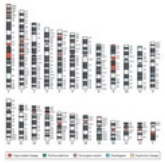
Genetics, Genomics and Human Origins



Introduction to Anthropogeny Lecture 2
Pascal Gagneux

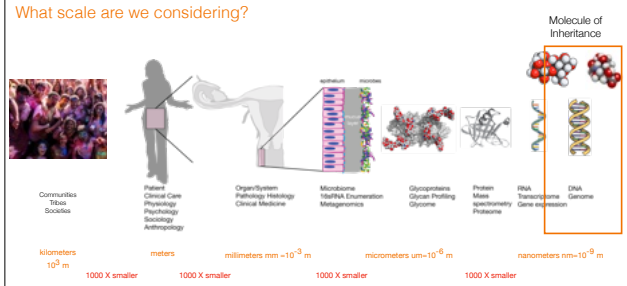
Wednesday, Oct 1, 2026

Question:

- What is the genetic difference between humans and apes in %?

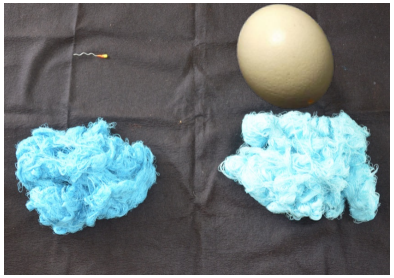
Scales: Societies to Atoms….

What scale are we considering?



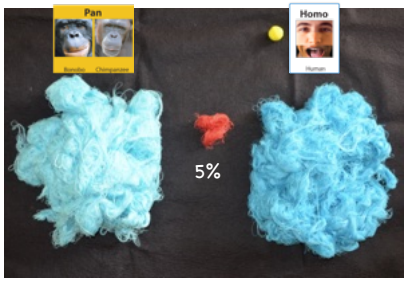
Our molecular nature ranges from societies to atoms. Genetic inheritance is mediated by molecules in the nanometer (10^{-9} m) range: DNA.
Sugars such as the disaccharide lactose and the monosaccharide sialic acid are about 1 to 2 nm in size, **1 billion time smaller than our bodies!**
You have to shrink 1000 times too be the size of a small blood vessel (millimiter, mm)
Another 100 times to be the size of a cell (micrometer, um)
And a third round of 1000 times to be in the realm of a sugar or DNA molecule (nanometer, nm)

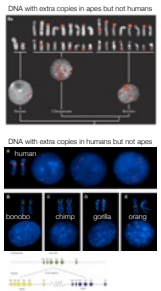
1000 x model of the genome in Sperm and Egg



Human egg 130 micron, magnified 1000 times = 130 mm = 13 cm
Somatic cell 20 micron, magnified 1000 times 2cm diameter
Sperm 60 micron long, magnified 1000 times = 6 cm long
Haploid human genome: 3 billion basepairs measuring ~1m magnified 1000 times = 1km

1000 x model of haploid genome

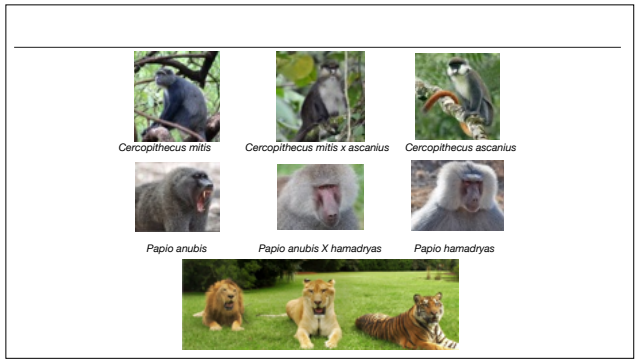




Including Insertions and deletions (pieces of the genomes that have no direct counterpart in the other genome: Human and chimpanzee genomes differ by ~ 5% of their DNA. Staining for chromosomes can reveal exceeds copies in apes (above) or new copies in humans (below) Example of a region where humans have two copies of a gene (MIC) but chimpanzees lost one of the two…

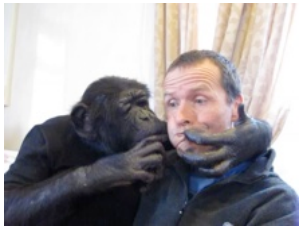
Question:

- How much genetic change is necessary for speciation to occur?



Many species in the wild hybridize. There is not linear metric of % DNA difference that can be applied to the possibility of viable hybridization.

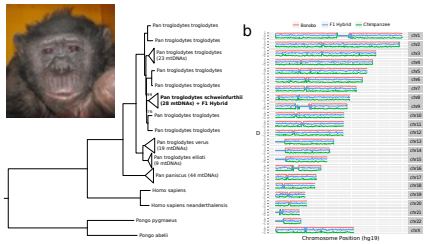
Hybridization happens



Female F1 hybrid of accidental cross between male bonobo and female common chimpanzee.

Genomic evidence for lack of pre-zygotic species barrier between chimpanzees and bonobos

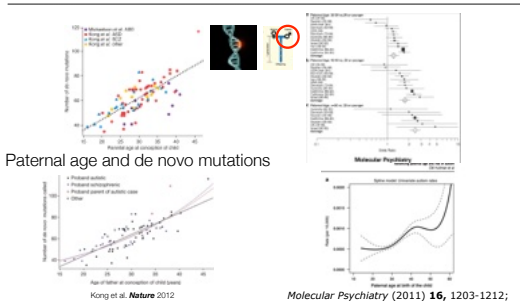
Samuel H. Vohr, Brendan O'Connell, Beatriz Freitas, Florence Ollivet-Courtois, Peter D. Heintzman, Pascal Gagneux, Alysson R. Muotri, and Richard E. Green submitted



Graph b shows that the DNA of the F1 Bonobo X chimpanzee hybrid has DNA that is equidistant from both chimp and bonobo parental chromosomes at all locations along each chromosome.

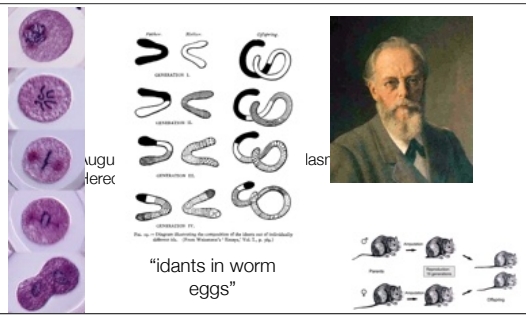
Question:

- How could culture/behavior affect genomic evolution?



Sequencing parent child trios for their whole genomes now allows to measure the rate of de novo mutation per generation.
About 50 new mutations are inherited by each child, unless the father is older, then many more de novo mutations are inherited.
Parental age is a strong risk factor for autism spectrum disorders.

Soma and Germ Line



The stuff of heredity, is particulate and is only passed on through specialized cell: the germ line. The rest of the body (the soma) serves as the vehicle to get these sexual cells to the right place: meeting with sex cells from other individuals.

Inheritance of acquired characteristics?

>200 generations of male circumcision, male babies still born with a prepuce!

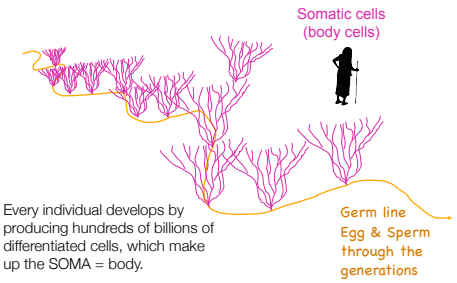
4500 years

> 220 generations

Relief depiction of circumcision from from the necropolis at Saqqara
Dating to Dynasty 6 and specifically to the Service of Nesi (Pharaoh) Teti (2355-2343 BCE)

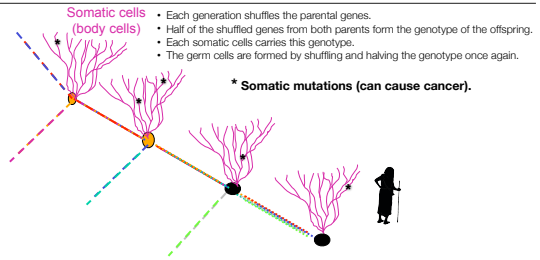
Just like the mouse tails cut off along successive generations by Weissmann, over 200 generation of male circumcision has not caused baby boys born without a prepuce. **It represents a culturally inherited practice!**

Germ Line and Soma



Every individual develops by producing hundreds of billions of differentiated cells, which make up the SOMA = body.

The Germ Line is not a simple line:



- Each generation shuffles the parental genes.
- Half of the shuffled genes from both parents form the genotype of the offspring.
- Each somatic cells carries this genotype.
- The germ cells are formed by shuffling and halving the genotype once again.

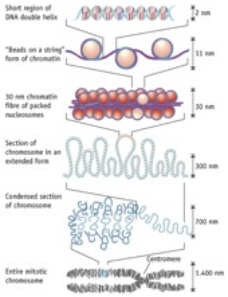
*** Somatic mutations (can cause cancer).**

The germ line is made up of shuffled pieces of DNA that meet and get taken apart again by sexual recombination.—-> **WE ARE DYNAMIC MOSAICS of mixed heritage!**

Each generation shuffles the parental genes. Half of the shuffled genes from both parents form the genotype of the offspring. Each somatic cells carries this genotype.
The germ cells are formed by shuffling and halving the genotype once again. Half of the chromosomes in the gametes are recombinant, the other half are as found in the parents.

