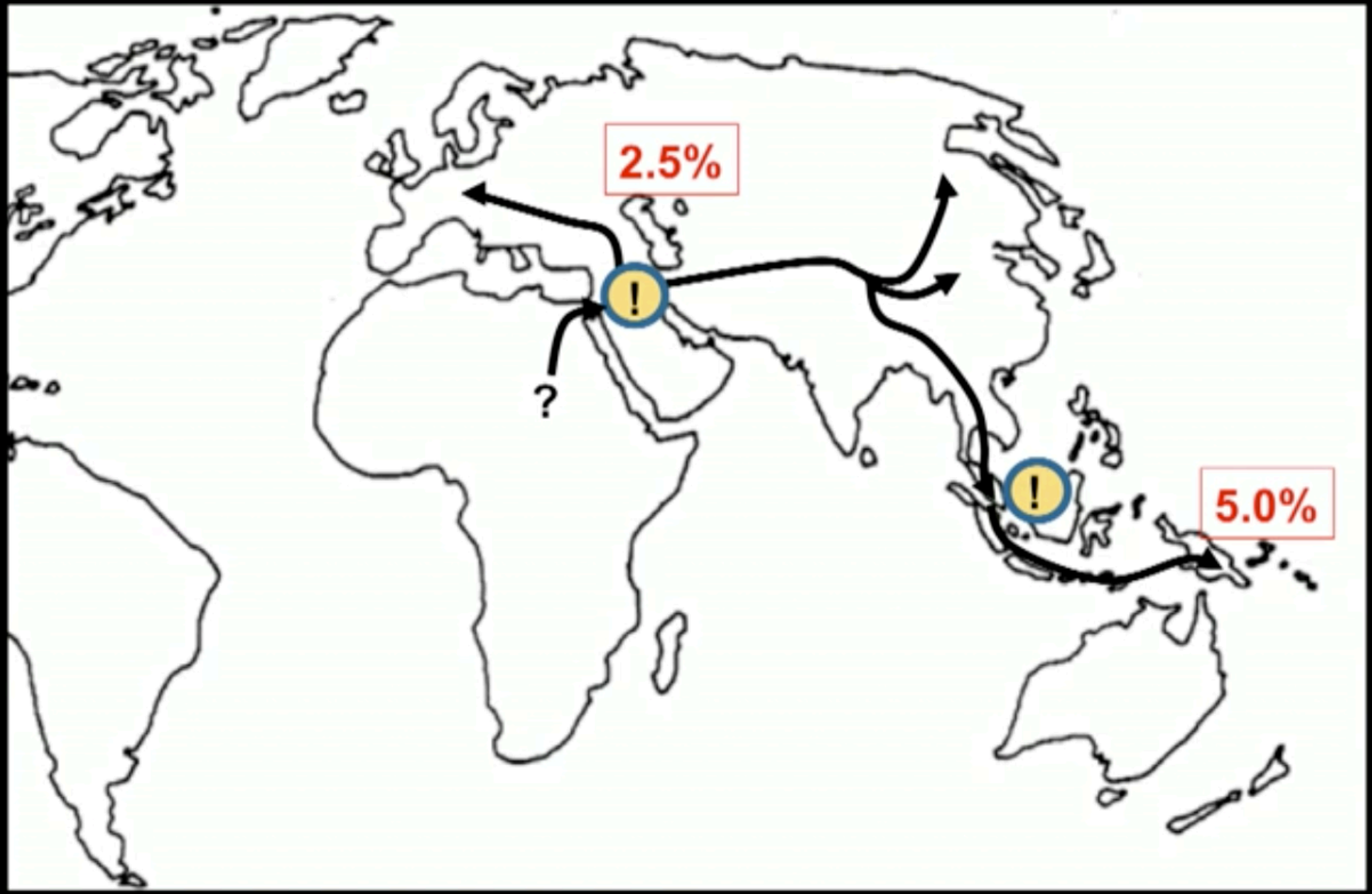
Timeline of ancient hominids



Timeline of ancient hominids



Timeline of ancient hominids



We have always mixed!

