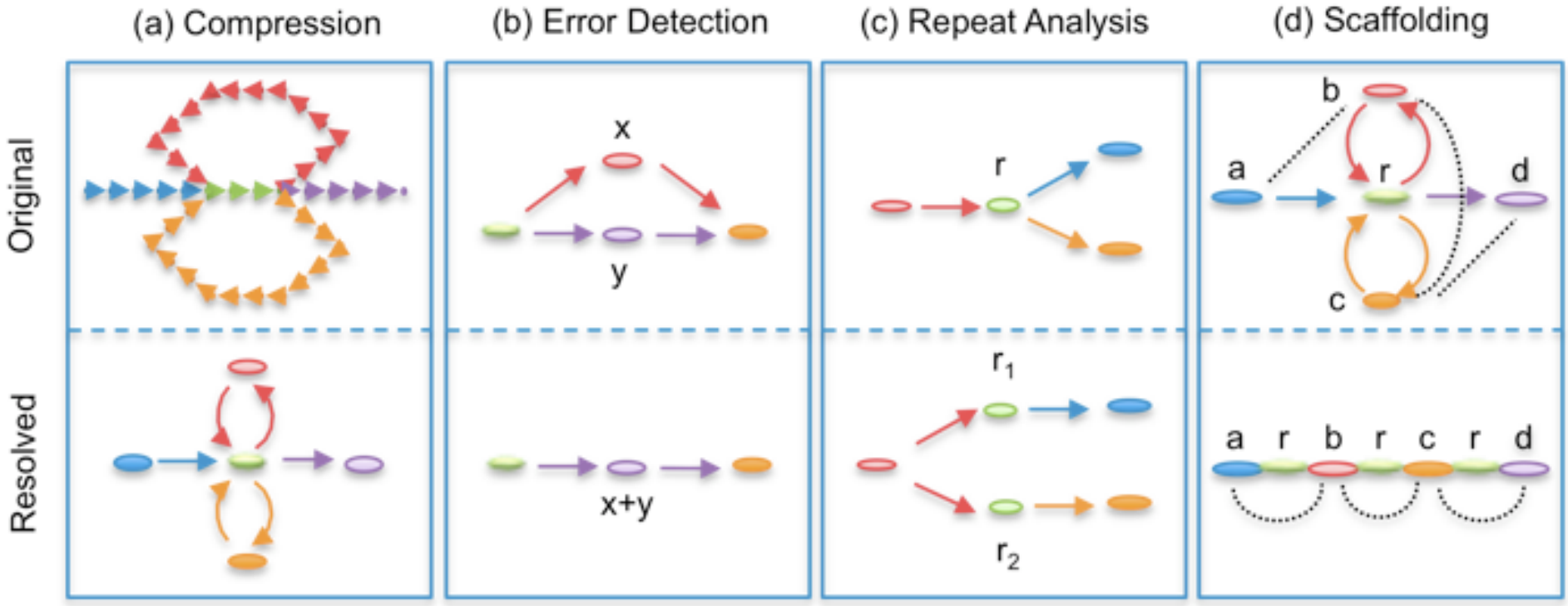
De Brujin Graph formulation



Quality of assemblies—mouse

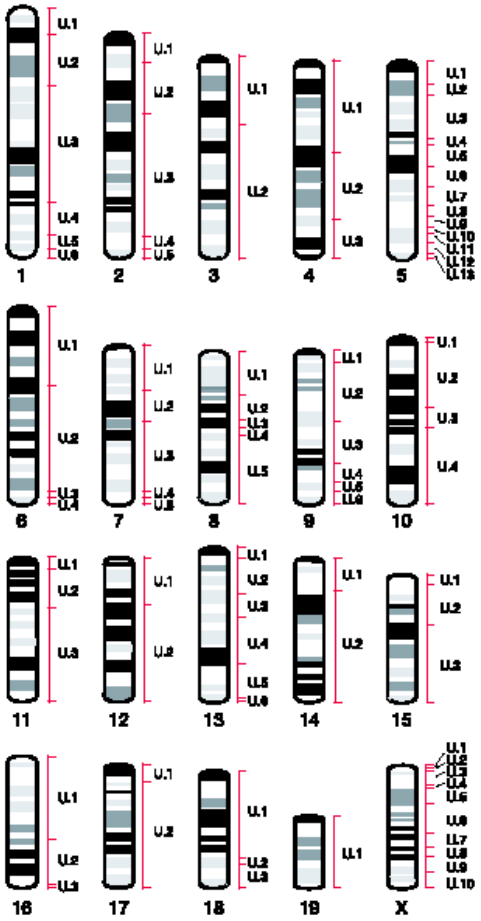


Figure 1 The mouse genome in 88 sequence-based ultracontigs. The position and extent of the 88 ultracontigs of the MGSCv3 assembly are shown adjacent to ideograms of the mouse chromosomes. All mouse chromosomes are acrocentric, with the centromeric end at the top of each chromosome. The supercontigs of the sequence assembly were anchored to the mouse chromosomes using the MIT genetic map. Neighbouring supercontigs were linked together into ultracontigs using information from single BAC links and the fingerprint and radiation-hybrid maps, resulting in 88 ultracontigs containing 95% of the bases in the euchromatic genome.

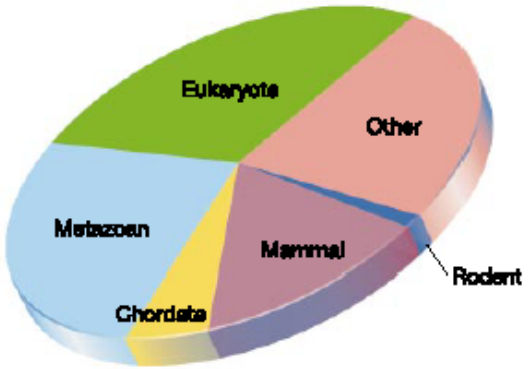
N50 length (kb)*	Bases (Gb)	Bases plus gaps (Gb)	Percentage of genome
25.9	2.372	2.372	94.9
18,600	2.372	2.477	99.1
50,600	2.372	2.493	99.7
2.3	0.106	0.106	–
18,700	2.352	2.455	98.2
22,900	1.955	2.039	81.6

des spanned gaps.
ercontigs with an N50 value of 3.4 kb. The N50 value for all contigs is 24.8 kb, and for all supercontigs is 16,900 kb (excluding gaps to gaps in the ultracontigs and are thus accounted for in the 'bases plus gaps' estimate.

Terminology: *N50 contig length*

If we sort contigs from largest to smallest, and start Covering the genome in that order, N50 is the length Of the contig that just covers the 50th percentile.

7.7X sequence coverage



Panda Genome

Table 1 | Summary of the panda genome sequencing and assembly

Step	Paired-end insert size (bp)*	Sequence coverage (×)†	Physical coverage (×)†	N50 (bp) ‡	N90 (bp) ‡	Total length (bp)
Initial contig	110–230; 380–570	38.5	96	1,483	224	2,021,639,596
Scaffold 1				32,648	7,780	2,213,848,409
Scaffold 2	Add 1,700–2,800	8.4	151	229,150	45,240	2,250,442,210
Scaffold 3	Add 3,700–7,500	6.5	450	581,933	127,336	2,297,100,301
Scaffold 4	Add 9,200–12,300	2.6	373	1,281,781	312,670	2,299,498,912
Final contig	All	56.0	1,070	39,886	9,848	2,245,302,481

Add denotes accumulative; for example, scaffold 2 uses data of 110–230, 380–570 and 1,700–2,800.

* Approximate average insert size of Illumina Genome Analyser sequencing libraries. The sizes were estimated by mapping the reads onto the assembled genome sequences.

† High-quality read sequences that were used in assembly. Coverage was estimated assuming a genome size of 2.4 Gb. Sequence coverage refers to the total length of generated reads, and physical coverage refers to the total length of sequenced clones of the libraries.

‡ N50 size of contigs or scaffolds was calculated by ordering all sequences then adding the lengths from longest to shortest until the summed length exceeded 50% of the total length of all sequences. N90 is similarly defined.

Assemblathon

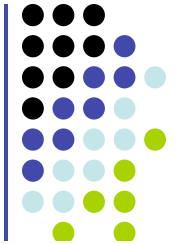


Table 2 Overview of sequencing data provided for Assemblathon 2 participants

Species	Estimated genome size	Illumina	Roche 454	Pacific biosciences
Bird (*Melopsittacus undulatus*)	1.2 Gbp	285x coverage from 14 libraries (mate pair and paired-end)	16x coverage from 3 library types (single end and paired-end)	10x coverage from 2 libraries
Fish (*Maylandia zebra*)*	1.0 Gbp	192x coverage from 8 libraries (mate pair and paired-end)	NA	NA
Snake (*Boa constrictor constrictor*)	1.6 Gbp	125x coverage from 4 libraries (mate pair and paired-end)	NA	NA

Table 1 Assemblathon 2 participating team details

Team name	Team identifier	Number of assemblies submitted			Sequence data used for bird assembly	Institutional affiliations	Principal assembly software used
		Bird	Fish	Snake			
ABL	ABL	1	0	0	4 + I	Wayne State University	HyDA
ABySS	ABYSS	0	1	1		Genome Sciences Centre, British Columbia Cancer Agency	ABySS and Anchor
Allpaths	ALLP	1	1	0	I	Broad Institute	ALLPATHS-LG
BCM-HGSC	BCM	2	1	1	4 + I + P[1]	Baylor College of Medicine Human Genome Sequencing Center	SeqPrep, KmerFreq, Quake, BWA, Newbler, ALLPATHS-LG, Atlas-Link, Atlas-GapFill, Phrap, CrossMatch, Velvet, BLAST, and BLASR
CBCB	CBCB	1	0	0	4 + I + P	University of Maryland, National Biodefense Analysis and Countermeasures Center	Celera assembler and PacBio Corrected Reads (PBcR)
CoBiG[2]	COBIG	1	0	0	4	University of Lisbon	4Pipe4 pipeline, Seqclean, Mira, Bambus2
CRACS	CRACS	0	0	1		Institute for Systems and Computer Engineering of Porto TEC, European Bioinformatics Institute	ABySS, SSPACE, Bowtie, and FASTX
CSHL	CSHL	0	3	0		Cold Spring Harbor Laboratory, Yale University, University of Notre Dame	Metassembler, ALLPATHS, SOAPdenovo
CTD	CTD	0	3	0		National Research University of Information Technologies, Mechanics, and Optics	Unspecified
Curtain	CURT	0	0	1		European Bioinformatics Institute	SOAPdenovo, fastx_toolkit, bwa, samtools, velvet, and curtain
GAM	GAM	0	0	1		Institute of Applied Genomics, University of Udine, KTH Royal Institute of Technology	GAM, CLC and ABySS
IOBUGA	IOB	0	2	0		University of Georgia, Institute of Aging Research	ALLPATHS-LG and SOAPdenovo
MLK Group	MLK	1	0	0	I	UC Berkeley	ABySS
Meraculous	MERAC	1	1	1	I	DOE Joint Genome Institute, UC Berkeley	meraculous
Newbler-454	NEWB	1	0	0	4	454 Life Sciences	Newbler
Phusion	PHUS	1	0	1	I	Wellcome Trust Sanger Institute	Phusion2, SOAPdenovo, SSPACE
PRICE	PRICE	0	0	1		UC San Francisco	PRICE
Ray	RAY	1	1	1	I	CHUQ Research Center, Laval University	Ray
SGA	SGA	1	1	1	I	Wellcome Trust Sanger Institute	SGA
SOAPdenovo	SOAP	3	1	1	I[2]	BGI-Shenzhen, HKU-BGI	SOAPdenovo
Symbiose	SYMB	0	1	1		ENS Cachan/IRISA, INRIA, CNRS/ Symbiose	Monument, SSPACE, SuperScaffolder, and GapCloser

Assemblathon

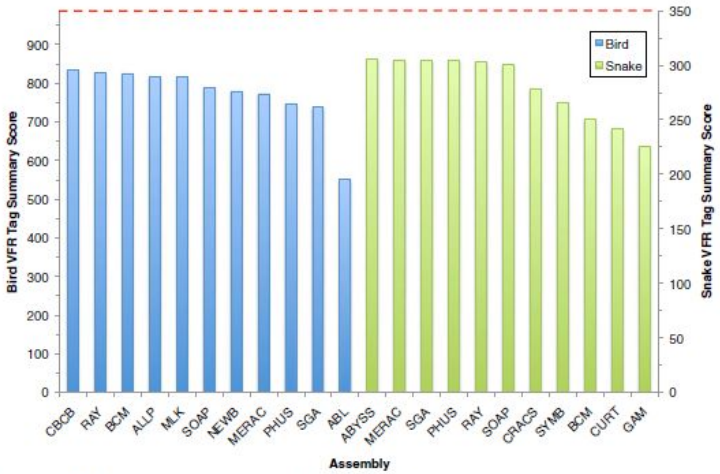


Figure 12 Short-range scaffold accuracy assessment via Validated Fosmid Regions. First, validated Fosmid regions (VFRs) were identified (86 in bird and 56 in snake, see text). Then VFRs were divided into non-overlapping 1,000 nt fragments and pairs of 100 nt 'tags' were extracted from ends of each fragment and searched (using BLAST) against all scaffolds from each assembly. A summary score for each assembly was calculated as the product of a) the number of pairs of tags that both matched the same scaffold in an assembly (at any distance apart) and b) the percentage of only the uniquely matching tag pairs that matched at the expected distance (± 2 nt). Theoretical maximum scores, which assume that all tag-pairs would map uniquely to a single scaffold, are indicated by red dashed line (988 for bird and 350 for snake).