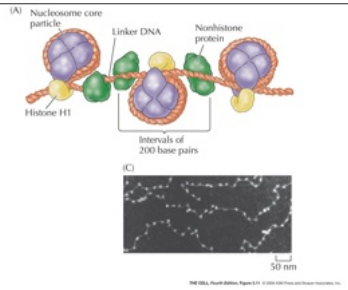
Good example of misperception: Blood as vehicle of inheritance? **Completely erroneous notions of “pure blood” and “contamination” are still around.**

Fitting a genome into a cell: genome packaging



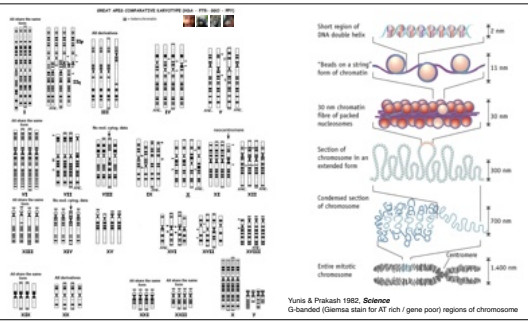
Highly efficient packaging of DNA into chromatin and chromosomes. Each chromosome is based on one segment of genomic DNA.

Genome packaging: chromatin

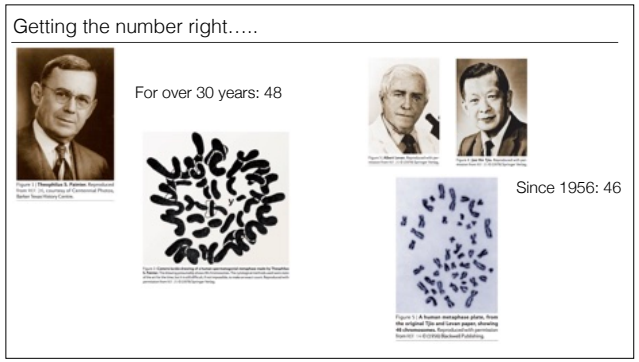


"Beads on a string" structure of chromatin, visible when chromosomes open up their tight coils in between cell divisions.

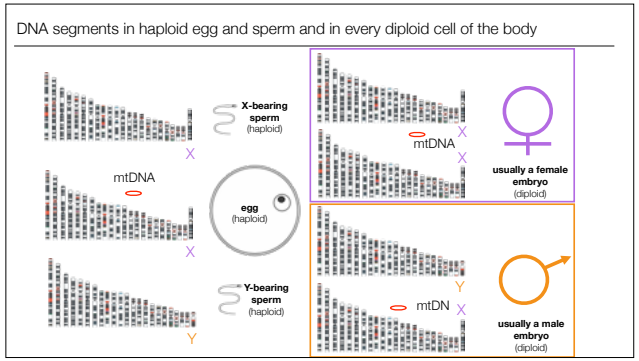
Fitting a genome into a cell: genome packaging



Comparison of human and great ape chromosomes. Humans have 46 chromosomes in each cell (sperm and eggs have 23 each, as they are haploid). The banning patterns obtained after staining with DNA specific dyes are highly similar between apes and humans.



It took a while to get the correct number of chromosome in human cells….
Diving human cells can easily be found in testes or white blood cells can be triggered to undergo cell division by adding special chemicals: lectin proteins from beans to trigger cell division and colchicine from a flower to arrest cell division in the stage where chromosomes have just been copied.

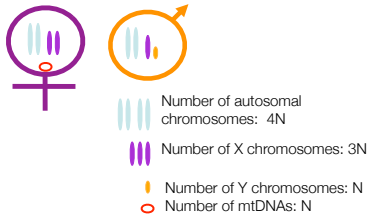


The complement of chromosomes in human gametes (haploid = single copy of each chromosome) and body (somatic) cells (diploid two copies of each chromosomes except for males who have a single X chromosome and a Y chromosome). 22 pairs of **autosomal** chromosomes and a pair of X chromosomes for females, or a single X- and single Y chromosome for males. X and Y chromosomes are known as **sex chromosomes** or **allosomes**. The **karyotype** (full complement of all chromosomes) of a female and male cell is illustrated on the right.

Ova fertilized by X-bearing sperm usually (98% of cases) give rise to females (persons with ovaries, fallopian tubes, uterus, cervix and vagina) whereas ova fertilized by sperm bearing an Y-chromosome usually give rise to males (persons with testes, epididymides, vas deferens, and a penis). About 2 % of the population develops with ambiguous genitalia for many possible reasons, most commonly hormonal congenital adrenal hypoplasia (**CAH**, a group of genetic disorders that affect the adrenal glands), or only vary rarely due to sex chromosome aneuploidy (absence or extra copy of a sex chromosome) see end of lecture.

Interesting fact: sperm do not actually carry chromosomes as the 23 segments of DNA of sperm cells are mostly removed from the histone proteins and tightly wrapped around protamine proteins so as to fit snuggly into a minuscule sperm head.

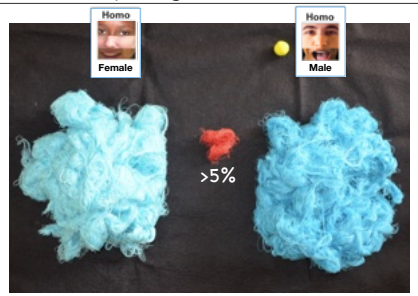
Different parts of the genome have different effective population sizes



Y-chromosome and mtDNA have shallower coalescence and they are affected by genetic drift more than the rest of the genome!

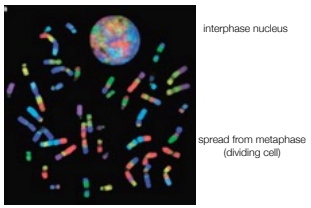
Due to their lower copy number in the “gene pool” Y chromosomes and mitochondrial variants are more easily lost from the population by chance (i.e. they are more strongly affected by genetic drift).

1000 x model of haploid genome



By virtue of having only one X chromosome (about 5.1 % of the genome) and a Y chromosome (~2% of the genome) human males differ from human females by more than 5%! in then haploid genome, given that humans are diploid and men have an X-chromosome, the diploid genomes differ by half as much.

Fitting a genome into a cell: genome packaging



Human chromosomes painted with probes from sorted gibbon chromosomes

Ferguson Smith 1997 ***European Journal of Human Genetics***

Chromosome painting is a form of fluorescence in situ hybridization (FISH) that has been highly productive in the construction of chromosome homology maps. The technique is described here using a gibbon–human comparison. A human metaphase and interphase nucleus is shown in panel a after hybridization with a chromosome-specific paint probe set that was derived from gibbon chromosomes. The probe set was made from a fluid suspension of gibbon chromosomes that were sorted and separated in a dual laser flow cytometer. Several hundred of each pair in the karyotype were collected in separate tubes. DNA in each tube was amplified by random-primed PCR and labelled with a combination of five fluorochromes so that each chromosome-specific DNA had a unique colour combination. A mixture of the complete set of labelled DNA probes was then hybridized in annealing conditions to denatured human metaphases that were fixed and air-dried onto microscope slides. Under these conditions, the gibbon paint probes anneal to complementary DNA sequences on human chromosomes, and the result (as shown in panel a) can be observed by digital fluorescence microscopy.

CHROMOSOME PAINTING
of Humans, Chimps, Gorillas, Orangs, and Macaque
Painted with fluorescent gibbon probes

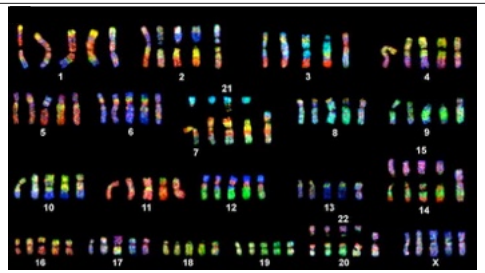
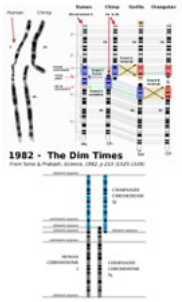


image by Stefan Müller. Universität München

Remarkable similarity between human and great ape chromosomes when painted with sorted and fluorescently labeled gibbon chromosomes.

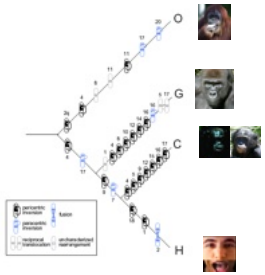
Chromosome fusion e.g. human chromosome 2



Details of the fusion of two ancestral ape chromosomes giving rise to the human chromosome 2.

Taxonomy by CHROMOSOMES

Cytogenetics:
the study of chromosomes
first visible DNA



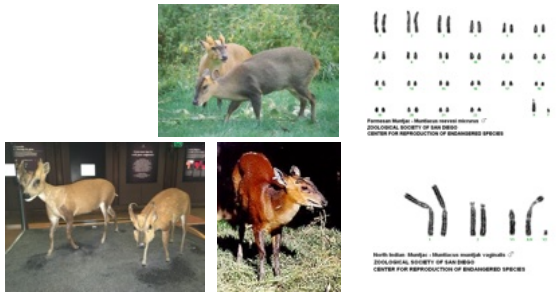
Modified from Mark A. Jobling ,Matthew Hurles ,Chris Tyler-Smith (2003)

Inversion, translocation, fusions and some more complicated changes in chromosome organization can be mapped onto the phylogeny of hominids.

Question:

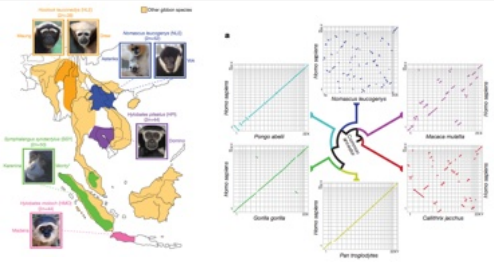
- Can similar species have radically different karyotypes (number and arrangement of chromosomes)?