Science

REPORTS

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

Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals

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

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

FOXP2 Analysis

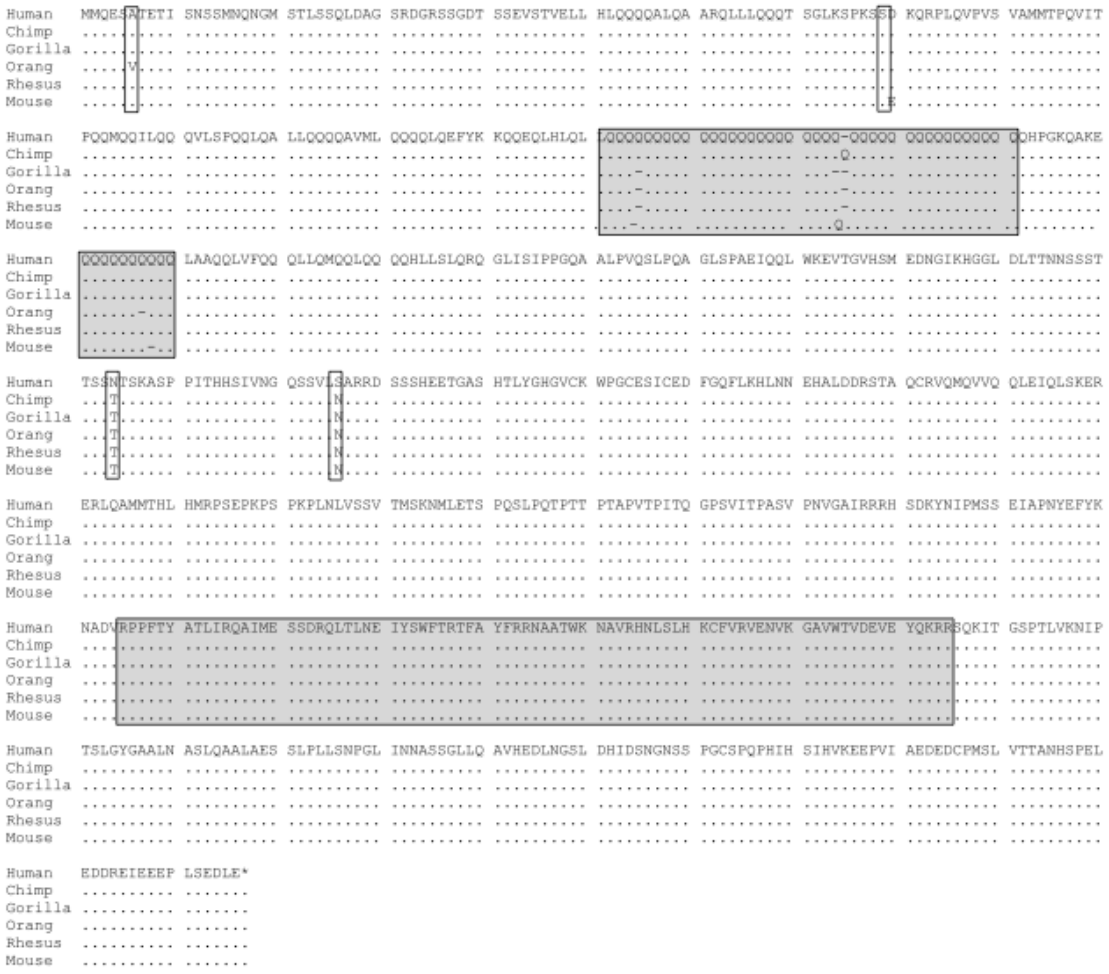


Figure 1 Alignment of the amino-acid sequences inferred from the *FOXP2* cDNA sequences. The polyglutamine stretches and the forkhead domain are shaded. Sites that differ from the human sequence are boxed.

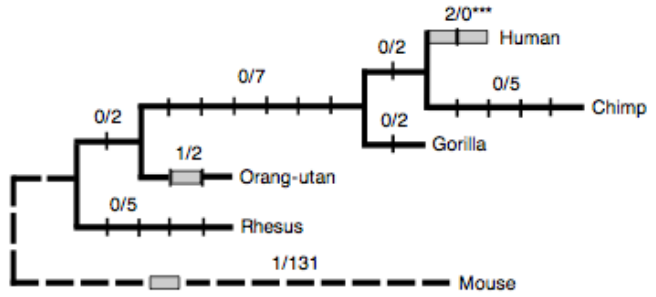
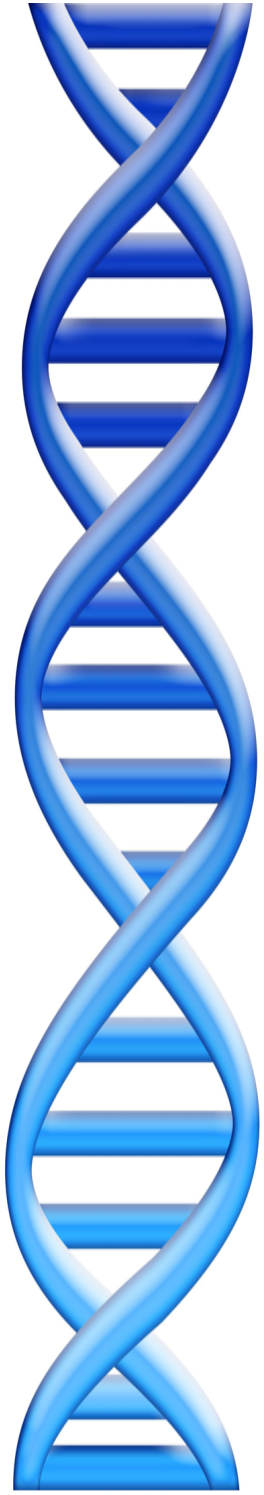


Figure 2 Silent and replacement nucleotide substitutions mapped on a phylogeny of primates. Bars represent nucleotide changes. Grey bars indicate amino-acid changes.