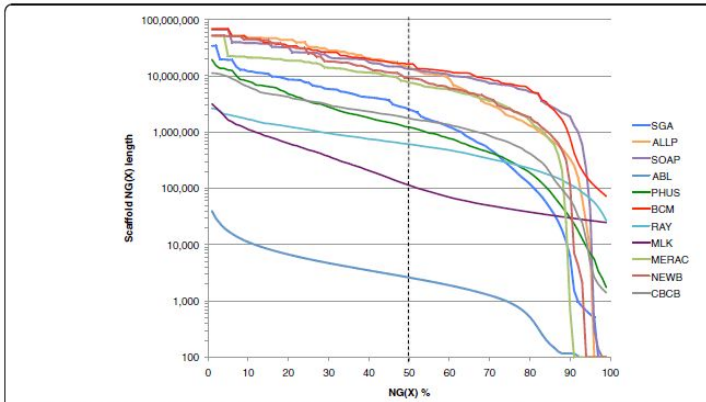


Figure 1 NG graph showing an overview of bird assembly scaffold lengths. The NG scaffold length (see text) is calculated at integer thresholds (1% to 100%) and the scaffold length (in bp) for that particular threshold is shown on the y-axis. The dotted vertical line indicates the NG50 scaffold length: if all scaffold lengths are summed from longest to the shortest, this is the length at which the sum length accounts for 50% of the estimated genome size. Y-axis is plotted on a log scale. Bird estimated genome size = ~1.2 Gbp.

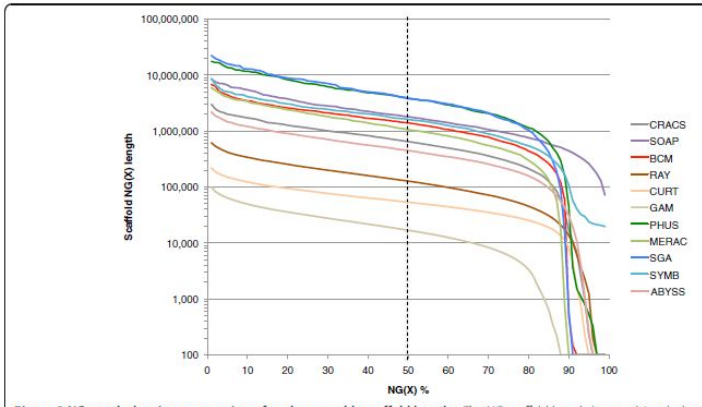
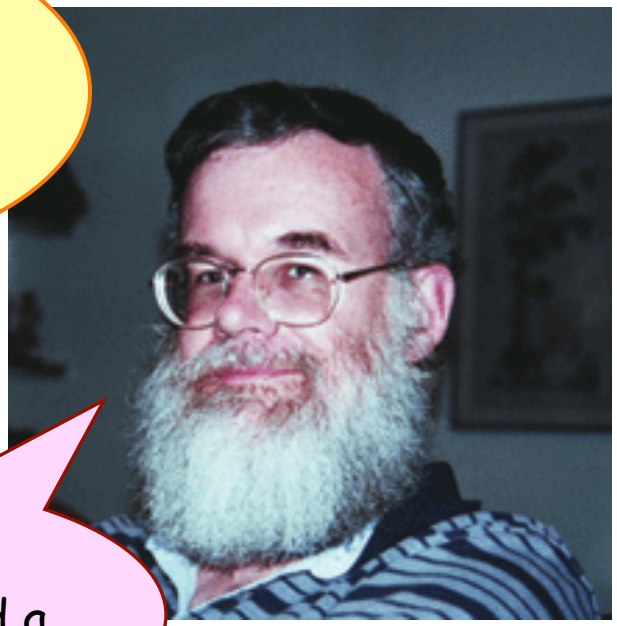


Figure 3 NG graph showing an overview of snake assembly scaffold lengths. The NG scaffold length (see text) is calculated at integer thresholds (1% to 100%) and the scaffold length (in bp) for that particular threshold is shown on the y-axis. The dotted vertical line indicates the NG50 scaffold length: if all scaffold lengths are summed from longest to the shortest, this is the length at which the sum length accounts for 50% of the estimated genome size. Y-axis is plotted on a log scale. Snake estimated genome size = ~1.0 Gbp.

History of WGA

1997

Let's sequence the human genome with the shotgun strategy



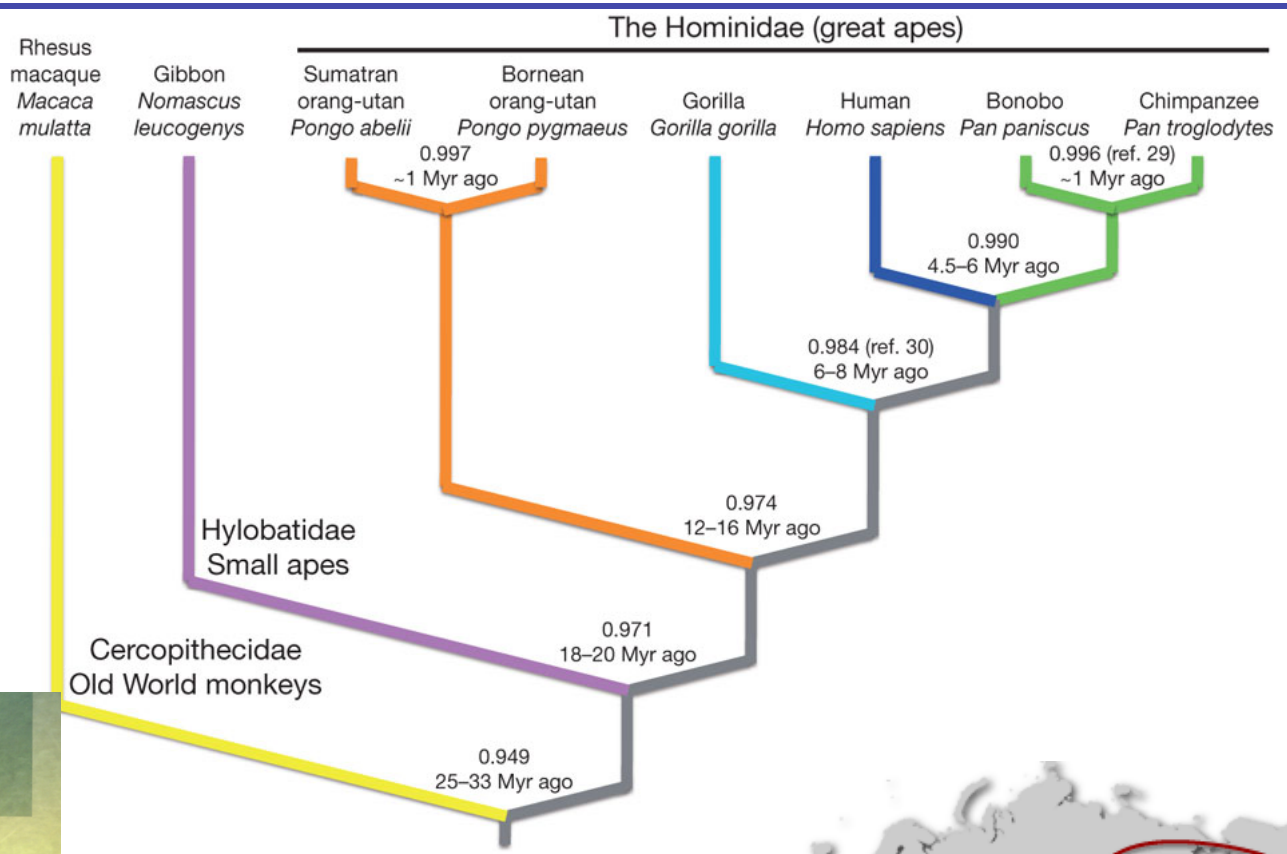
That is impossible, and a bad idea anyway

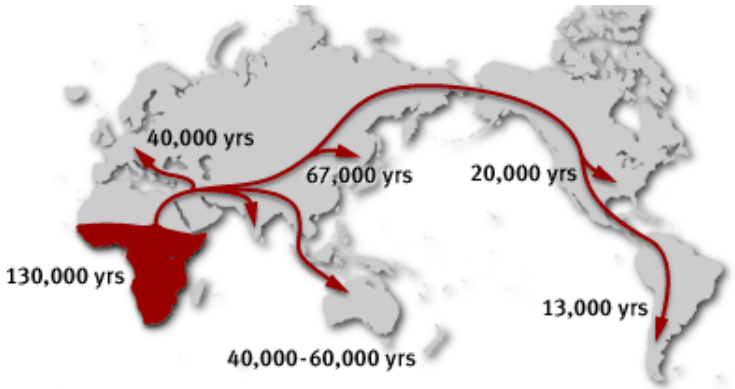
Gene Myers

Phil Green

Human Genome Diversity, Coalescence & Haplotypes

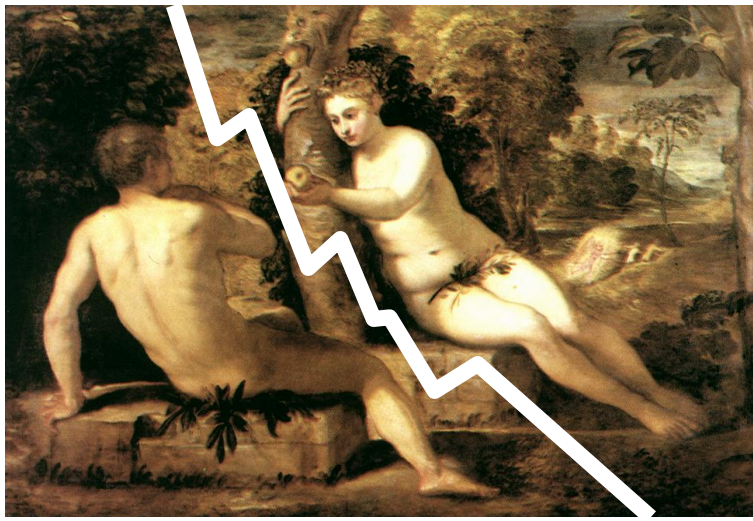
The Hominid Lineage



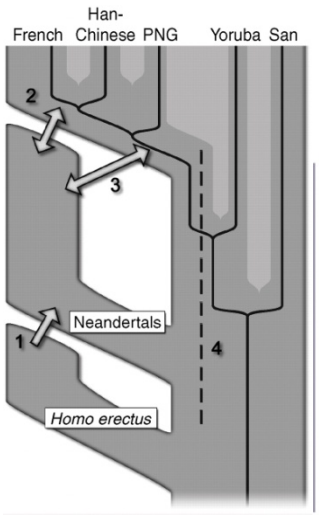


Human population migrations

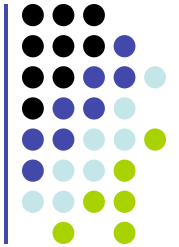
- Out of Africa, Replacement
  - Single mother of all humans (Eve) ~99,000 – 150,000yr
  - Single father of all humans (Adam) ~120,000 - 340,000yr
  - Humans out of Africa ~50000 years ago replaced others (e.g., Neandertals)