Chromosomal (karyotypic) evolution
(still poorly understood)



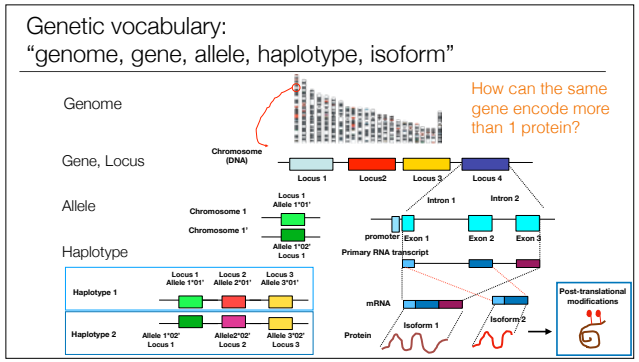
Similar looking mammals (barking deer) have either 46 or 7 chromosomes!

Gibbon Karyotype Evolution



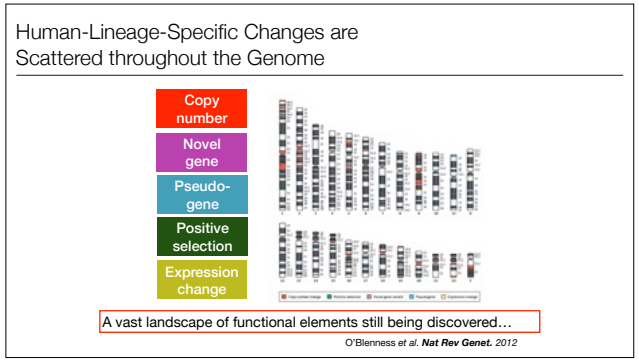
Gibbon genome and the fast karyotype evolution of small apes, ***Nature*** 2014 September

Assembly and analysis of a northern white-cheeked gibbon (Nomascus leucogenys) genome. We describe the propensity for a gibbon-specific retrotransposon (LAVA) to insert into chromosome segregation genes and alter transcription by providing a premature termination site, suggesting a possible molecular mechanism for the genome plasticity of the gibbon lineage. We further show that the gibbon genera (Nomascus, Hylobates, Hoolock and Symphalangus) experienced a near-instantaneous radiation ~5 million years ago, coincident with major geographical changes in southeast Asia that caused cycles of habitat compression and expansion. Finally, we identify signatures of positive selection in genes important for forelimb development (TBX5) and connective tissues (COL1A1) that may have been involved in the adaptation of gibbons to their arboreal habitat.

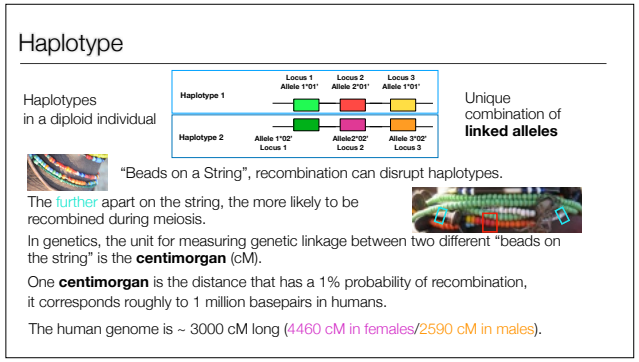


Some key terms used in genetics: genome, gene, locus ("site" in Latin), allele, haplotype, promoter, exon, intron, mRNA, alternative splicing, protein isoform, post-translational modification

Haplotypes are long stretches of DNA that carry unique combinations of genetic variants (alleles). Post-translational modifications of proteins include the addition of sugar (glycosylation), or phosphate (phosphorylation) etc… Among other things, these modification regulate the function of proteins.



Ape genomes are about 2 meters long 3 billion basepairs on 23 or 24 different chromosomes. Human changes are scattered throughout this vast landscape: 5% of our DNA differs from that of the chimpanzee genome if deletions and duplications are included. The search is on identifying key genetic changes. Important changes range from single nucleotide (DNA "letters") changes to large duplications or deletions, differences in gene copy number and altered regulatory DNA. In addition, identical genes can be expressed at different times and in different tissues, creating large differences in the organism.



The human female genome is 4460 cM long, while the male genome is only 2590 cM long

Morgan's Chromosomal Theory of Inheritance

T.H. Morgan:

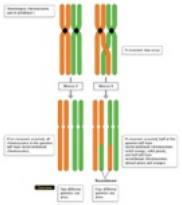
Chromosomal Theory of Inheritance:
Genes are located on chromosomes like beads on a string.

Some genes are physically linked (same chromosome), and thus are (almost) always inherited together (early description of linkage disequilibrium).

"Exceptions" to Mendel's Second Law: Independent assortment
Noticed that "linked" traits would occasionally separate.

Other traits on the same chromosome showed little detectable linkage.

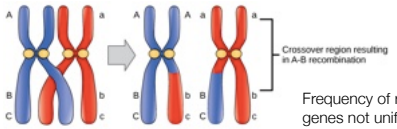
Proposed that recombination, might explain his results.
Proposed that the two paired chromosomes could "cross over"
to exchange information during prophase of meiosis,
and this produces different combinations of alleles in the gametes.



33

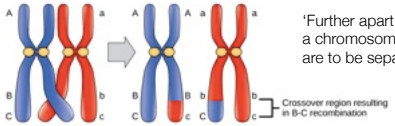
Morgan's insights gained first at **Columbia University** in New York and then at **Caltech** in Pasadena while working with fruit flies.
Genetic Linkage is defined as genetic variants along the same strand of DNA being inherited together.

Recombination



Frequency of recombination between genes not uniform.

'Further apart two genes are on a chromosome, the more likely they are to be separated by recombination'



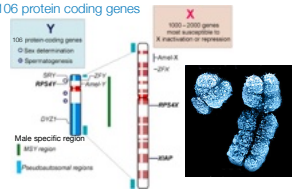
https://www.nature.com/articles/scientificamerican0406-44/figures/1

34

Example of most common type of recombination by crossing over between **homologous chromosomes**.

The human Y-chromosome, mostly one haplotype

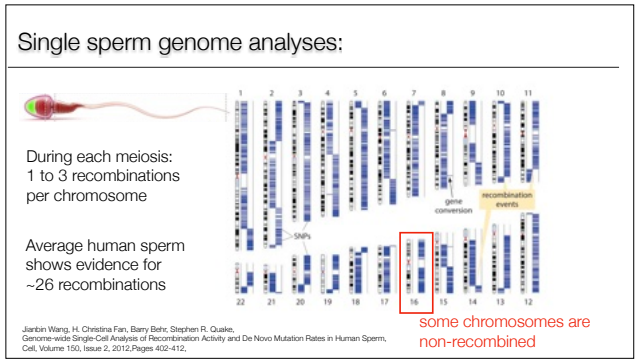
lates T2T Y data:
106 protein coding genes



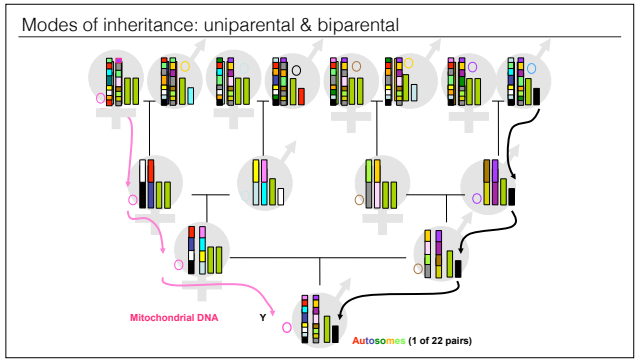
Human **mitochondrial DNA** is also one haplotype:

just small caps of the Y chromosome,
the pseudoautosomal regions (PAR),
recombine with homologous regions on the X!

The Y-chromosome: **largest haplotype** in the genome. Recent telomere to telomere sequencing (T2T) has produced the complete sequence for the 24th human chromosome, the Y.
Both X and Y have a small "**pseudoautosomal**" region, which can recombine.

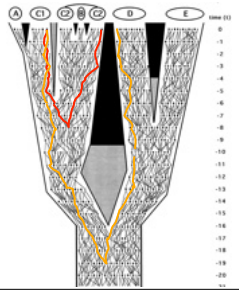


Detection of recombination events based on sequencing of a **single sperm cell**.
The two columns in each chromosome represent the two homologous chromosomes carried by the subject.
The source of the sperm single chromosome copy can be traced to one or the other homologous chromosome based on single nucleotide polymorphisms that appear in one chromosome but not the other. Blue lines show the association of the sperm sequence to the two chromosome sets based on those single nucleotide polymorphisms. Each switch (**haplotype block**) indicates a recombination event. (Adapted from J. Wang, et al., Cell 150:402,2012.)
Human females have ≈**50% higher recombination rates** than males (42 versus 28 on average in onerecent study: Hussin J, Roy-Gagnon M-H, Gendron R, Andelfinger G, Awadalla P (2011) Age-Dependent Recombination Rates in Human Pedigrees. ***PLoS Genet*** 7(9): e1002251. https://doi.org/10.1371/journal.pgen.1002251.
So even though you tend to get **more of your single base mutations from your father**, your **crossovers come mostly courtesy of your mother.**



Most DNA is on autosomal chromosomes (all but the X or Y sex chromosomes).
These autosomal chromosomes get reshuffled when eggs and sperm are produced.
Each one of us is a genetic mosaic that only exists in its present combination once in the history of the universe!
Most of the Y-chromosome and the mitochondria DNA are not subject to recombination.