- Mutations of FOXP2 cause a severe speech and language disorder in people
- Versions of FOXP2 exist in similar forms in distantly related vertebrates; functional studies of the gene in mice and in songbirds indicate that it is important for modulating plasticity of neural circuits.
- Outside the brain FOXP2 has also been implicated in development of other tissues such as the lung and gut.

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

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



Part II:
Modern Humans

ARTICLE

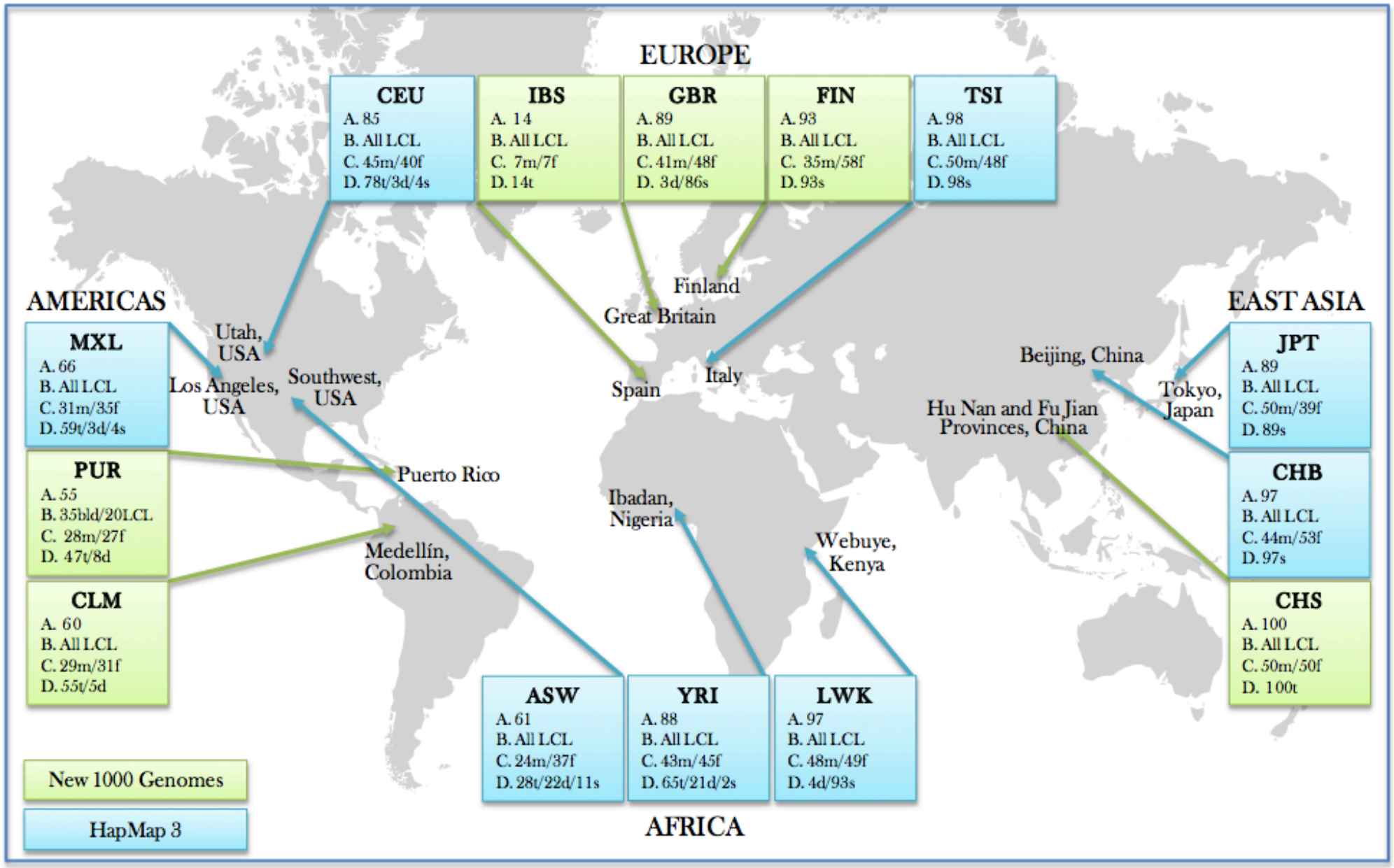
doi:10.1038/nature11632