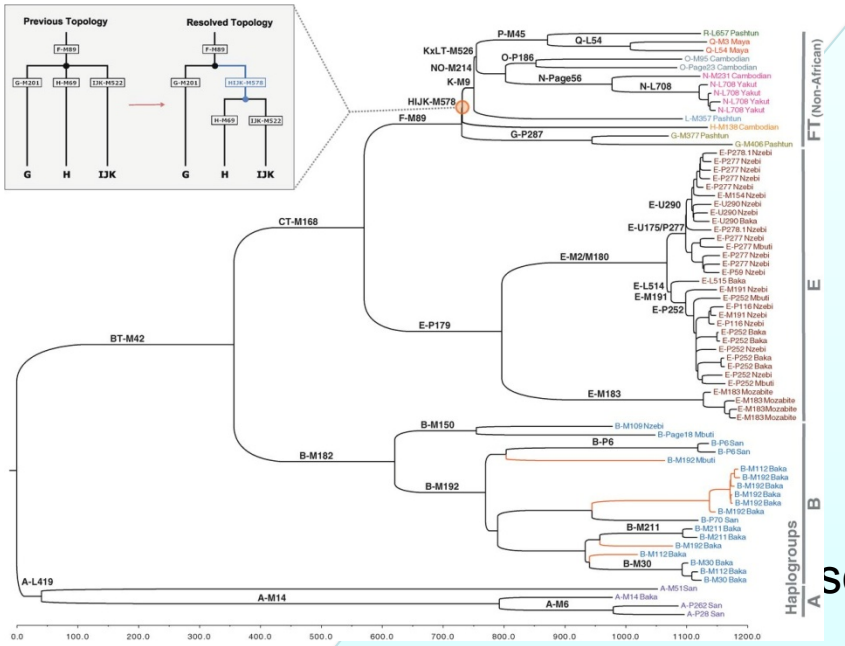


- Multiregional Evolution
  - Generally debunked, however,
  - ~5% of human genome in Europeans, Asians is Neanderthal, Denisova

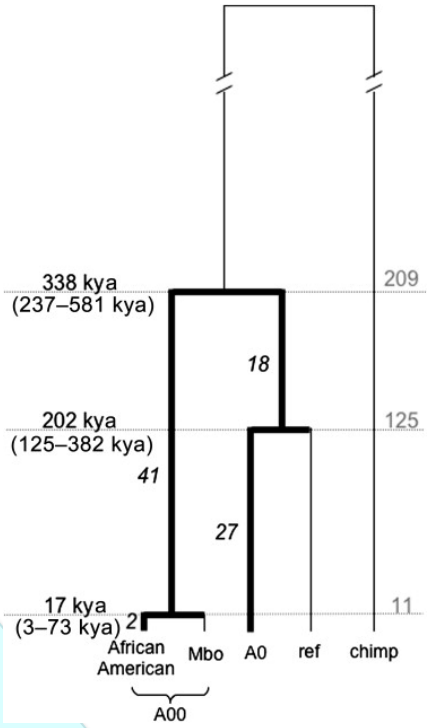


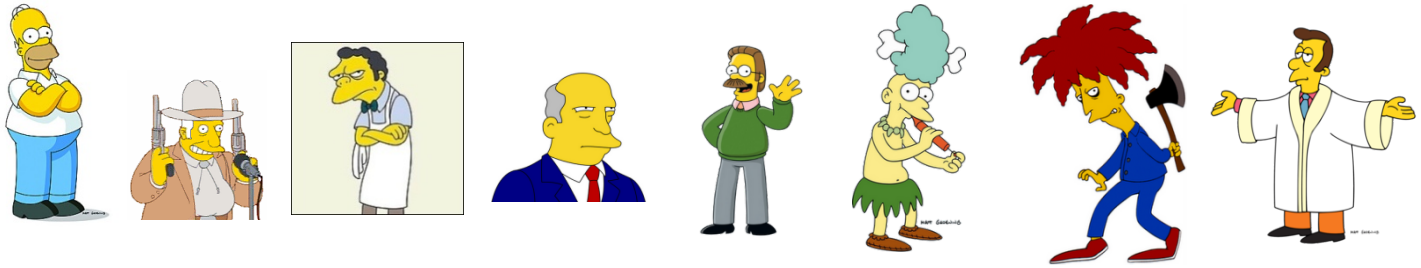
Coalescence





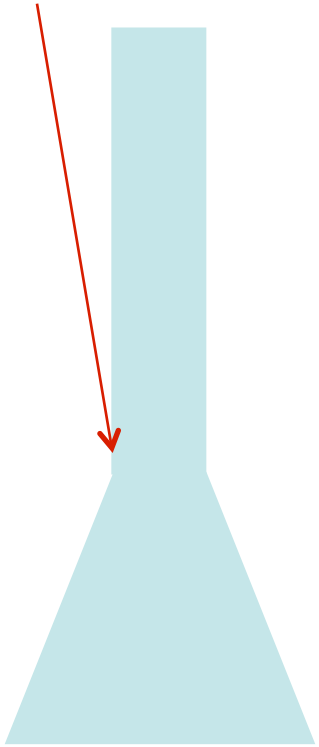
some coalescence

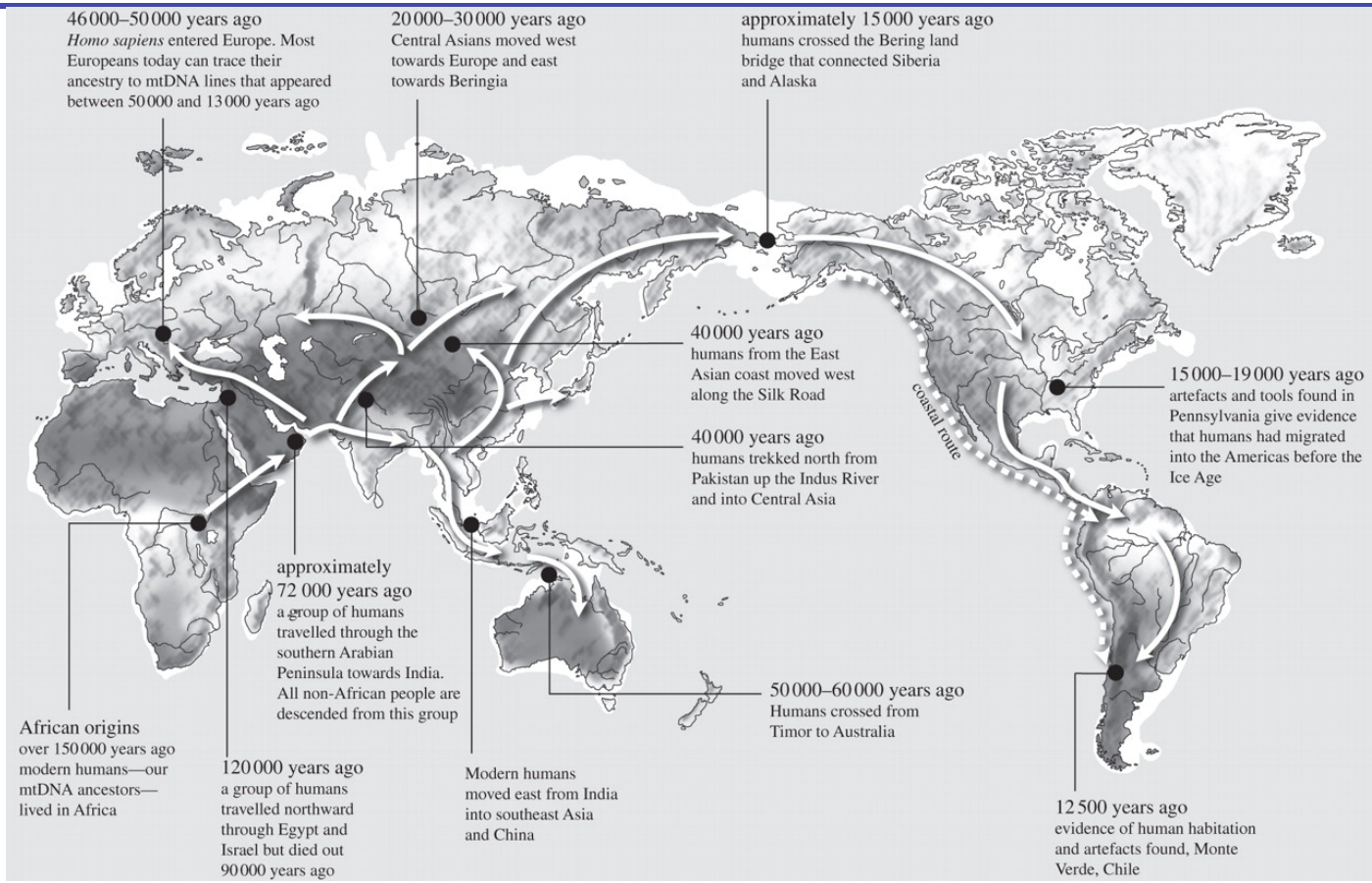




Why humans are so similar

Out of Africa





Oppenheimer S Phil. Trans. R. Soc. B 2012;367:770-784

The Neanderthal Genome

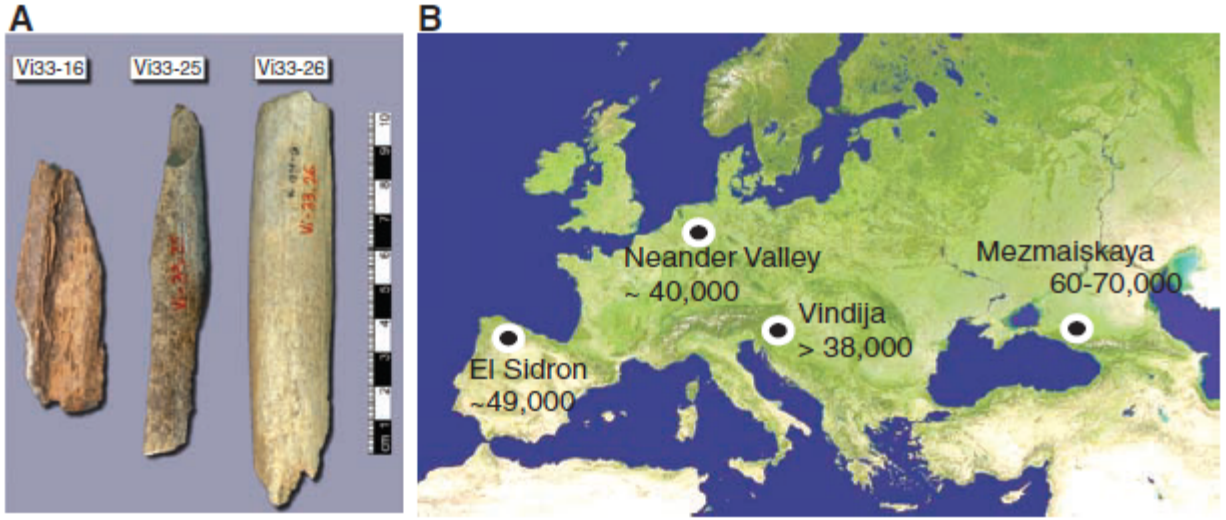


Fig. 1. Samples and sites from which DNA was retrieved. (**A**) The three bones from Vindija from which Neandertal DNA was sequenced. (**B**) Map showing the four archaeological sites from which bones were used and their approximate dates (years B.P.).