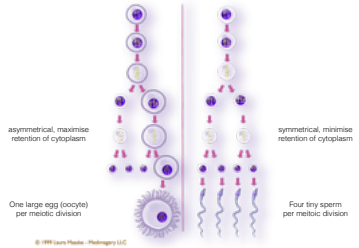
DNA trickling through generations

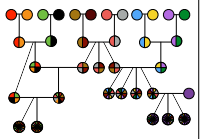


Different pieces of our genomes share common ancestors at various time depths (number of past generations back)

Sex: Meiosis (reduction division, from diploid to haploid)

Mixing of genomes via meiosis and fusion of gametes



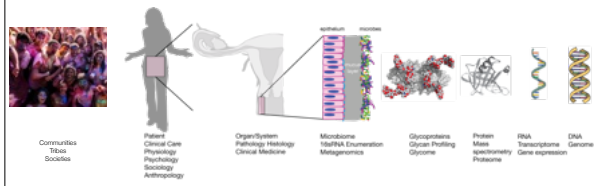


Each generation, half of the reshuffled DNA from two parents forms a new individual

The core aspects of sex in eukaryotes. For simplicity, the figure shows a hypothetical organism in which the whole genome is carried in a single chromosome. The sexual cycle starts with a diploid cell that contains two different copies of the genome on a pair of homologous chromosomes. Each chromosome is first replicated to produce two genetically identical chromatids. The chromosomes then line up and exchange genetic material through recombination, producing chromatids that contain a mix of genetic material from both chromosomes. A two-stage meiotic division then leads the production of haploid gametes, each containing a single chromatid—half of the genetic material of the original diploid cell. Completion of the sexual cycle requires that diploidy is restored through the fusion of two gametes, usually from two different individuals. The production of eggs (ova) maximizes the size and content of cytoplasm, including nutrients by generating a single egg from a meiotic division, that in the case of sperm produces four tiny sperm, each with its unique combination of reshuffled and halved genome.

Scales: Societies to Atoms….

What scale are we considering?



kilometers
km=10^3 m

meters

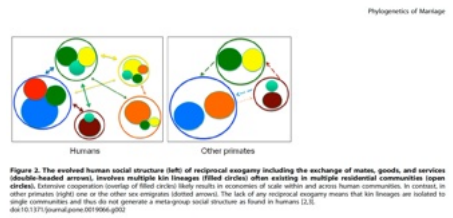
millimeters
mm =10^{-3} m

micrometers
um=10^{-6} m

nanometers
nm=10^{-9} m

Our molecular nature ranges from atoms to societies.
Glycosylation is one of the most common forms of PTM, each monosaccharide measures around 1 nm (1 meter divided by 1000 three times in a row).

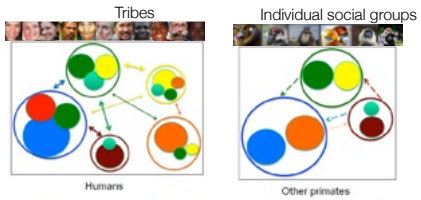
Reciprocal exogamy and meta-groups



Walker et al. ***PlosOne*** 2011

Meta-groups (groups of groups) do not seem to exist in other primates. Together with personal names and language, these groups allow the emergence of tribes, sharing belief systems, symbolic identity and language.

Small groups - large networks



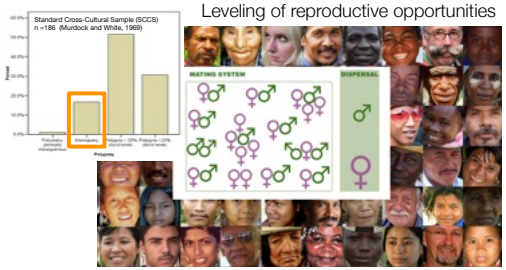
reciprocal exogamy
multiple kin lineages in multiple communities
socio-cultural niches

Walker et al. **PlosOne** 2011

The evolved human social structure (left) of reciprocal exogamy including the exchange of mates, goods, and services (double-headed arrows), involves multiple kin lineages (filled circles) often existing in multiple residential communities (open circles). Extensive cooperation (overlap of filled circles) likely results in economies of scale within and across human communities. In contrast, in other primates (right) one or the other sex emigrates (dotted arrows). The lack of any reciprocal exogamy means that kin lineages are isolated to single communities and thus do not generate a meta-group social structure as found in humans. Kin lineages in humans are directly identified by language., essentially allowing the invention of “tribes”.

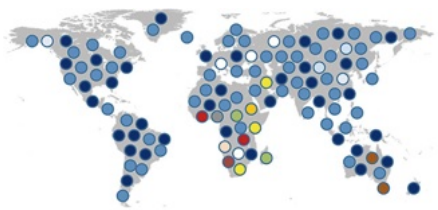
Practice question: Explain the term reciprocal exogamy.
Repeated exchanges of mates, goods and services between different social groups.

Finding a Mate



The majority of documented traditional human societies tolerated some degree of polygamy by men. Individually, the majority of men appear to have been (serially) monogamous. Human mating system are remarkably diverse and do include some with polyandry (one woman with multiple male partners).

Genetic diversity in current human populations

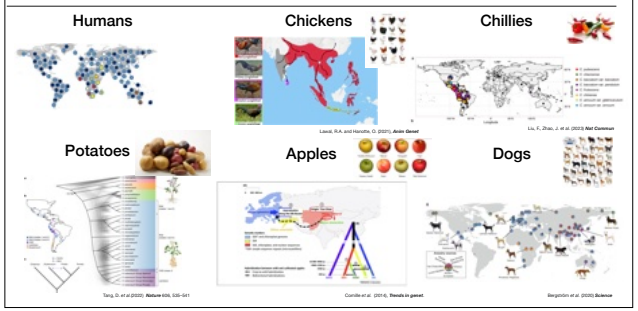


The rest of the world is a subsample of Africa !

The figure shows genetic diversity in current human populations. The colored circles highlight different genetic variants. Due to the African origins of mankind, Sub-Saharan populations have greater genetic diversity than other populations and greater effective population size. The genetic diversity observed outside Africa is a subset of African diversity due to the founder effect that took place some 50,000 years ago with the first great human migration. / David Comas

---

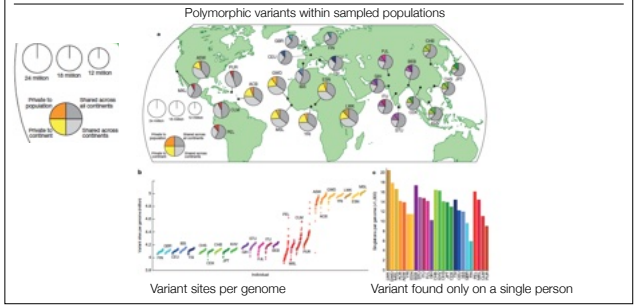
Highest Genetic Diversity points to Place of Origins



The figure shows genetic diversity in current human populations. The colored circles highlight different genetic variants. Due to the African origins of mankind, Sub-Saharan populations have greater genetic diversity than other populations and greater effective population size. The genetic diversity observed outside Africa is a subset of African diversity due to the founder effect that took place some 50,000 years ago with the first great human migration. / David Comas

---

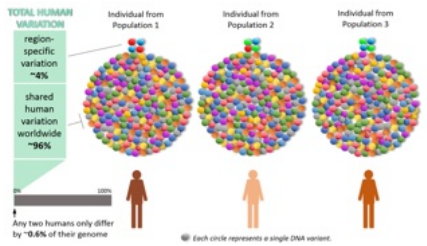
The largest Fraction of Diversity is found Within Populations



Population sampling.
a, Polymorphic variants within sampled populations. The area of each pie is proportional to the number of polymorphisms within a population. Pies are divided into four slices, representing variants private to a population (darker colour unique to population), private to a continental area (lighter colour shared across continental group), shared across continental areas (light grey), and shared across all continents (dark grey). Dashed lines indicate populations sampled outside of their ancestral continental region.

b, The number of variant sites per genome.
c, The average number of singletons per genome (In a large set of genetic samples, a singleton is a variant that appears in only one person.)

DNA variation specific to regions: a tiny fraction

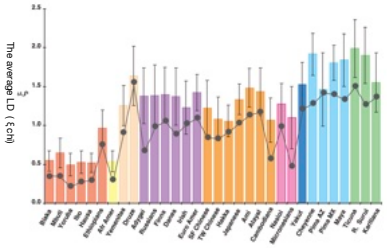


Scale of the DNA differences between populations of different regions of the world. Only a tiny fraction of the whole of the natural human DNA variation is actually more specific to certain geographic regions. And this total DNA variation among humans is, in itself, only a tiny fraction of the genome.

The average Linkage Disequilibrium (LD) for 83 SNPs across 21 haplotypes for 32 populations

based on published data on CD4, DM1, DRD2, DRD4, PAH and COMT plus unpublished data.

Increase in LD with distance from Africa provides one of the strongest evidence of our common African ancestry.

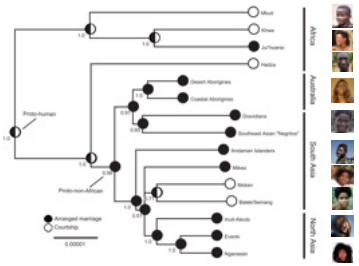


Tishkoff & Kidd VOLUME 36 | NUMBER 11 | NOVEMBER 2004 NATURE GENETICS SUPPLEMENT

The average LD (ξ chi) for 83 SNPs (single nucleotide polymorphisms) across 21 haplotypes for 32 populations. LD is measured as the ξ coefficient, a **standardized measure of overall nonrandomness of alleles at the sites in the haplotypes**. The bars are the mean values of ξ across the same 21 independent haplotype systems in all populations. The standard errors of the means are given as the error bars and the median values are plotted as dots connected by the line. Bars are color-coded by geographic region of origin of the populations, from left to right as sub- Saharan Africa, African Americans, Southwest Asia, Europe, East Asia, Pacific, Siberia, North America and South America. Population and sample descriptions are in ALFRED69.
Different samples of populations with the same name are distinguished by initials: SF, San Francisco; TW, Taiwan; AZ, Arizona; MX, Mexico; R, Rondonian. The haplotyped loci were

chosen with no prior knowledge of LD values at the locus. The number of sites per haplotyped locus varied from 2 to 7 for a total of 83 SNPs. The graph is based on published data on CD4, DM1, DRD2, DRD4, PAH and COMT plus unpublished data. These results show less LD in African populations than elsewhere and greater LD in the Native American populations than in other regions, as well as variation in LD within geographic regions.