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

The 1000 Genomes Project Consortium*

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

1000 Genomes Populations



1000 Genomes Populations

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

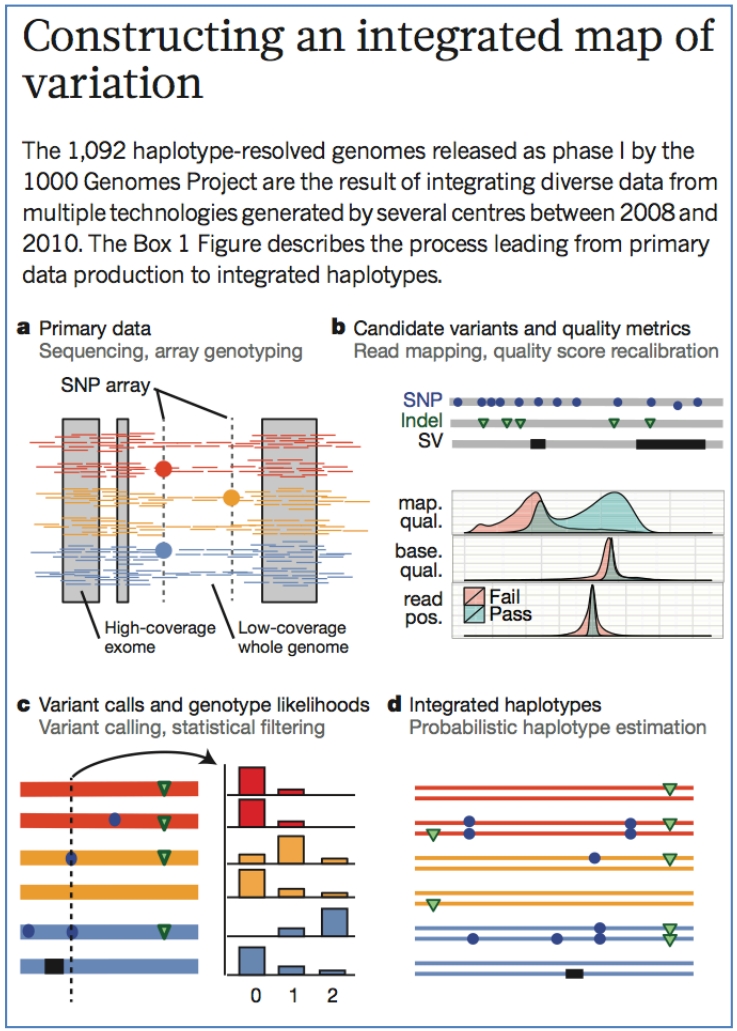
26 populations from 5 major population groups

1000 Genomes: Human Mutation Rate

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

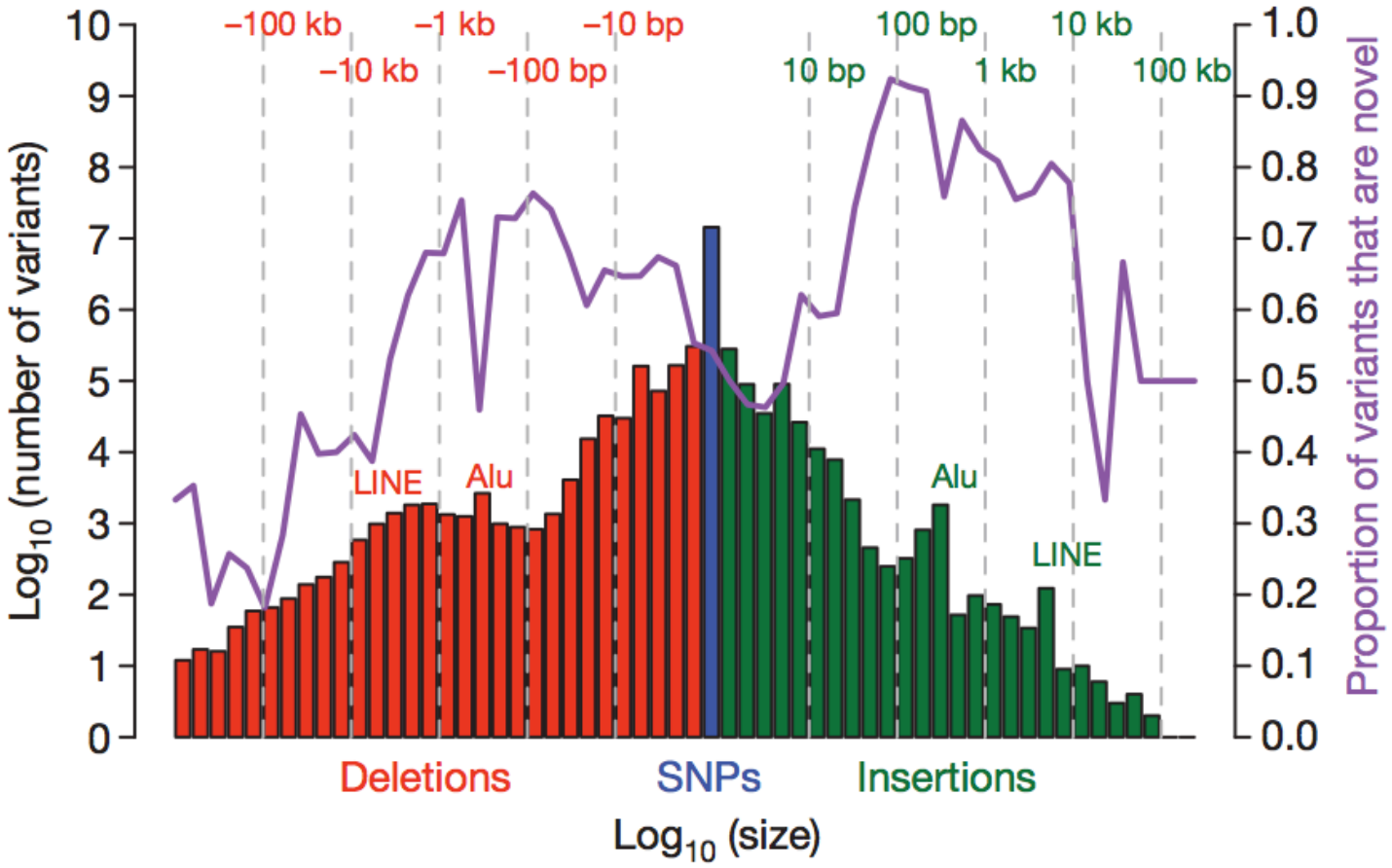
- Overall SNP rate between any two people is ~1/1200bp to ~1/1300
  - ~3M SNPs between me and you (.1%)
  - ~30M SNPs between human to Chimpanzees (1%)