- From bones, compared genomes of three different Neanderthals with five genomes from modern humans from different areas of the world

Figure 1- R. E. Green et al., Science 328, 710-722 (2010)

Neanderthal Genome

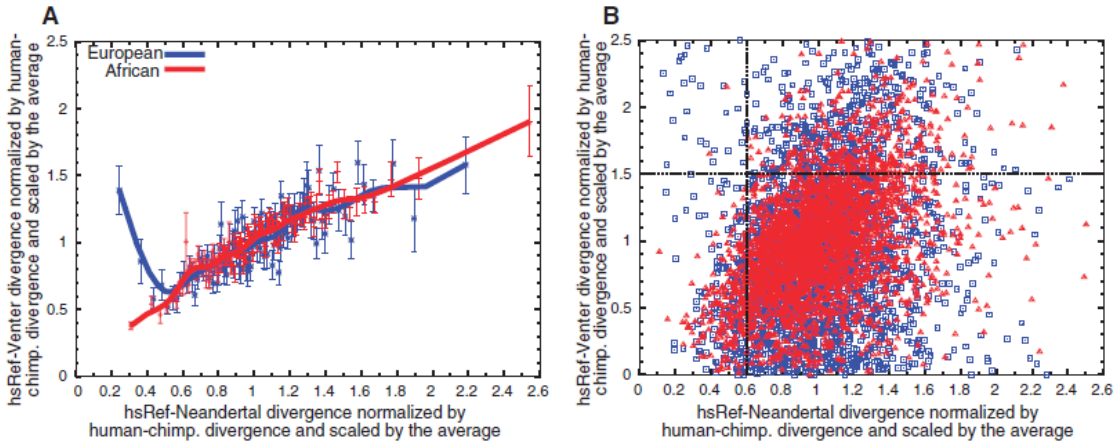


Fig. 5. Segments of Neandertal ancestry in the human reference genome. We examined 2825 segments in the human reference genome that are of African ancestry and 2797 that are of European ancestry. (**A**) European segments, with few differences from the Neandertals, tend to have many differences from other present-day humans, whereas African segments do not, as expected if the former are derived from Neandertals. (**B**) Scatter plot of the segments in (A) with respect to their divergence to the Neandertals and to Venter. In the top left quandrant, 94% of segments are of European ancestry, suggesting that many of them are due to gene flow from Neandertals.

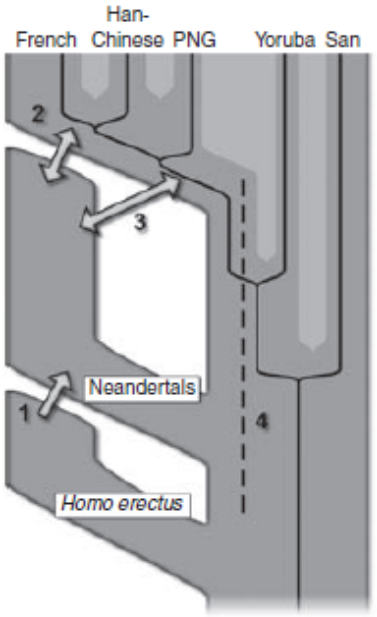


Fig. 6. Four possible scenarios of genetic mixture involving Neandertals. Scenario 1 represents gene flow into Neandertal from other archaic hominins, here collectively referred to as *Homo erectus*. This would manifest itself as segments of the Neandertal genome with unexpectedly high divergence from present-day humans. Scenario 2 represents gene flow between late Neandertals and early modern humans in Europe and/or western Asia. We see no evidence of this because Neandertals are equally distantly related to all non-Africans. However, such gene flow may have taken place without leaving traces in the present-day gene pool. Scenario 3 represents gene flow between Neandertals and the ancestors of all non-Africans. This is the most parsimonious explanation of our observation. Although we detect gene flow only from Neandertals into modern humans, gene flow in the reverse direction may also have occurred. Scenario 4 represents old substructure in Africa that persisted from the origin of Neandertals until the ancestors of non-Africans left Africa. This scenario is also compatible with the current data.

Denisovan – Another human relative

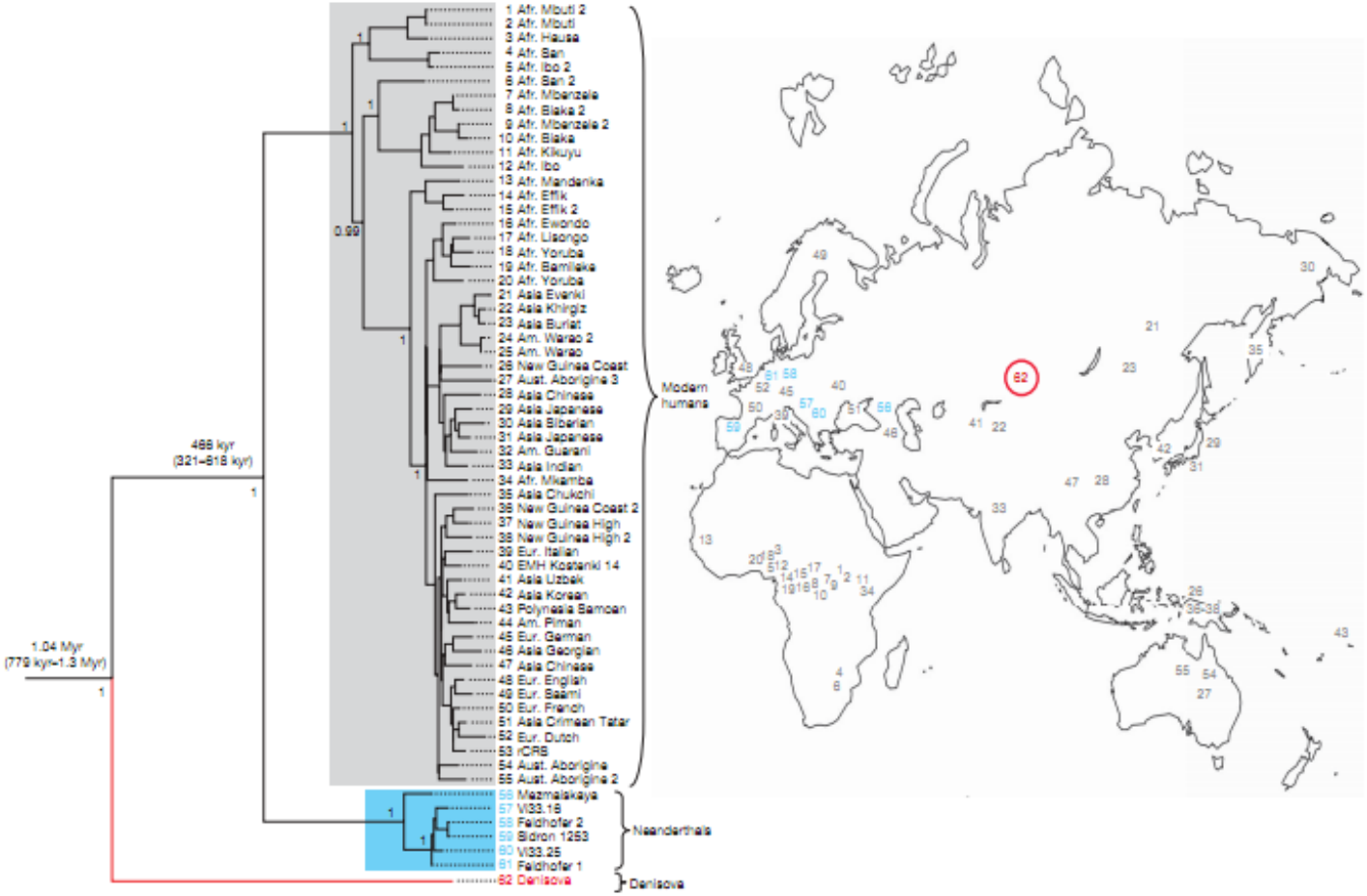
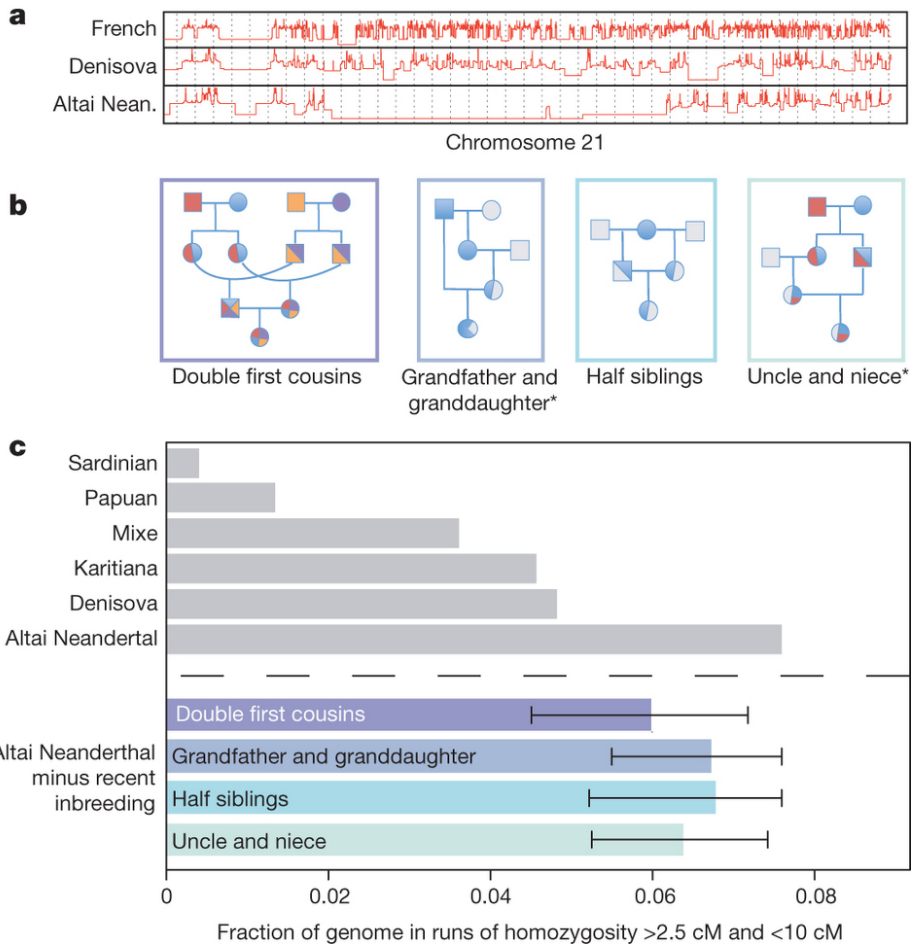
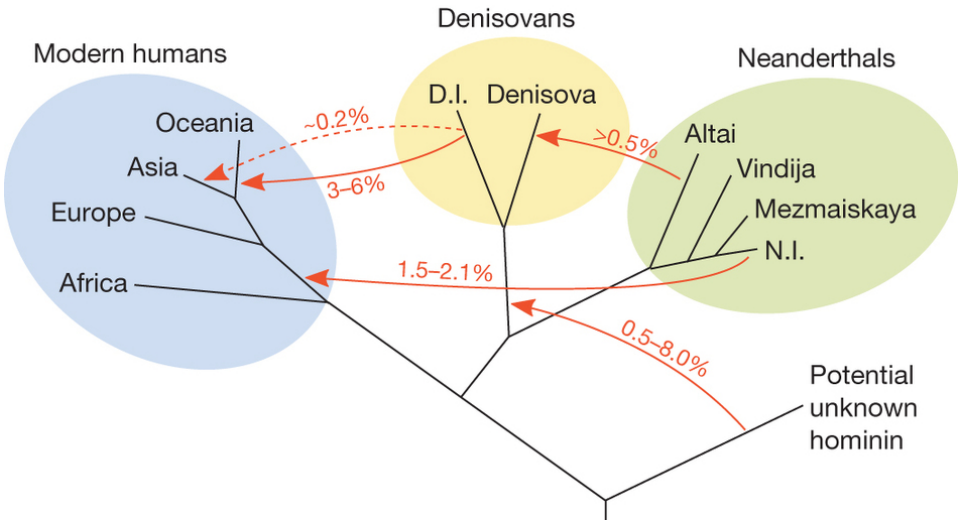


Figure 3 | Phylogenetic tree of complete mtDNAs. The phylogeny was estimated with a Bayesian approach under a GTR+1+Γ model using 54 present-day and one Pleistocene modern human mtDNA (grey), 6 Neanderthals (blue) and the Denisova hominin (red). The tree is rooted with a chimpanzee and a bonobo mtDNA. Posterior probabilities are given for each major node. The map shows the geographical origin of the mtDNAs (24, 25, 32, 44 are in the Americas). Note that two partial mtDNAs sequenced from Teshik Tash and Okladikov Cave in Central Asia fall together with the complete Neanderthal mtDNAs in phylogenies[4] (not shown).