Evolutionary History of Hunter-Gatherer Marriage Practices

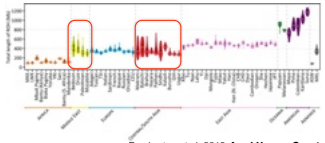


Walker *et al.* ***PlosOne*** 2011

Many documented hunter and gatherer societies have arranged marriages.

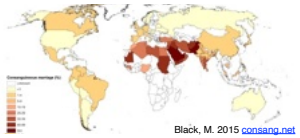
Human culture shaping the genome via marriage patterns

Length of runs of homozygosity



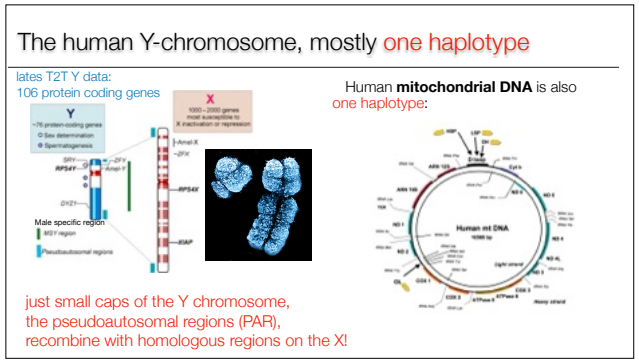
Pemberton et al. 2012 ***Am J Human Genet***

Prevalence of consanguineous marriages

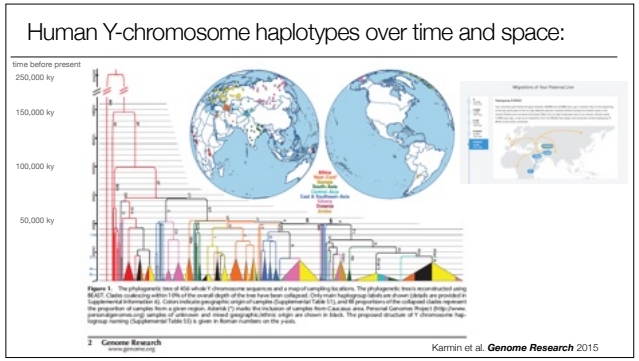


Black, M. 2015 consang.net

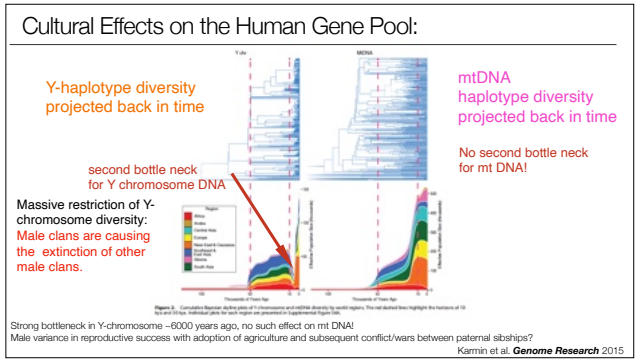
Impact of close-relative marriage (uncle niece or first cousins) on homozygosity (length of DNA segments with no evidence of recombination). Marriage patterns are literally helping to mold human genomes.



The Y-chromosome: **largest haplotype** in the genome
Both X and Y have a small "**pseudoautosomal**" region, which can recombine.

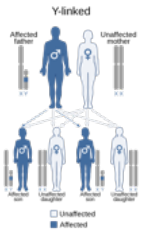


Phylogeny of Y-chromosomes showing deeper divergences (highest levels of variation) in Africa.
One of the many pieces of strong evidence for **our shared African ancestry**.



Extinction of entire Y-chromosome lineage due to lethal conflicts between paternal kinship clans?
Coalescence analysis combined with **ancient DNA** allows the reconstruction of past changes in diversity
(estimates of **effective population size**, or N_e, a theoretical concept about genetic diversity and population size)

Y-linked genes and disease



The majority of the ~60 Y-chromosome genes are located in the "non-recombining region"

They are inherited together as one haplotype!

- ASMTY (acetylserotonin methyltransferase), melatonin biosythesis
- TSPY (testis-specific protein), 35 copies exist, spermatogenesis
- IL3RAY (interleukin-3 receptor),
- SRY (sex-determining region),
- ZFY (zinc finger protein),
- PRKY (protein kinase, Y-linked),
- ANT3Y (adenine nucleotide translocator-3 on the Y),
- AZF2 (azoospermia factor 2),
- BPY2 (basic protein on the Y chromosome) Male germ cell development
- AZF1 (azoospermia factor 1),
- DAZ (Spermatogenes is deleted in azoospermia),
- RBM1 (RNA binding motif protein, Y chromosome, family 1, member A1), spermatogenesis
- RBM2 (RNA binding motif protein 2), spermatogenesis
- UTY (ubiquitously transcribed TPR gene on Y chromosome) Minor Histocompatibility Antigen
- USP9Y (spermatogenesis, mutate s assoc with Sertoli cell only syndrome)
- AMELY (biomineralization e.g. tooth enamel)

Inheritance pattern allows easy identification of Y-linked genes

The vast landscape of the genome

The "Gene" Crowd

Zócalo in Mexico City, Spencer Tunick 2007

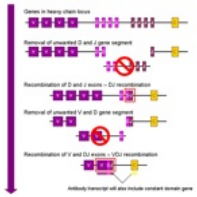
- ~23,000 **protein coding genes** (UCSD undergrads in 2012!)
- 1000s of **long non-coding RNA** "genes"
- > 1.5 Million **enhancers = cis regulating elements** in genome

(friends and relatives of UCSD undergrads!)

The notion of "genes" defined as stretches of DNA encoding a protein, have become limiting. There are thousands of other functional elements in our genomes that have regulating functions. Networks of co-expressed genes with hub-like transcription factor proteins that orchestrate levels of co-expression. The new definition of a gene is "a segment of DNA encoding proteins or RNA". The majority of genes have multiple functions: they are **pleiotropic.** Such pleiotropic effects can take place in different tissues and at different times during our lives, they can even be "good" early in life and "bad" late in life, in which case they are known as antagonistic pleiotropy. Senescence (aging) has been explained by antagonistic pleiptropy, whereby genes with positive effects really in life are selected even though they may have negative effects much later in life. Williams G.C. (1957). "Pleiotropy, natural selection, and the evolution of senescence". ***Evolution***. 11 (4): 398–411.

Question:

- It is estimated that each of us makes over 10^{12} different antibodies.
- How can a genome code for more proteins than it has genes for?



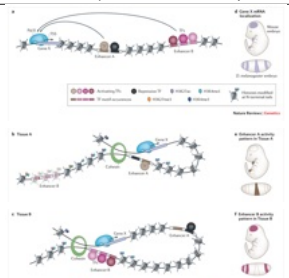
Individual B-cells and T-cells can diversify their receptors (T-cell receptors and Antobodies) by shuffling cassettes of genes/ somatic recombination.

Enhancers (more than a million) and Transcription factors

What are Transcription Factors (TFs)?

orchestrators of gene expression

Transcription Factors:
- Proteins
- Pressure
- Cytoarchitecture
- Social partners



Daria Shlyueva, Gerald Stampfel & Alexander Stark ***Nature Reviews Genetics*** 15, 272–286 (2014)

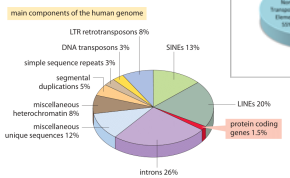
Each protein-coding gene in our genomes is influenced by hundreds of enhancer sequences: Non-coding DNA sequences that are recognized by transcription factors and influence the expression of genes near and far!

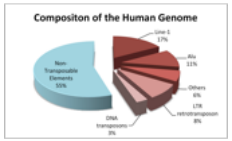
Enhancers are distinct genomic regions (or the DNA sequences thereof) that contain binding site sequences for transcription factors (TFs) and that can up-regulate (that is, enhance) the transcription of a target gene from its transcription start site (TSS). Along the linear genomic DNA sequence, enhancers can be located at any distance from their target genes on the same chromosome, which makes their identification challenging. b,c In a given tissue, active enhancers (Enhancer A in part b or Enhancer B in part c) are bound by activating TFs and are brought into proximity of their respective target promoters by looping, which is thought to be mediated by cohesin and other protein complexes. Moreover, active and inactive gene regulatory elements are marked by various biochemical features: active promoters and enhancers are characterized by a depletion of nucleosomes, which is the structural unit of eukaryotic

chromatin. Nucleosomes that flank active enhancers show specific histone modifications, for example, histone H3 lysine 4 monomethylation (H3K4me1) and H3K27 acetylation (H3K27ac). Inactive enhancers might be silenced by different mechanisms, such as by the Polycomb protein-associated repressive H3K27me3 mark (part b) or by repressive TF binding (part c). d–f Complex patterns of gene expression result from the additive action of different enhancers with cell-type- or tissue-specific activities.

45% of Genome comprised of Transposons (“parasitic DNA”)

A.k.a. “jumping Genes” discovered in maize by Barbara Mc Clintock





Repetitive DNA, was initially excluded from DNA comparisons, thus only 1. 3 % difference reported between human and chimp DNA.
If one includes repetitive DNA (mostly transposable elements) the difference is ~5%.
Half of the human genome is made up of “parasitic DNA”/transposons.

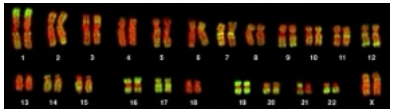
Gene evolution - How many genes?