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



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

Human Mutation Types



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

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

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

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

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

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

Homozygous LoF Mutations

LETTER

doi:10.1038/nature22034

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

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

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

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

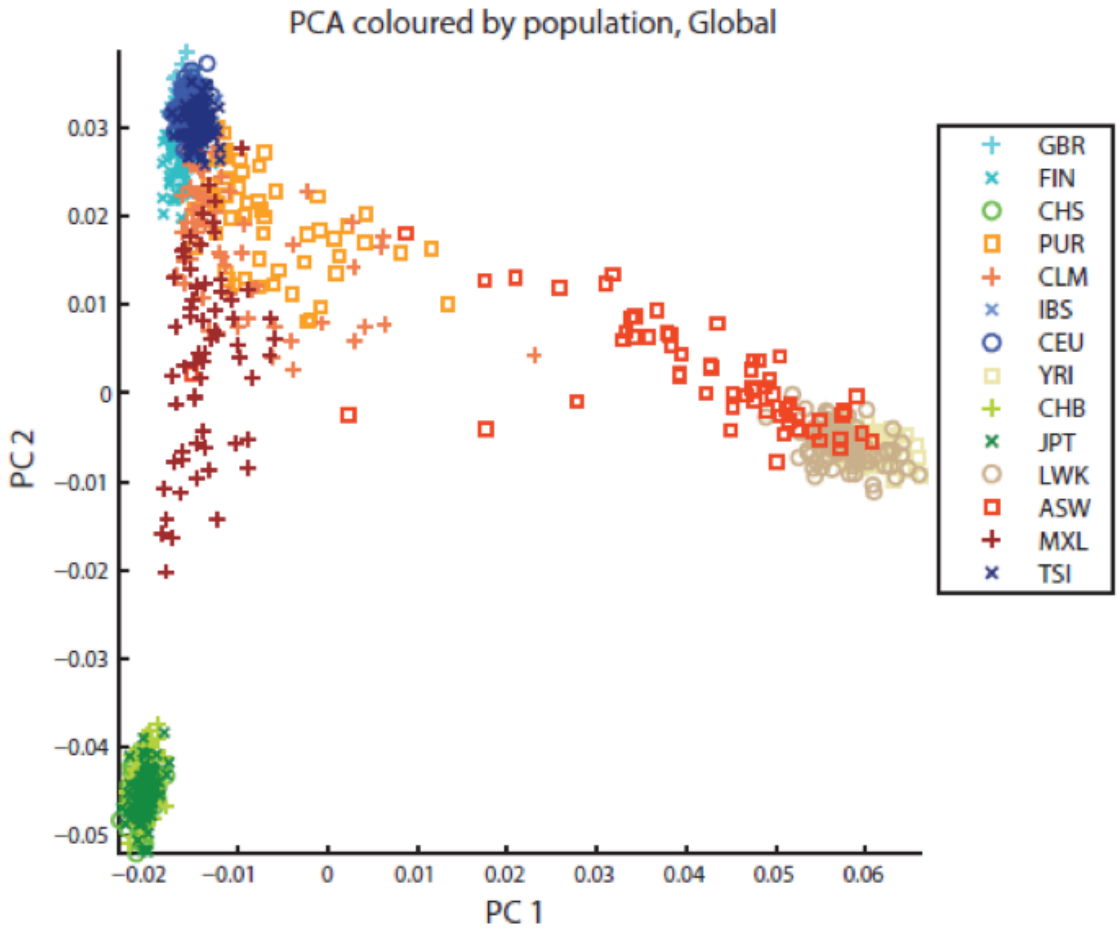
Across all 10,503 PROMIS participants, both copies of 1,317 distinct genes were predicted to be inactivated owing to pLoF mutations. A full listing of all 1,317 genes knocked out, the number of knockout participants for each gene, and the specific pLoF mutation(s) are provided in Supplementary Table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