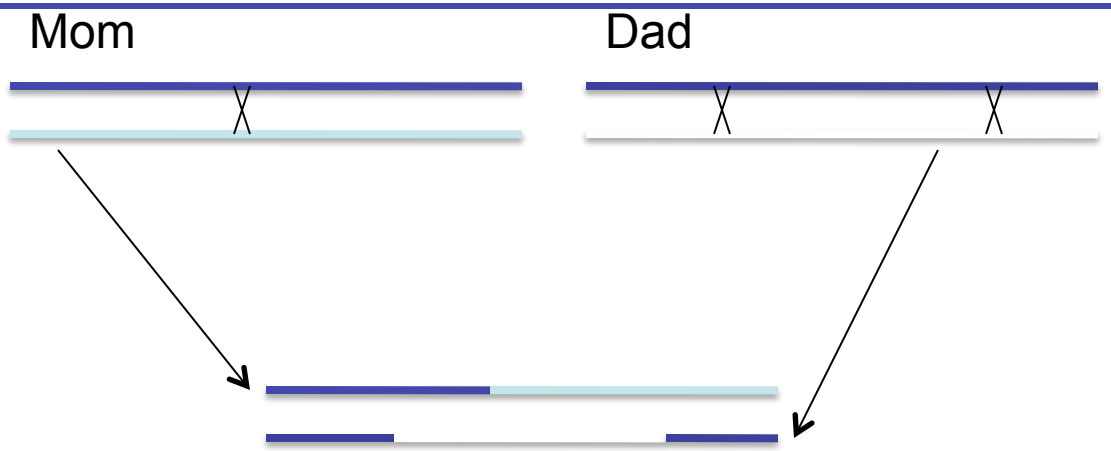
The Neanderthal Whole Genome





Some Key Definitions

```
Mary: AGCC G/G CG
John: AGCC G/G CG
Josh: AGCC G/T CG
Kate: AGCC G/G CG
Pete: AGCC G/G CG
Anne: AGCC G/G CG
Mimi: AGCC G/G CG
Mike: AGCC T/T CG
Olga: AGCC T/G CG
Tony: AGCC T/G CG
```



Alleles: G, T

Major Allele: G
Minor Allele: T

Heterozygosity:
Prob[2 alleles picked at random with replacement are different]

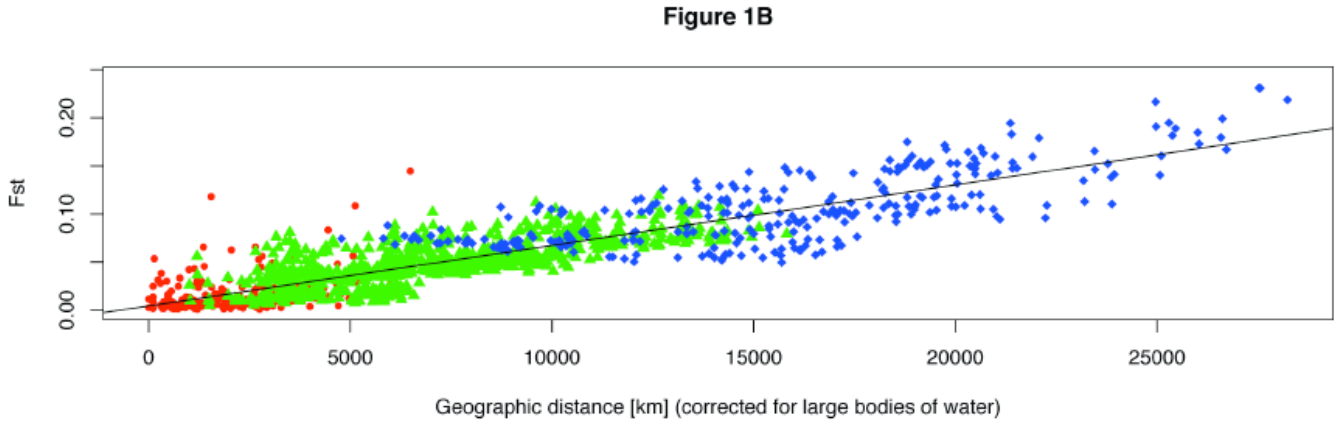
2*.75*.25 = .375

H = 4Nu/(1+4Nu)

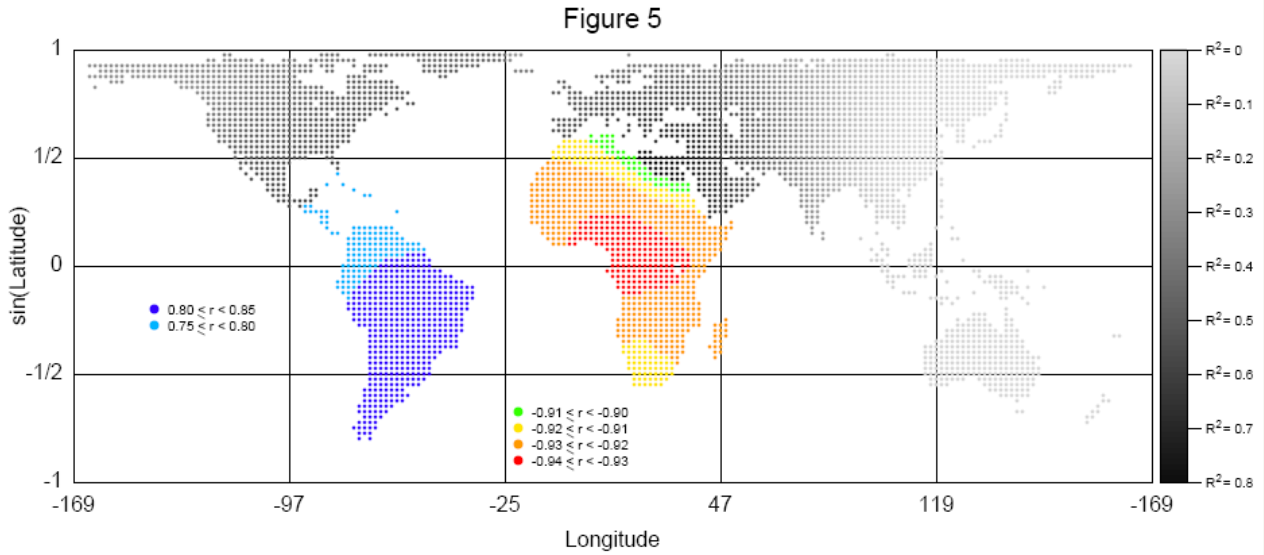
Recombinations:
At least 1/chromosome
On average ~1/100 Mb

Linkage Disequilibrium:
The degree of correlation between two SNP locations

The Fall in Heterozygosity

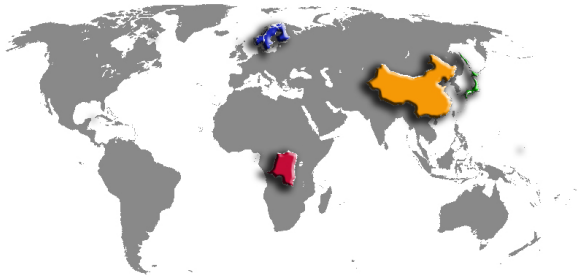


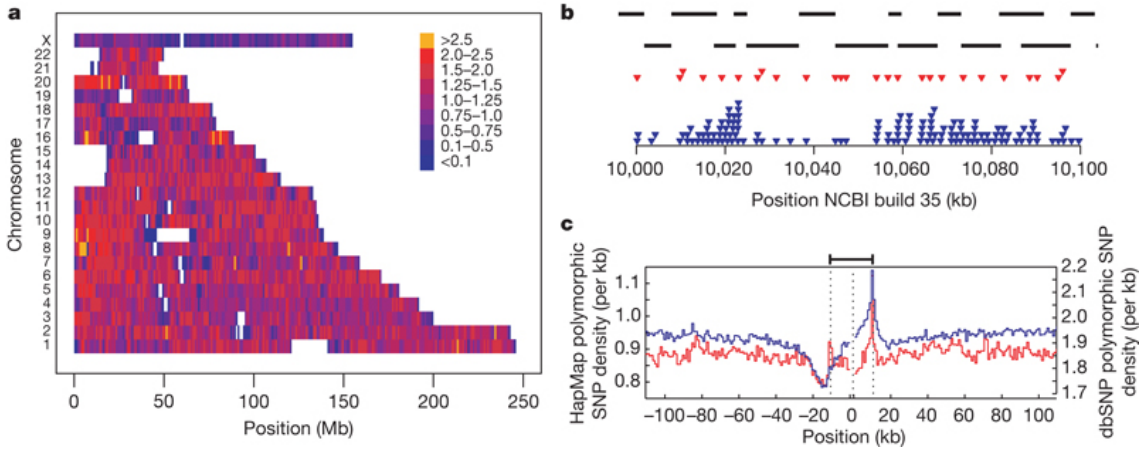
$$F_{ST} = \frac{H - H_{POP}}{H}$$



The HapMap Project

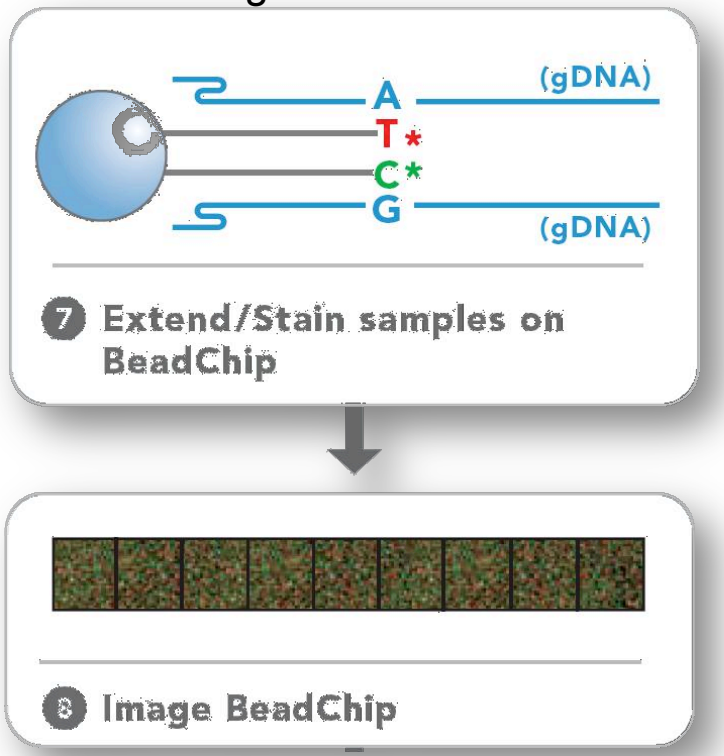
ASW	African ancestry in Southwest USA	90
CEU	Northern and Western Europeans (Utah)	180
CHB	Han Chinese in Beijing, China	90
CHD	Chinese in Metropolitan Denver	100
GIH	Gujarati Indians in Houston, Texas	100
JPT	Japanese in Tokyo, Japan	91
LWK	Luhya in Webuye, Kenya	100
MXL	Mexican ancestry in Los Angeles	90
MKK	Maasai in Kinyawa, Kenya	180
TSI	Toscani in Italia	100
YRI	Yoruba in Ibadan, Nigeria	100