- 3 billion base pairs in haploid genome
- 22 000 genes (180,000 axons)
- Alternative splicing (40-60 % of genes)
- Protein coding ~ 20 000? (1.5% of genome)
- Structural RNAs ~ 3 000?
- 2 % highly conserved non protein coding
- Gene structure
- Gene expression
- Post-translational modification
- Microbiome genomes

A few stats about our genomes

Alu elements in primates

Over 1 million *Alu* element insertions in the human genome!

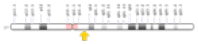


Karyotype from a female human lymphocyte (46, XX). Chromosomes were hybridized with a probe for Alu elements (green) and counterstained with TOPRO-3 (red). Alu elements were used as a marker for chromosomes and chromosome bands rich in genes.

Alu elements are named after *Arthrobacter luteus* bacteria. An enzyme from this bacteria cuts DNA at a sequence carried by all these million of copies of a ~ 300 basepair element.

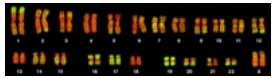
Neurofibromatosis, *NF1* gene

(tumor suppressor gene, "control of cell division")



Neurofibromin protein acts as a tumor suppressor.

Insertion of transposable elements called **Alu elements**, a type of Short Interspersed Nuclear Element (**SINE**).

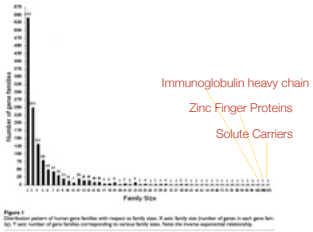


Alu elements, chunks of "jumping DNA" make up almost 40% of our genome! Their activity can disrupt genes and contributes to evolution. The green fluorescence on this karyotype of a human cell indicates the presence of *Alf* elements.

Bolzer et al., (2005) ***PLoS Biol*** 3(5)

The NF1 gene provides instructions for making a protein called neurofibromin.Neurofibromin acts as a **tumor suppressor protein.** Tumor suppressors normally prevent cells from growing and dividing too rapidly or in an uncontrolled way. This protein appears to prevent cell overgrowth by turning off another protein (called RAS) that stimulates cell growth and division.
NF1 is the most common neurological disorder caused by a single gene; occurring in one in every 3,000 children born.
Autosomal dominant.

Gene Families and their sizes

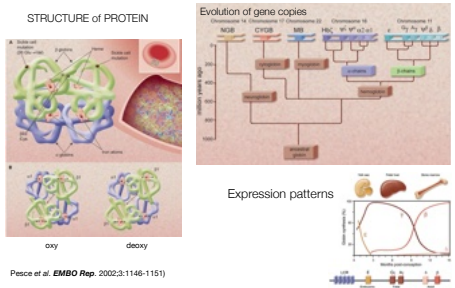


Over half of all genes exist in more than one copy…..

Sharma et al, ***BMC Genomics*** 2005;6:83

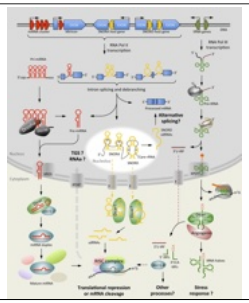
Multiple copies of the same gene allow for adaptation by tweaking the function of each copy…..

Hemoglobin Evolution



The X-ray determined structure of the hemoglobin molecule and a representation of its very high concentration in the erythrocyte. (A) The arrangement of the -helices (shown as tubes) in each unit—one on the left and one, 180° rotated, on the right—is shown, as are the 4 heme groups with their iron atoms where gas molecules bind. The site of the sickle mutations on mutant -chains as well as the 93 conserved cysteine residues is also shown. Hemoglobin molecules in the red blood cell, shown in an inset on the right, are very tightly packed (at a concentration of approximately 34 g/dL) and have little access to solvent; this allows efficient oxygen transport by each cell but also affects the chemical behavior of the molecules, such as promoting sickle cell hemoglobin polymerization upon slight deoxygenation. (B) A representation of the quaternary structural changes in the hemoglobin tetramer, in a top-down view, in the transition from the oxy conformation (left) to the deoxy conformation (right). The iron atoms shift relative to the planes of the heme groups and a central cavity between the -chains opens, facilitating 2,3 BPG binding. These diagrams are based on drawings of Irving M. Geis. Figure 2. A diagram of the proposed evolutionary relationships of the human globin proteins as inferred from sequence analyses. NGB, neuroglobin; CYGB, cytoglobin; MB, myoglobin.

RNA complications:



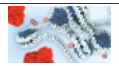
tRNA (transfer)
rRNA (ribosomal)
micro RNAs
sdRNA (sub-genomic)
SnoRNA (small nucleolar)
siRNA (small interfering)
piRNA (piwi interacting)
CRISPER RNA (Clustered Regularly Interspaced Short Palindromic Repeats)

Martens-Uzunova et al. ***Cancer Letters*** 2013

Cross-talk between the pathways of biogenesis and function of miRNAs, snoRNAs, tRNAs, sdRNAs and tRFs. MiRNAs are encoded in clustered genomic loci or in the introns of other genes and are transcribed by RNA polymerase II (RNA Pol Pri-miRNA transcripts are processed to individual pre-miRNAs in the nucleus by Drosha. SnoRNAs and intron-encoded miRNAs are produced after splicing, debranching, and exonucleolytic trimming. Pre-miRNAs are exported to the cytoplasm by Exportin 5 (XPO5) and further processed by Dicer to mature miRNAs that enter the RNA-induced silencing complex (RISC). Cytoplasmically matured miRNAs in complex with proteins from the Ago family (AGO) may be imported back in the nucleus possibly by Importin 8 (IPO8) to participate in the processes of transcriptional gene silencing (TGS) or RNA activation (RNAa). SnoRNAs assemble with snoRNP-core proteins (not shown) and enter the nucleolus where they participate in the chemical modification of ribosomal RNA (rRNA) and other RNA species. SnoRNAs may be exported to the cytoplasm by unknown transporter proteins, where they are cleaved possibly by Dicer to short ~22 nt long sdRNAs and are loaded into RISC. Alternatively, snoRNAs may also be cleaved by unknown nucleases in the nucleus or nucleolus, to sdRNAs with a different size. Longer sdRNAs of ~27 nt do not exit the nucleus, but instead participate in the regulation of alternative splicing. tRNAs are transcribed from individual tRNA genes by RNA polymerase III (RNA Pol III). Pre-tRNA transcripts are processed by the endonucleases RNase P and RNase Z to remove 5′- and 3′-trailer sequences, and after chemical modification, CCA addition, and aminoacylation, are exported to the cytoplasm by Exportin-t (XPOT) to participate in protein synthesis. 3′U tRFs are produced by RNase Z after trimming of the 3′-trailer sequence. Stress factors may induce cleavage in the anticodon loop of mature tRNAs to tRNA halves performed by the endonuclease Angiogenin. Shorter 5′tRFs and 3′CCA tRFs may be produced from 5′- and 3′-ends of mature tRNAs by Dicer and associate with AGO proteins to participate in various processes of transcriptional and post-transcriptional regulation. sdRNA = subgenomic RNA SnoRNA=small nucleolar RNA MiRNA

Genome “syntax”

Chromatin Remodeling
“genome packaging” and its effects on gene expression via access for transcritpion factors, enhancers and transcriptional machinery


Packaging

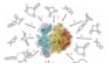
Histone Modification
annotation of histone and effects on gene expression. methylation, acetylation, ubiquitination, O-GlcNAcylation

Annotation of packaging

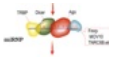
DNA Methylation
annotation of DNA, silencing of paternal or maternal allele, or both.

Annotation of DNA

non-coding RNA
(micro, piwi, nc, circular RNA etc.. interact with ribosomal proteins, transcription factors, messenger RNA)


RNA with novel function

RNA-binding Proteins
Approximately 1000 RBP in nucleus, cytoplaams and mitochondria regulate splicing, translation, degaradation


RNA - Protein interactions

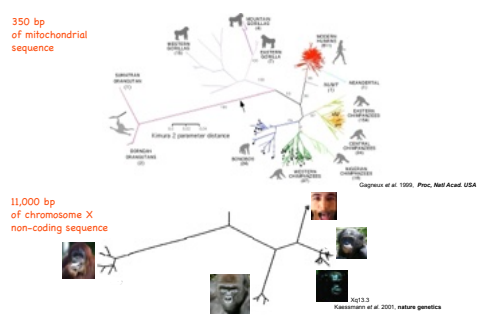
THE GENOME is much **more complicated than most of us think**, possibly more complicated than we can think!
Information in the genome includes **several layers above the simple DNA sequence**: packaging, modification of histones, modification of DNA, microRNA and RNA-binding proteins all affect regulation of gene expression.

Information from comparative genomics:
Demographic History vs Biology

- Whole genome sequencing of multiple individuals:
  - Reconstruction of phylogeny
  - Reconstruction of past population dynamics (demography)
  - Reconstruction of past episodes of selection
  - Comparative genomics to look for shared features, uniquely derived features and their effects on each species

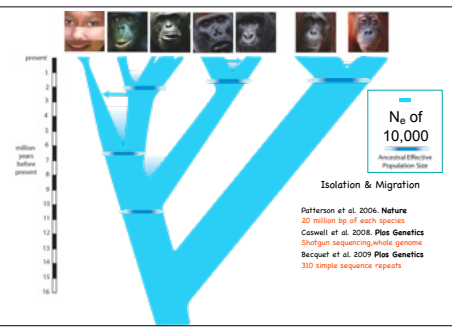
To reconstruct the evolutionary history, there more parts of the genome is available, the better the reconstruction will become.
Currently, it is realistic to obtain whole genome data fro most species. Reconstructing population dynamics relies on the use of parts of the genome that are “selectively neutral”. Most of the genome actually does not see to be under strong natural selection and the majority of mutations are neutral (no advantage and no down side). Detecting past natural selection in present day genomes is difficult. Most commonly, one looks for sequence that appear more different between closely related species than one would expect under “neutral evolution” which is time since last common ancestor times the mutation rate. Once identified, uniquely evolved parts of the genome can be studied in model animals or cell culture for their effects on biology.