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

(2017) Nature. doi:10.1038/nature22034

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

Variation across populations

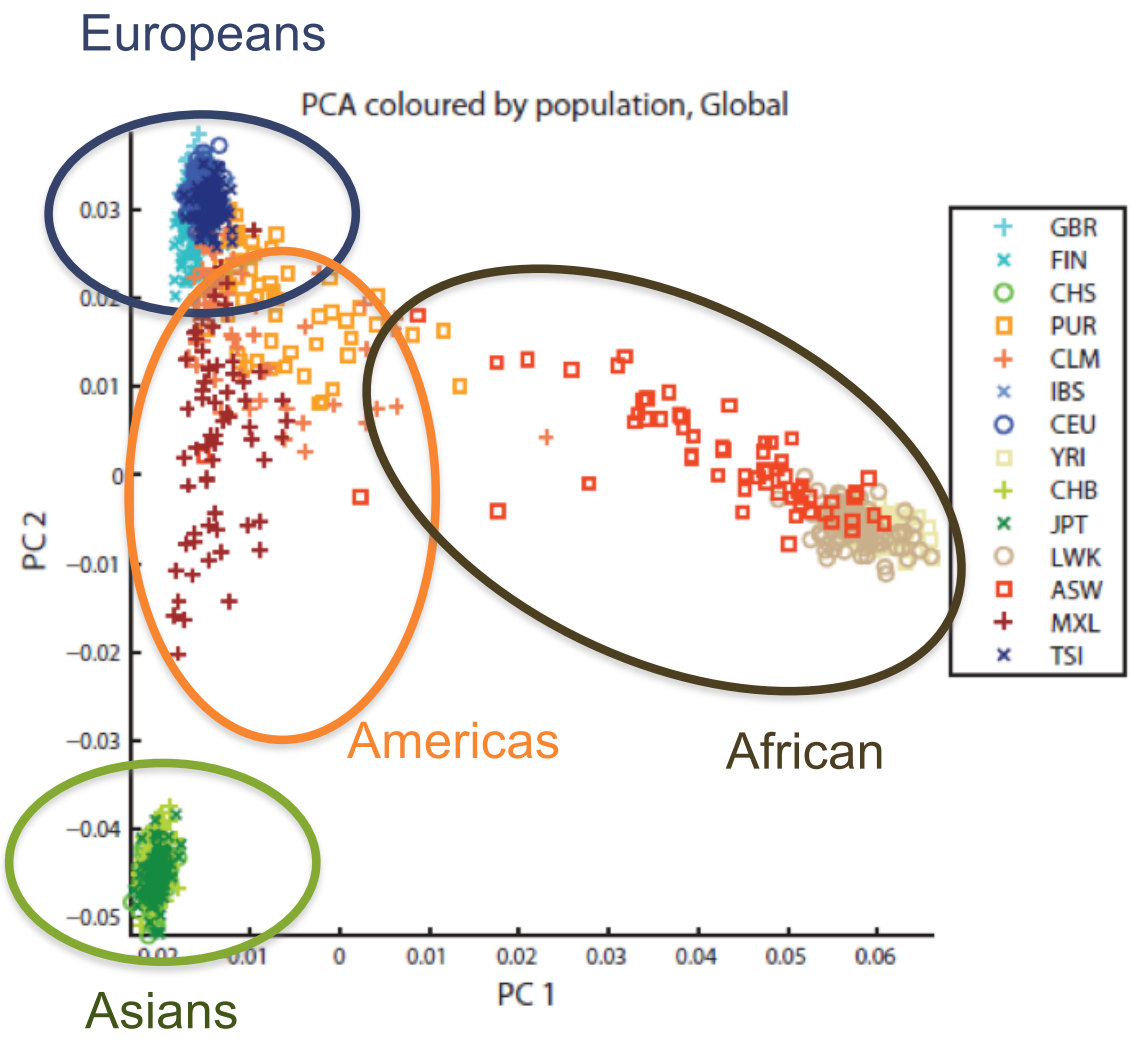


LEVEL	POP_PAIR	# of Highly differentiated SNPs	% in transcribed regions*
AFR	ASW-LWK	258	46.8
AFR	LWK-YRI	251	50.2
AFR	ASW-YRI	213	45.8
ASN	CHS-JPT	275	48.1
ASN	CHB-JPT	176	43.7
ASN	CHB-CHS	79	38.7
EUR	FIN-TSI	343	42.6
EUR	CEU-FIN	201	40.7
EUR	FIN-GBR	197	43.2
EUR	GBR-TSI	100	38.9
EUR	CEU-TSI	57	53.8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