Genotyping:

Probe a limited number (~1M) of known highly variable positions of the human genome



Linkage Disequilibrium & Haplotype Blocks

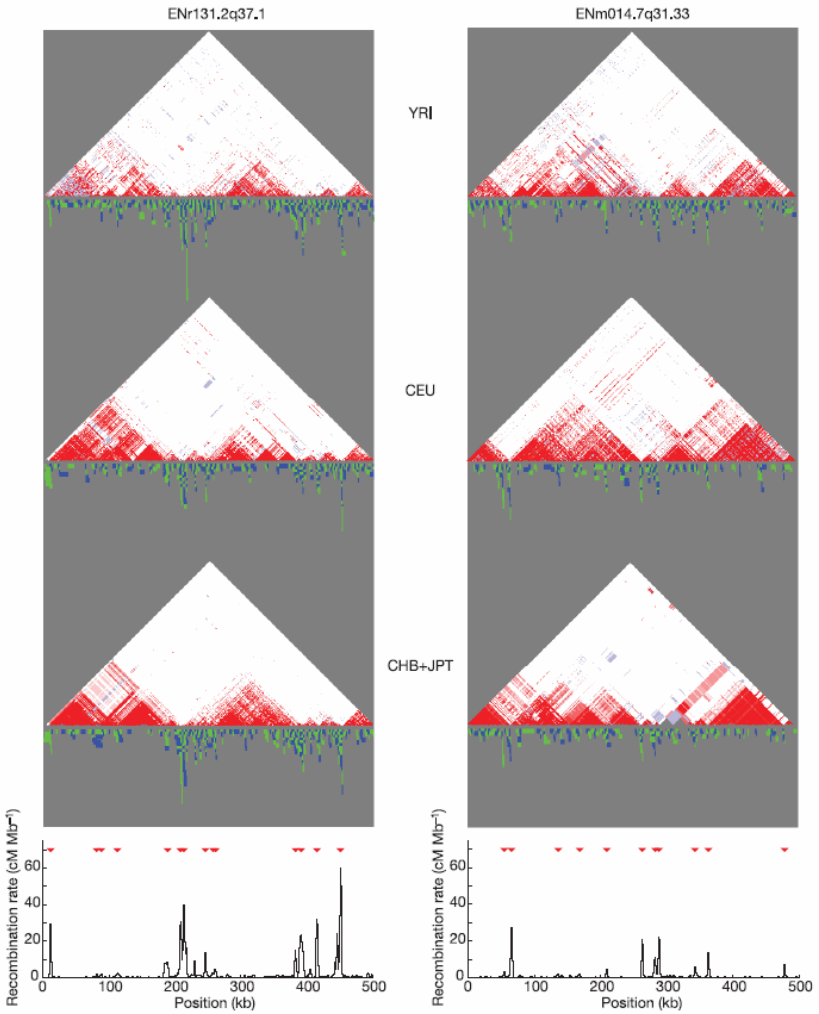


Figure 8 | Comparison of linkage disequilibrium and recombination for two ENCODE regions. For each region (ENr131.2q37.1 and ENm014.7q31.33), D' plots for the YRI, CEU and CHB+JPT analysis panels are shown: white, $D' < 1$ and LOD < 2; blue, $D' = 1$ and LOD < 2; pink, $D' < 1$ and LOD ≥ 2; red, $D' = 1$ and LOD ≥ 2. Below each of these plots is shown the intervals where distinct obligate recombination events must have occurred (blue and green indicate adjacent intervals). Stacked intervals represent regions where there are multiple recombination events in the sample history. The bottom plot shows estimated recombination rates, with hotspots shown as red triangles[46].

Minor allele: A G

p_A p_G

Linkage Disequilibrium (LD):

$D = P(A \text{ and } G) - p_A p_G$

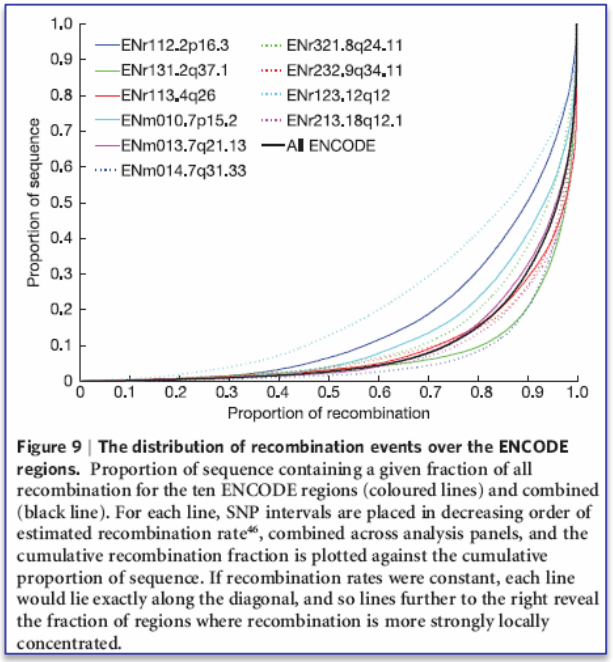
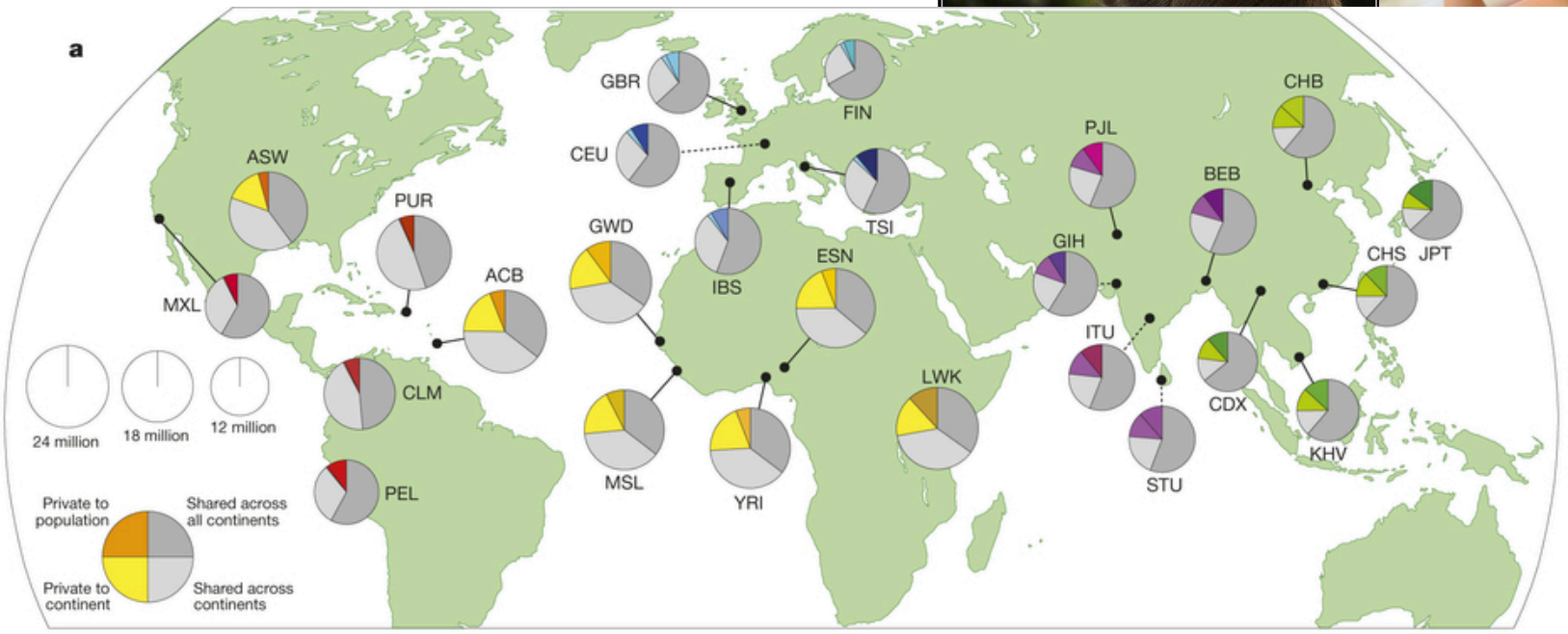
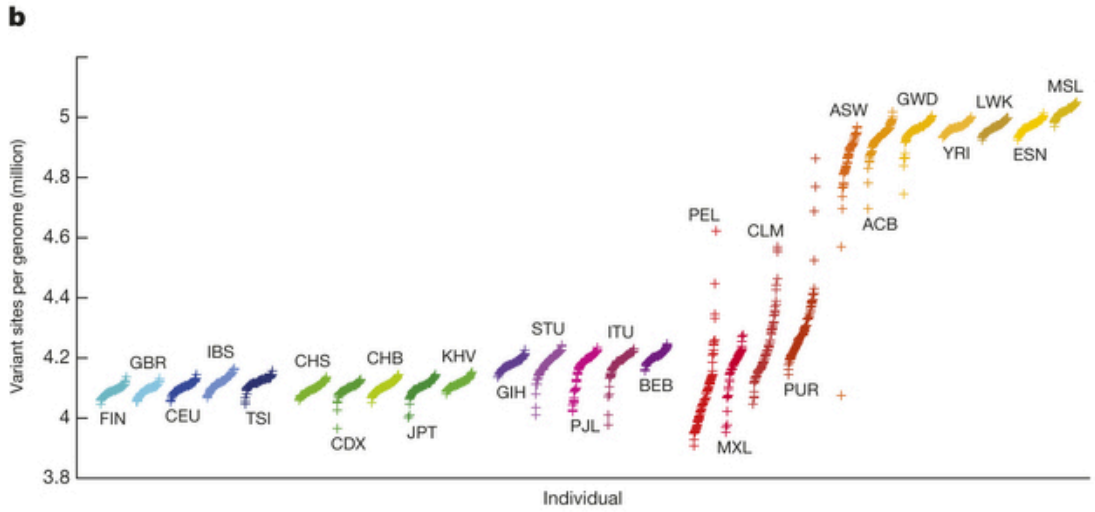
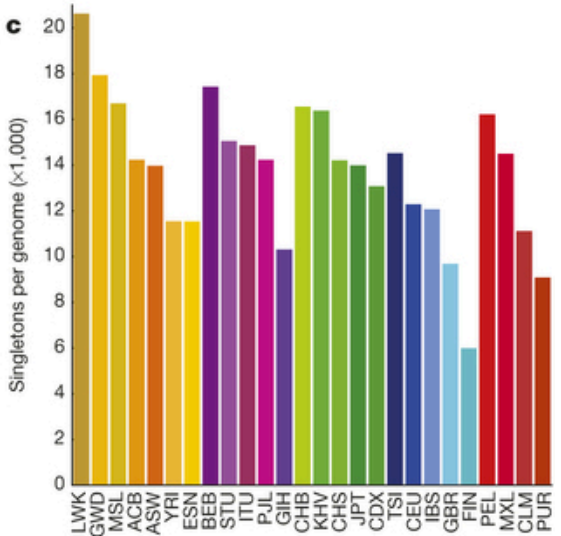


Figure 9 | The distribution of recombination events over the ENCODE regions. Proportion of sequence containing a given fraction of all recombination for the ten ENCODE regions (coloured lines) and combined (black line). For each line, SNP intervals are placed in decreasing order of estimated recombination rate[46], combined across analysis panels, and the cumulative recombination fraction is plotted against the cumulative proportion of sequence. If recombination rates were constant, each line would lie exactly along the diagonal, and so lines further to the right reveal the fraction of regions where recombination is more strongly locally concentrated.