Hominid phylogenies, mitochondrial and X-linked DNA



Years ago, I collaborated with a large group of people to compare the DNA sequences of a small stretch of mitochondrial DNA. We reported that each of the grat ape species showed much more genetic vaiation than 800 humans from populations from all around the world. A few years later, Svante Pääbo’s group sequenced a strecth of DNA 30 times longer on noncoding parts of the X chromosome and found a similar pattern. Now we have whole genomes for all these players including Neanderthals.

Hominid Effective Population Sizes (N_e): Past and Present



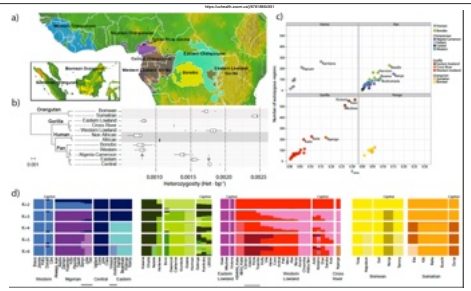
N sub e Definition:

The number of breeding individuals in an idealized population whose genetic diversity is subject to the same effects of genetic drift and inbreeding as the population under consideration.

Uncertainty due to recombination, selection? Different parts of the genome have differing histories. The more variance there is in coalescence times (number of generations back to the last common ancestor) across the genome, the larger the ancestral populations were.

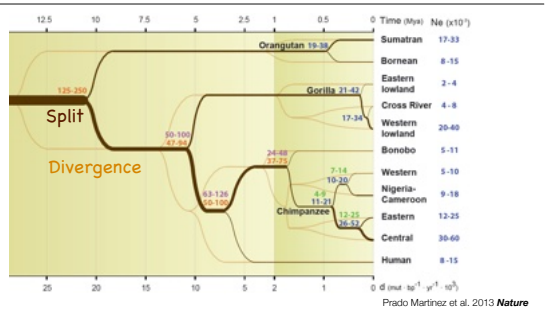
Great ape genetic diversity and population history

whole genome of 78 great apes Prado Martinez et al 2013 NATURE



Samples, Heterozygosity and genetic diversity. a. Geographical distribution of great ape populations across Indonesia and Africa sequenced in this study. The formation of the islands of Borneo and Sumatra resulted in the speciation of the two orangutan populations. The Sanaga River forms a natural boundary between Nigeria-Cameroon and Central chimpanzee populations while the Congo River separates the bonobo population from the Central and Eastern chimpanzees. Eastern lowland gorillas and Western lowland gorillas are both separated by a large geographical distance. b. Heterozygosity estimates of each of the individual species and subspecies are superimposed onto a neighbor-joining tree from genome-wide genetic distance estimates. Arrows indicate heterozygosities previously reported[4] for Western and Central chimpanzee populations respectively. An almost fourfold range of heterozygosity is observed among different great ape populations. c. Runs of Homozigosity among great apes. Relationship between the coefficient of inbreeding (F_{ROH}) and the number of autozygous >1Mbp segments. Bonobos, and Eastern lowland gorillas show an excess of inbreeding compared to the other great apes, suggesting small population sizes or fragmented population. d. Genetic structure based on clustering algorithm of great apes. All individuals (columns) are grouped in a different cluster (K=2 to K=6, rows) colored by species and according to their common genetic structure. Some groups, such as Western lowland gorillas, present a transitional clustering pattern, while other groups, such as chimpanzee, show a clear distinct pattern according to the known subspecies. Most captive individuals, labeled on top, present a complex admixture from different wild populations. A signature of admixture, for example, is clearly observed in the known hybrid, Donald, a second-generation captive where we predict 15% admixture of Central chimpanzee on a Western background consistent with its pedigree. A gray line has been added at the bottom of the new groups according to location of origin or ancestral admixture.

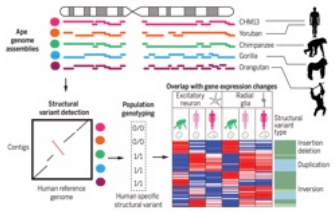
Divergence, Split and Effective Population Sizes (N_e)



Prado Martinez et al. 2013 ***Nature***

Inferred population history. Population splits and effective population sizes (Ne) during great ape evolution. Split times (dark brown) and divergence times (light brown) are plotted as a function of divergence (d) on the bottom and time on top. Time is given using a single mutation rates ($\mu = 1 \cdot 10^{-9}$ mut/(bp·year)). The ancestral and the current effective population sizes are also given depending on μ; the methods used in different periods of time COALHMM and PSMC are colored in orange and blue, respectively. The chimpanzee split times are estimated using the ABC method. The x-axis is rescaled for divergences larger than $2 \cdot 10^{-3}$ to provide more resolution in recent splits.

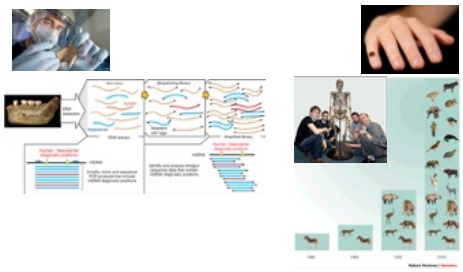
High coverage ape genomes (65X)



Kronenberg let al., ***Science***, 2018

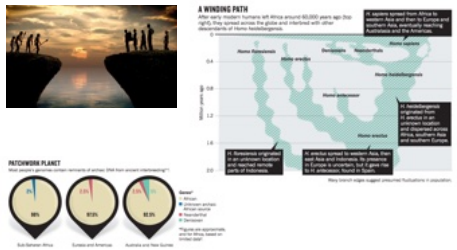
SMRT assemblies and SV analyses. (Top) Contiguity of the de novo assemblies. (Bottom, left to right) For each ape, SV detection was done against the human reference genome as represented by a dot plot of an inversion). Human-specific SVs, identified by comparing ape SVs and population genotyping (0/0, homozygous reference), were compared to single-cell gene expression differences [range: low (dark blue) to high (dark red)] in primary and organoid tissues. Each heatmap row is a gene that intersects an insertion or deletion (green), duplication (cyan), or inversion (light green).

Ancient DNA



Several research group have been sequencing ancient DNA with increasing success. To date, the oldest hominin DNA sequenced is from Spain (Atapuerca), where DNA from *Homo heidelbergensis*, ancestor to Neanderthals from 400,000 years ago was obtained.
The partial finger bone from a young female yielded a good quality genome, that revealed a new group of extinct hominids called Denisovans, after the South Siberian Cave in which it was found.
Ancient protein from enamel from *Homo antecessor* 800,000 years old found in the Gran Dolina of Atapuerca. Several juvenile and cannibalized skeletons!

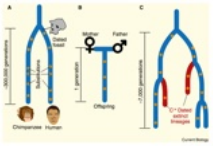
Archaic admixture



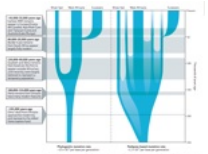
Comparisons between modern human genome sequences and those of two extinct hominids, Neanderthals that went extinct about 45 kya and Denisovans who's date of extinction is not known have shown that the genomes of present day humans outside Africa contain a few % from these extinct forms. Inside Africa, there is evidence of a similar admixture of archaic forms.

Mutation rates

- inferred from species comparisons: number of observed mutations, divergence estimates and generation times
- parent offspring trios: approx. 50 new mutations per individual, most of paternal origin.
- paleogenetics: comparing DNA sequences dating back 40, 30, 20 10 and 2 thousand years.



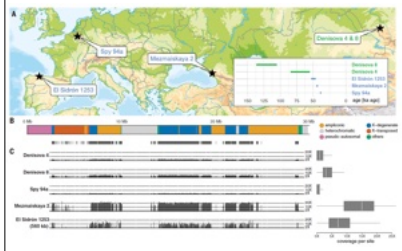
Green & Shapiro, ***Current Biology*** 2014

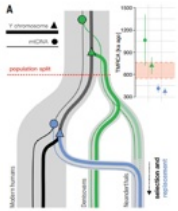


Veeramah & Hammer, ***Nature Reviews Genetics***, 2014

Key fossil evidence for nearly or fully anatomically modern humans (AMHs) is described on the left, and approximate date range estimates are indicated by the grey shading. West Africans are relatively genetically homogenous modern-day Niger–Kordofanian- and Nilo-Saharan-speaking populations that are often represented by the Yoruba of Nigeria. Eurasians encompass all modern-day non-African populations. Divergence times that are estimated using the faster phylogenetic mutation rate under the assumption of relatively instantaneous population splitting are mostly consistent with the fossil evidence. To preserve the correspondence between fossil dates and population divergence times under the slower pedigree-based mutation rate, this model assumes long-term gene flow among subdivided ancestral populations (represented by the gradient of blue shading), which leads to older divergence time estimates.

How Neanderthals lost their Y





Petr, M *et al.* The evolutionary history of Neanderthal and Denisovan Y chromosomes. ***Science*** 2020

3. Proposed model for the replacement of Neanderthal Y chromosomes and mtDNA. (A) Relationships between archaic and modern human mtDNA and Y chromosomes. The semitransparent Neanderthal lineage indicates a (as yet unsampled) hypothetical Y chromosome replaced by an early lineage
related to modern human Y chromosomes. Most recent common ancestors with modern human lineages are shown for mtDNA (circles) and Y chromosomes
(triangles). The inset shows TMRCAs for the four nodes in the diagram: Y chromosome TMRCAs as estimated by our study and mtDNA TMRCA estimates
from the literature (7, 8). The red shaded area highlights the 95% CI for the population split time between archaic and modern humans, shown as the dotted red horizontal line (6). (B) Probability of replacement of a nonrecombining,
uniparental Neanderthal locus over time, assuming a given level
of fitness burden relative to its modern human counterpart. Trajectories are based on forward simulations across a grid of parameters (figs. S27 to S29) (14),

with Ne of modern humans and Neanderthals fixed at 10,000 and 1000, respectively. Modern human introgression was simulated in a single pulse at 5%. Replacement probabilities from a wider range of model parameters are shown in fig. S31.