Table S12A Summary of sites showing high levels of population differentiation

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

Variation across populations

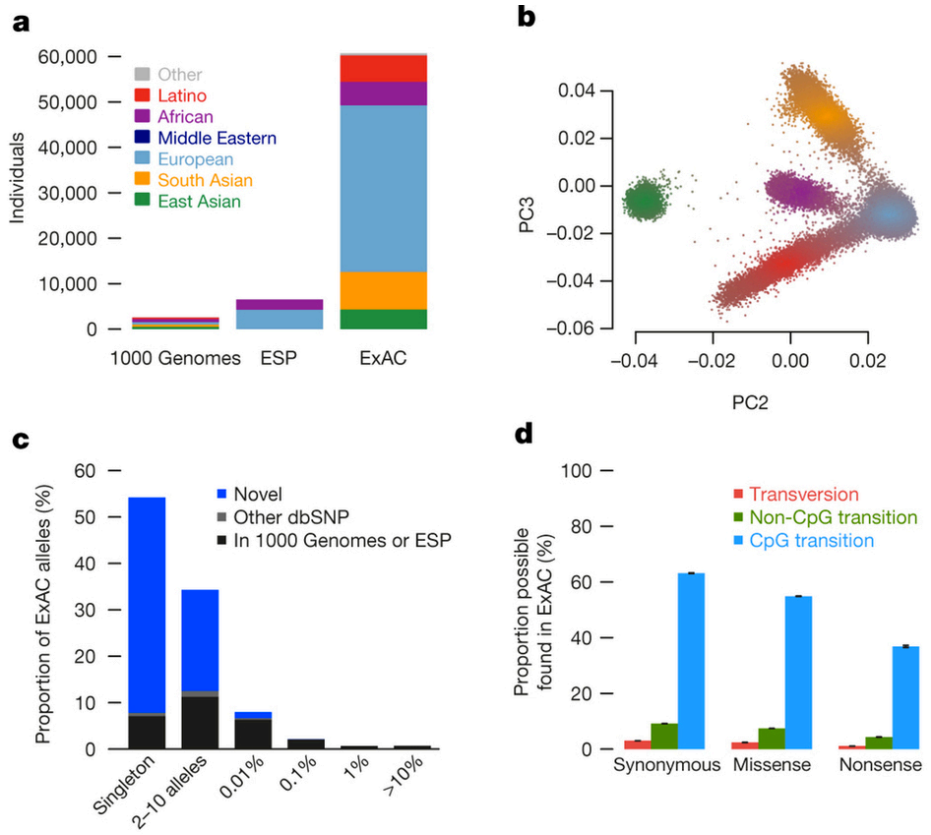


LEVEL	POP_PAIR	# of Highly differentiated SNPs	% in transcribed regions*
AFR	ASW-LWK	258	46.8
AFR	LWK-YRI	251	50.2
AFR	ASW-YRI	213	45.8
ASN	CHS-JPT	275	48.1
ASN	CHB-JPT	176	43.7
ASN	CHB-CHS	79	38.7
EUR	FIN-TSI	343	42.6
EUR	CEU-FIN	201	40.7
EUR	FIN-GBR	197	43.2
EUR	GBR-TSI	100	38.9
EUR	CEU-TSI	57	53.8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