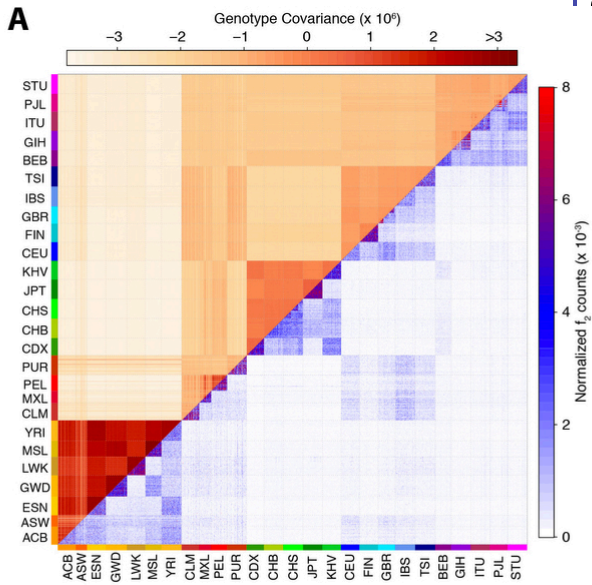
Population Sequencing – 1000 Genomes Project

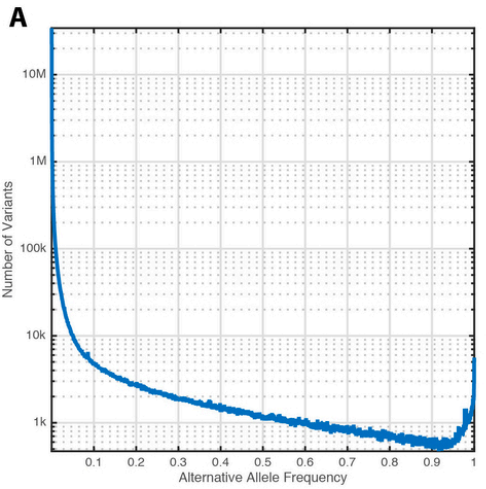


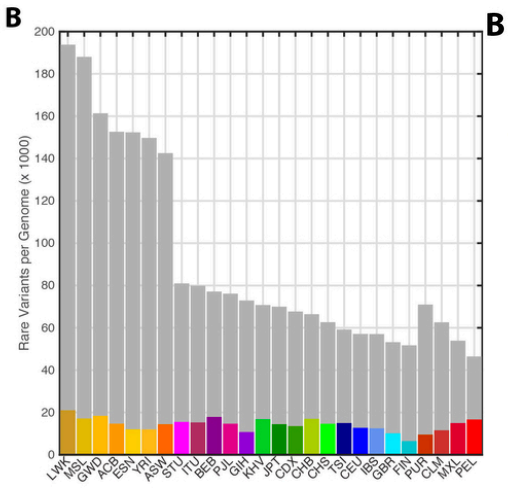


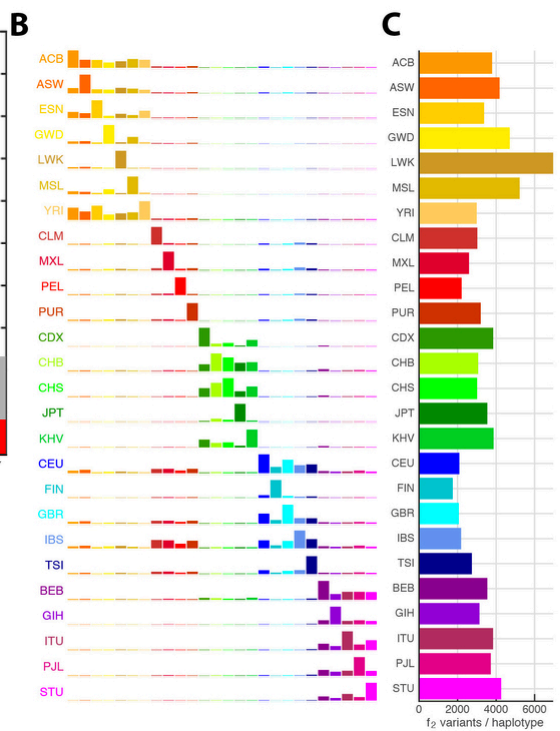


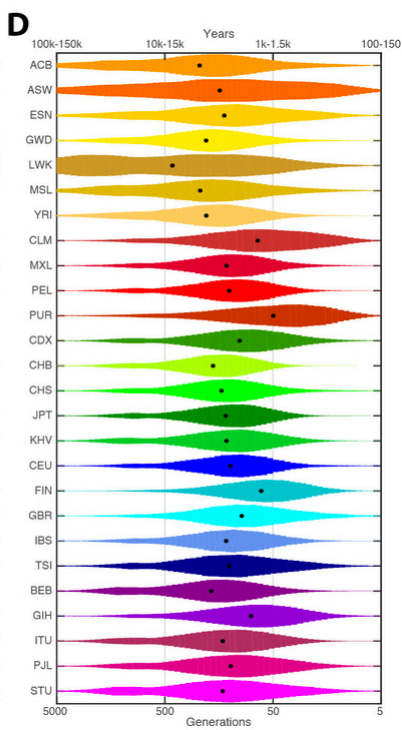
Population Sequencing – 1000 Genomes Project



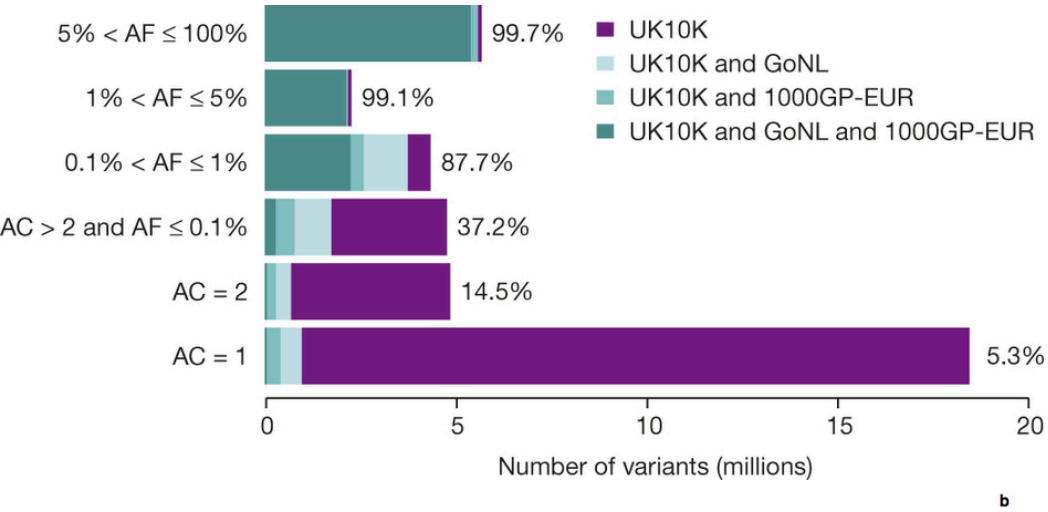




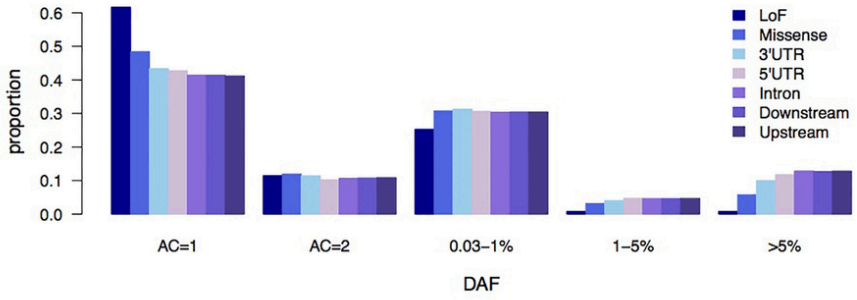


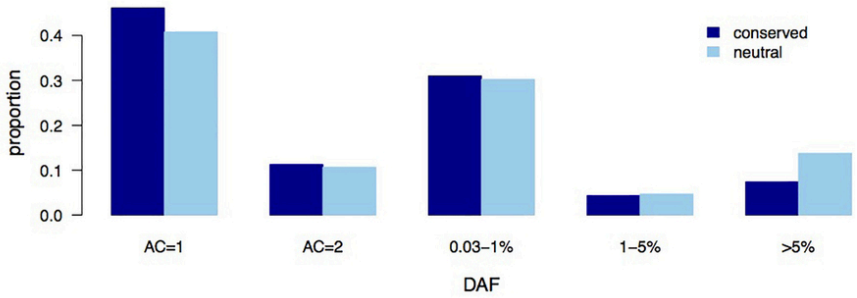


Population Sequencing – UK10K

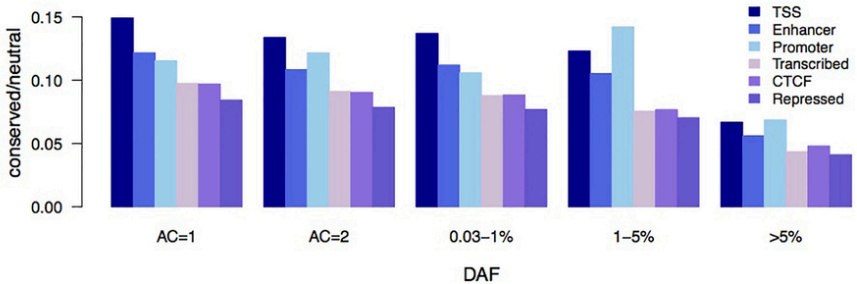


b

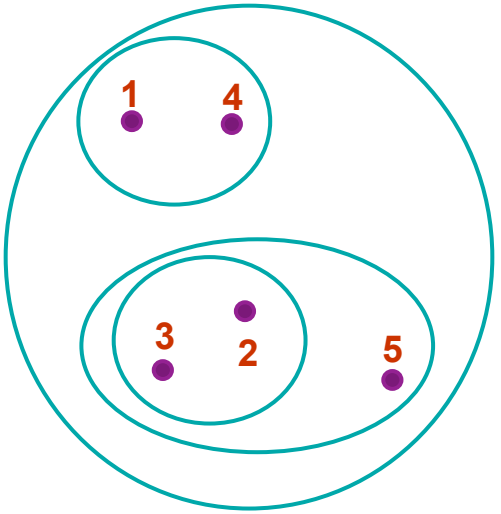


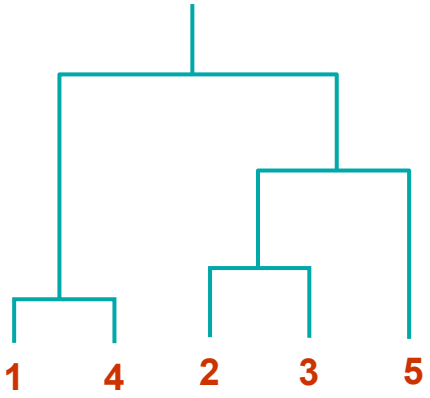


c

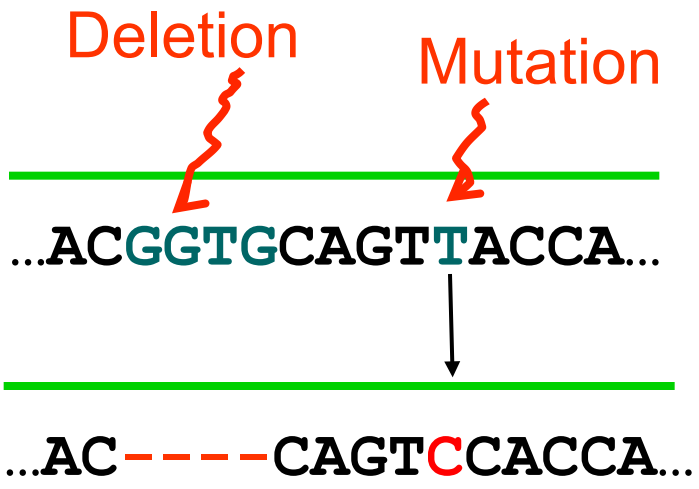


Molecular Evolution and Phylogenetic Tree Reconstruction



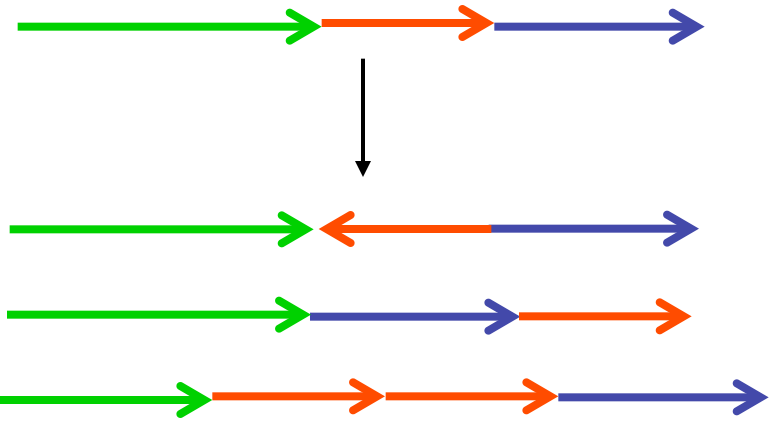


Evolution at the DNA level

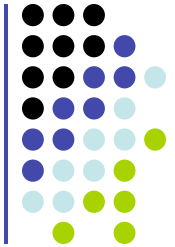


SEQUENCE EDITS

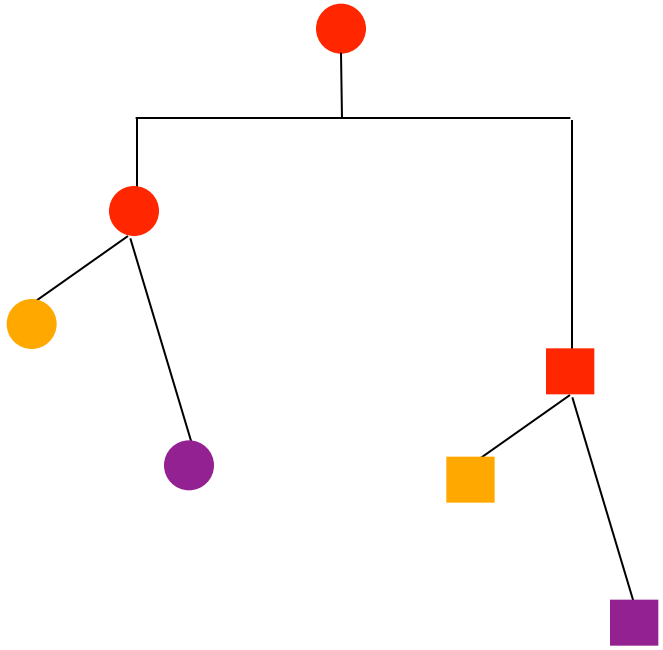
REARRANGEMENTS



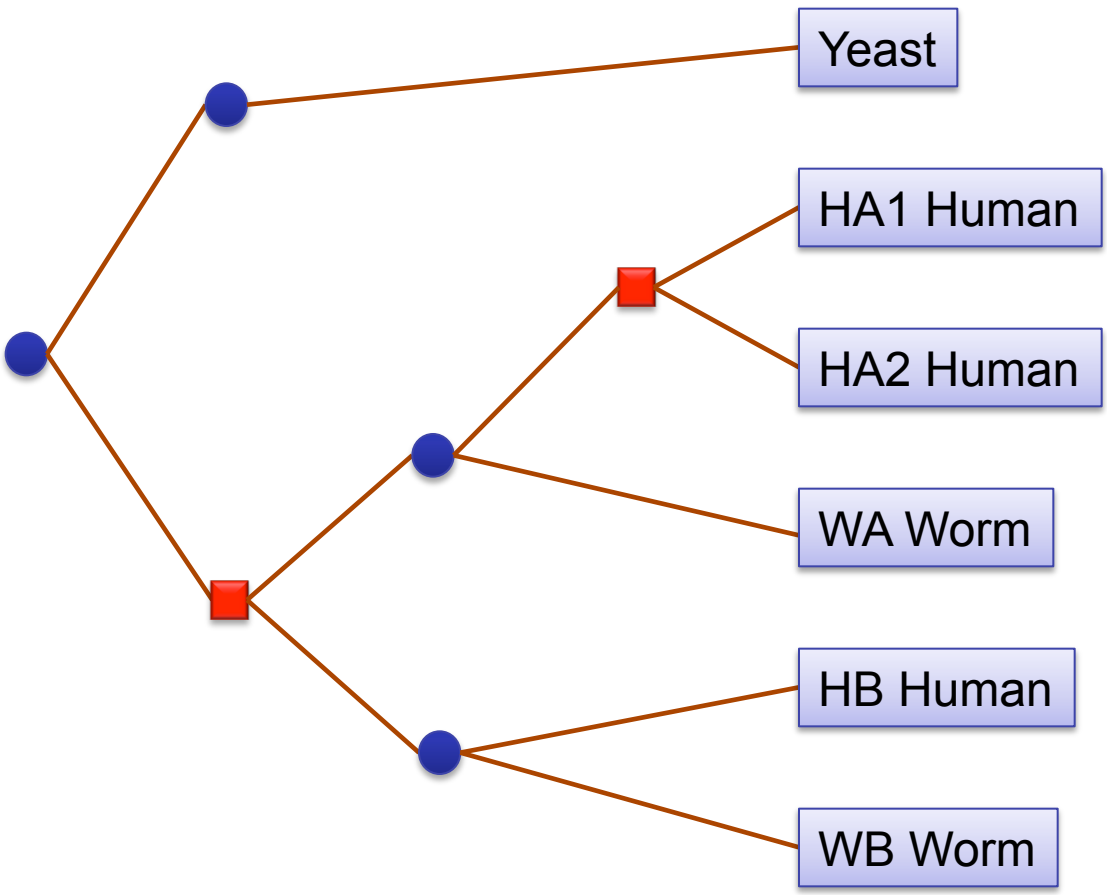
Protein Phylogenies



- Proteins (genes) evolve by both duplication and species divergence



Orthology and Paralogy



Orthologs: *Derived by speciation*

Paralogs: *Everything else*

Orthology, Paralogy, Inparalogs, Outparalogs

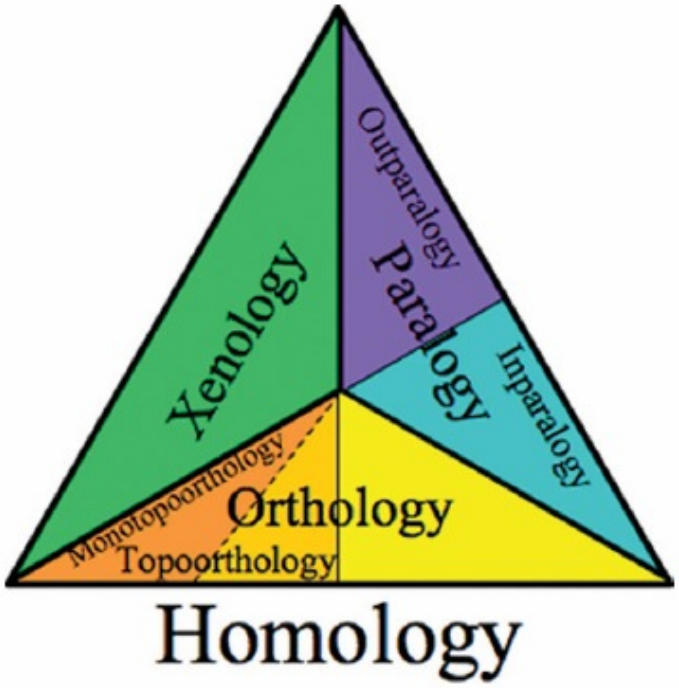


Figure 1. Refinements of homology.

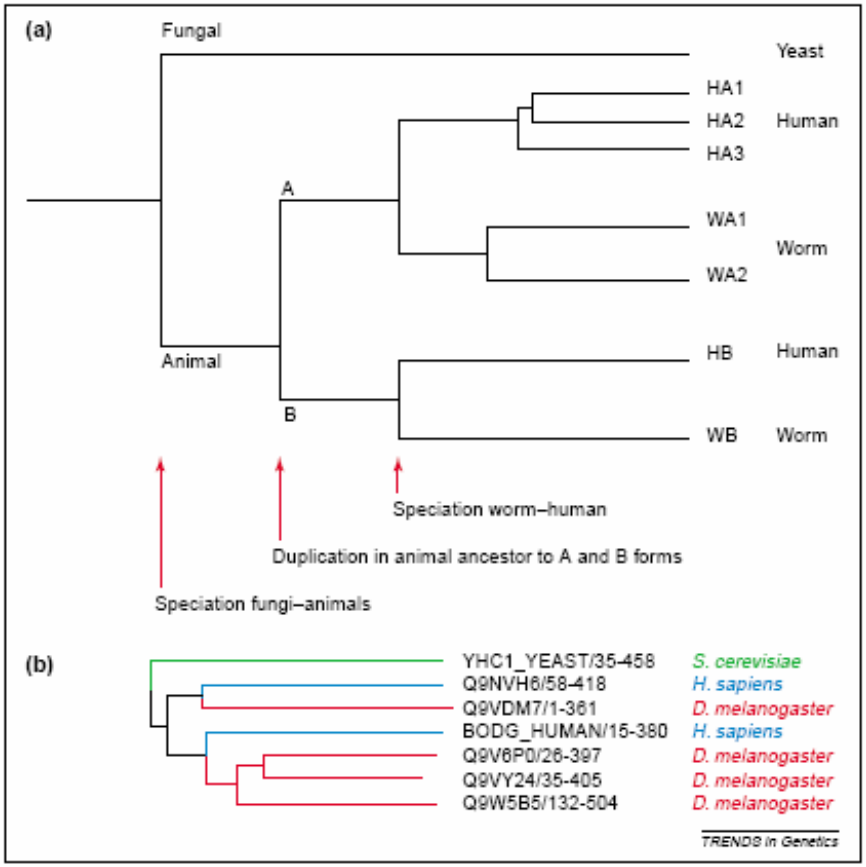


Fig. 1. The definition of inparalogs and outparalogs. (a) Consider an ancient gene inherited in the yeast, worm and human lineages. The gene was duplicated early in the animal lineage, before the human–worm split, into genes A and B. After the human–worm split, the A form was in turn duplicated independently in the human and worm lineages. In this scenario, the yeast gene is orthologous to all worm and human genes, which are all co-orthologous to the yeast gene. When comparing the human and worm genes, all genes in the HA* set are co-orthologous to all genes in the WA* set. The genes HA* are hence 'inparalogs' to each other when comparing human to worm. By contrast, the genes HB and HA* are 'outparalogs' when comparing human with worm..However, HB and HA*, and WB and WA* are inparalogs when comparing with yeast, because the animal–yeast split pre-dates the HA*–HB duplication. (b) Real-life example of inparalogs: γ-butyrobetaine hydroxylases. The points of speciation and duplication are easily identifiable. The alignment is a subset of Pfam:PF03322 and the tree was generated by neighbor-joining in Belvu. All nodes have a bootstrap support exceeding 95%.

Phylogenetic Trees

- Nodes: species
- Edges: time of independent evolution
- Edge length represents evolution time
  - AKA genetic distance
  - Not necessarily chronological time