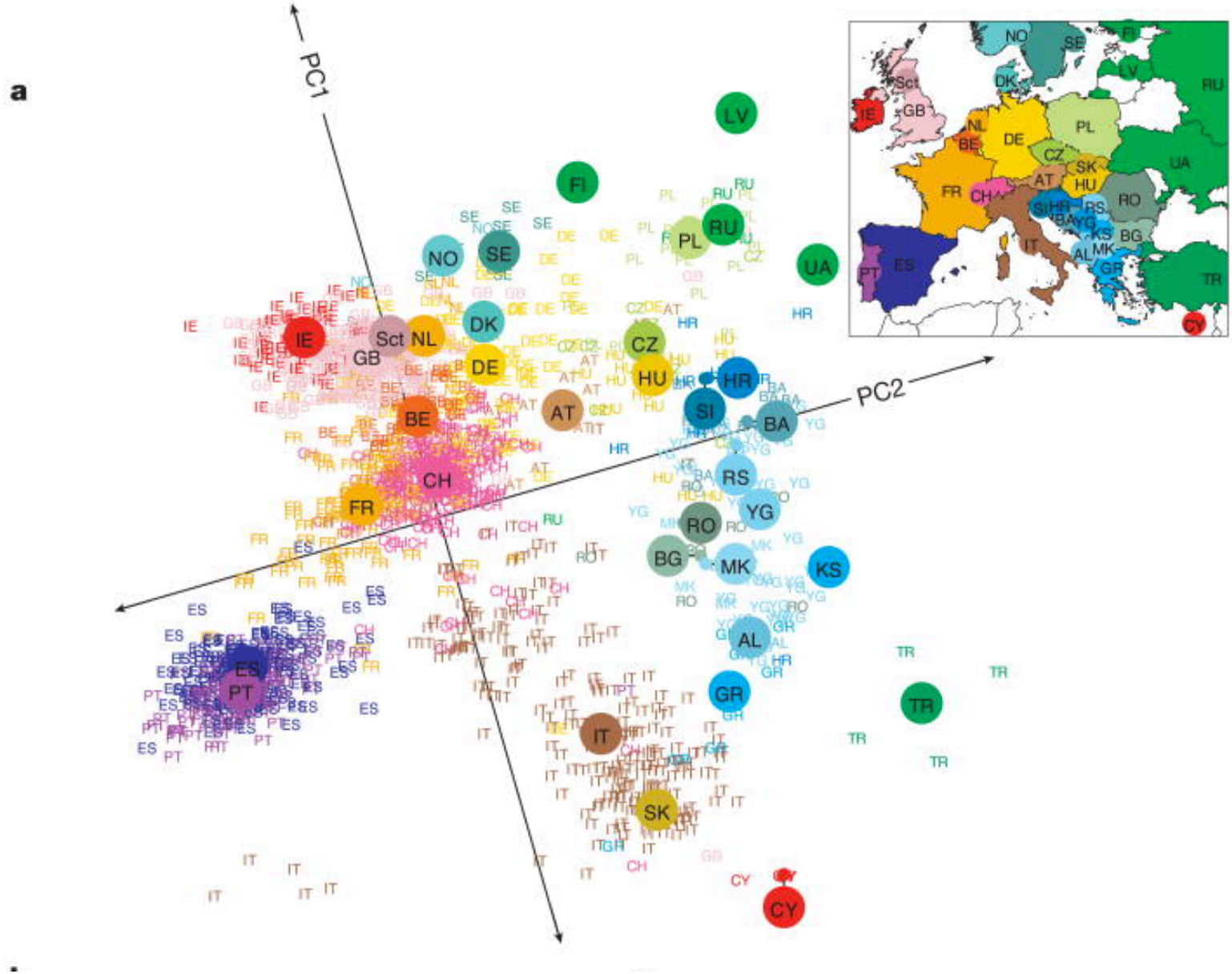
ExAC: Exome Aggregation Consortium



- 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

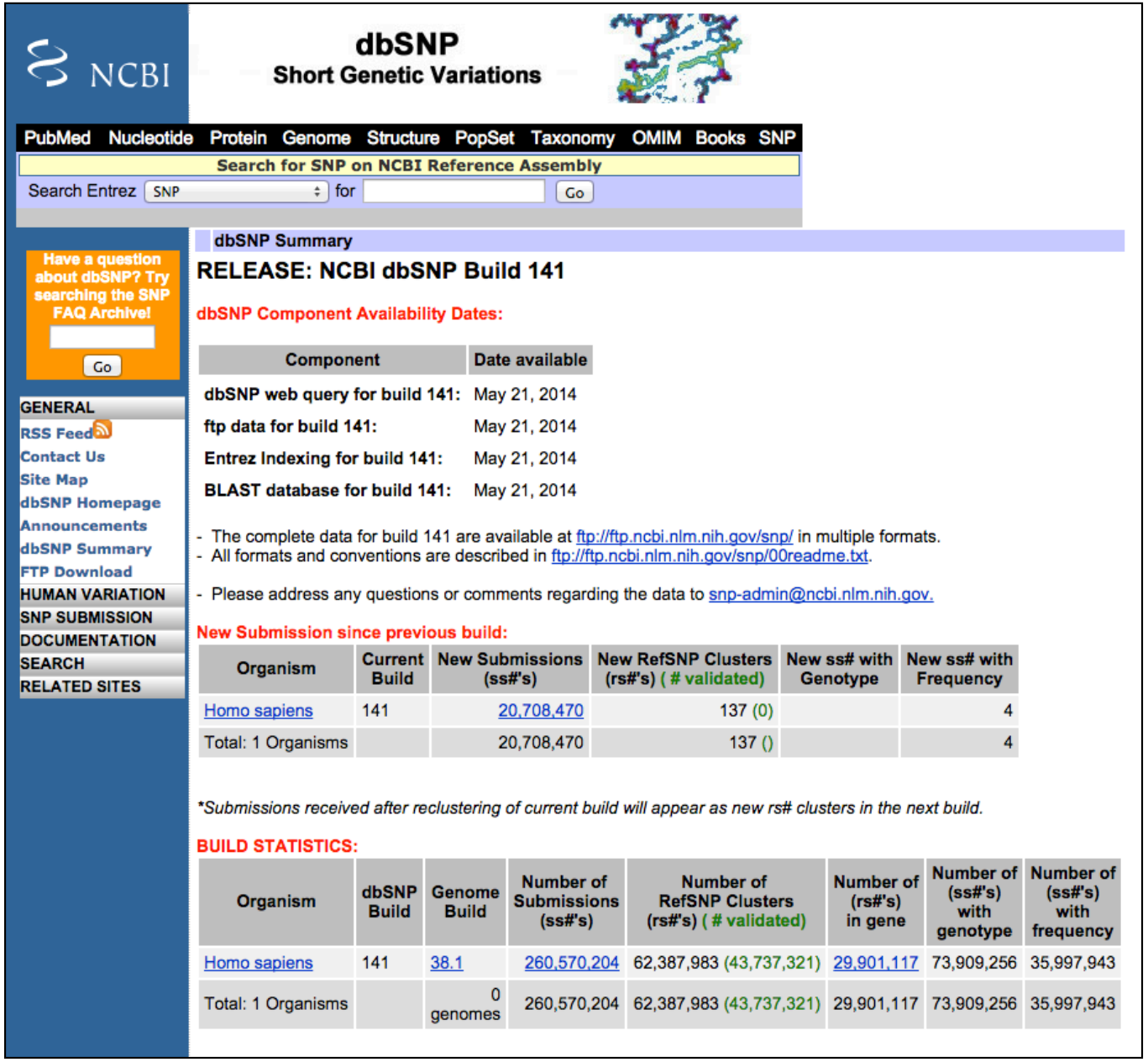
Lek et al (2016) Nature. doi:10.1038/nature19057



Genes mirror geography within Europe

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

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.