Types of Genetic Changes

size of DNA sequences involved:

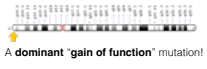
- Chromosomal rearrangements — 10^7 bp
  Deletions, inversions, duplications, fusion, Fission, translocation
- Segmental duplication — 10^4 - 10^6 bp
  intra- and inter chromosomal
- Endogenous retroviruses — 10^4 bp
- Transposable elements — 10^2 - 10^4 bp
  LINE, SINE
- Simple Sequence Repeats — 1 - 5 bp
  (Microsatellites)
- SNP = Single Nucleotide Polymorphisms — 1 bp

Mutations are changes to the genome. These can be gigantic changes involving fusion or inversion of million of basepairs during chromosome rearrangements , or much smaller events all the way down to single basepair mutations, occurring as copying errors during DNA replication.

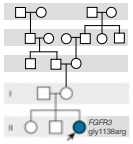
Achondroplasia, *FGFR3*

(gene for growth factor receptor, "signaling" kept on permanently)

Dr. Michael Craig Ain
Orthopedic Surgeon,
Johns Hopkins



A **dominant** "**gain of function**" mutation!



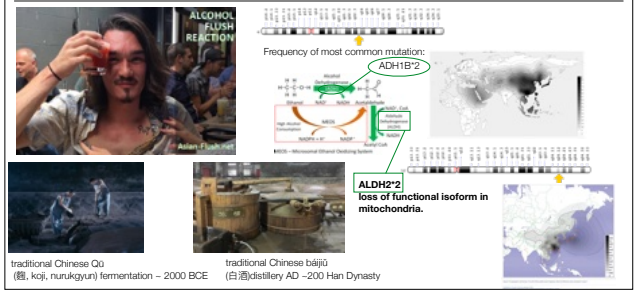
Vosoritide,
voxzogo?
from
Marin Bioscience

The most common form of dwarfism, achondroplasia is caused by mutations in the gene for FGFR3, fibroblast growth factor receptor 3. The most common one is a gain of function mutation with frequencies similar across different population the world over, most cases are due to ***de novo*** (new) mutations in one of the parents, usually the father. G1138A Mutation is found in over 90% of spontaneous achondroplasia cases.

Vosoritide (trade name Voxzogo), a drug manufactured by Biomarin, reduces the activity of FGF3R and was approved in 2021 in the USA and in Europe.

Alcohol flush, *ADH1B, ALDH2*

(genes for alcohol metabolizing enzymes, step 1 and step 2)

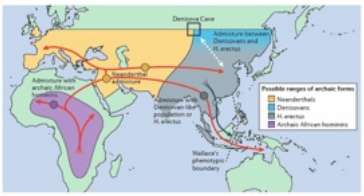


Alcohol is a natural toxin produced as a waste product when yeast ferment sugars. Humans have fermented sugars (diluted honey) or malted grain since more than 10 thousands of years to make alcohol.

genetic variation at two genes coding for alcohol metabolism alcohol dehydrogenase 1b that turns alcohol to acetaldehyde and aldehyde dehydrogenase 2 that turns acetaldehyde into acetate can strongly affect an individual's ability to metabolize alcohol and in doing so limit the toxic effect of alcohol.

These genes act in a **co-dominant manner**: one allele of a poorly active enzyme reduces metabolism, two copies lead to much stronger effect: individuals with two copies of inactive/or slowly active enzyme get classical facial flushing after just a small amount of alcohol. The reasons for the high frequency of these alleles in East Asia are not understood, but could include ***social selection*** against alcoholism, known to be very costly to societies. East Asia, where distillation was wide-spread early in history. East Asian populations have lower rates of

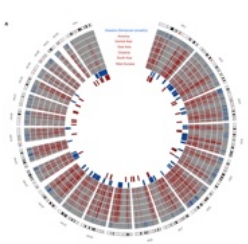
Hybridization happened



Veeramah & Hammer, ***Nature Reviews Genetics***, 2014

A possible model of archaic introgression based on the latest analysis using second-generation sequencing. Red arrows indicate initial colonization events across the Old World after the origination of anatomically modern humans (AMHs) in Africa, including two movements into Asia. Approximate positions of introgression events are represented by colored circles and are not intended to be accurate. This model portrays the hypothesis that portions of the Denisovan genome entered the human gene pool through hybridization with more widespread populations of archaic hominins (such as Homo erectus), which also interbred with the Denisovan population. The black arrow shows a more recent expansion of Asian farming populations (that is, <10,000 years ago) that did not carry introgressed Denisovan alleles and that replaced much of the indigenous resident population up to Wallace's phenotypic boundary (shown by the dashed line), which lies just east of Wallace's biogeographical line. This hypothesis may explain the lack of evidence for Denisovan

Mapping genes on the genome



Neanderthal girl paleo-reconstruction,
(Kennis and Kennis)
Non-overlapping 100 kb windows with inferred archaic ancestry in each of six populations (blue, Denisovans; red, Neanderthal).
In the innermost rings plots "gene deserts" (windows >10 Mb).

Sankararaman *et al.*, 2016, ***Current Biology*** 26, 1241–1247
The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans

Fine-Scale Maps of Denisovan and Neanderthal Introgression (**archaic introgression**)
(A) Non-overlapping 100 kb windows that have non-zero inferred archaic ancestry in each of six populations (blue, Denisova; red, Neanderthal). In the innermost rings, plot "**gene deserts**", ice, region relatively devoid of protein coding sequences (windows >10 Mb).

Map representing Neanderthal geographical distribution in groups.

Fabre, V.; Condemi, S.; Degioanni, A. (2009). "Genetic evidence of geographical groups among Neanderthals". PLOS ONE.

Ancient DNA from Neanderthal suggest that these lived in small populations and with some major subdivisions. They evolved independently from modern humans fro over 500,000 years.

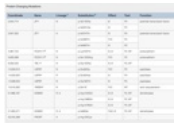
---

more evidence for New genome liabilities:
Y-chromosome, mt DNA?

Protein-coding differences between Neandertal and modern human Y chromosomes, including potentially damaging changes to PCDH11Y, TMSB4Y, USP9Y, and KDM5D. Three of these changes are missense mutations in genes that produce male-specific minor histocompatibility (H-Y) antigens. Antigens derived from KDM5D, for example, are thought to elicit a maternal immune response during gestation. It is possible that incompatibilities at one or more of these genes played a role in the reproductive isolation of the two groups.

Mendez *et al.* 2016 ***AJHG***

Many gene-rich areas contain zero admixture signal
Selection against incompatibility?



Mito-nuclear incompatibilities?

Sharbrough *et al.* 2017
Genome Biol. Evo

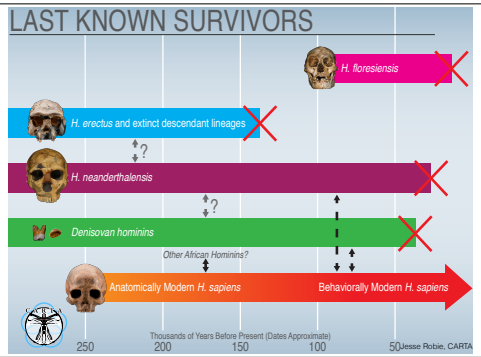
Both, the neanderthal Y-chromosome and the mitochondrial genome seem to have contained elements that gave rise to conflicts with the modern human genome.

Adaptive gifts from archaic hominids:

EPAS1 high altitude in Tibetans
BNC2 pigmentation in Europeans
POU2F3 keratinocyte proliferation in Europeans
HLA-C 15:05 allele disease resistance in Asia & Europe
TLR1/6/10 (Toll like receptors) innate immunity Asia & Europe
SLC16A11: Lipid metabolism (protection from starvation) Asia

Neanderthal liabilities - variants that are associated with negative outcomes:
Type 2 diabetes
Crohn's disease
Lupus
biliary cirrhosis
tobacco smoking
depression
Covid-19 severity
Lower athletic performance

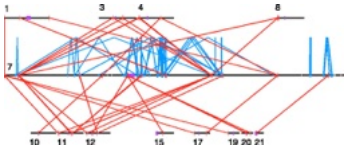
There are however, several genes that have been actively retained in the genomes of modern humans, some just in relatively few local populations: EPAS1 in Tibetans and Sherpas in the Himalayan plateau appears to be a Denisovan variant that is highly adaptive for high altitude. Several disease -resistance genes have also been co-opted after hybridization.
Genetic variants that protected against starvation are now genetic risks for type 2 diabetes, an example of genomic mismatch between archaic genes and modern human lifestyle….



The big picture remains one, where we Homo sapiens have taken over the entire planet and remain the last hominin standing. Our precise role if any in the demise of several other hominins, the Neanderthals, Denisovans and Homo floresiensis is hotly debated.

Segmental Duplications – Crucibles of Primate Evolution?

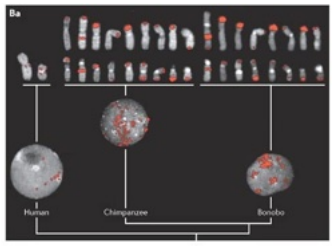
Recent (>40kb) segmental duplications on human chromosome 7



E. E. Eichler *et al.*, ***Science*** 301, 793 -797 (2003)

A schematic of recent segmental duplications (copy and pasting within or across chromosomes of large chunks of DNA) on human chromosome 7. The distribution of both interchromosomal (red) and intrachromosomal (blue) duplications is shown for human chromosome 7 (drawn ×50 to scale). Duplications (>90% sequence identity and >40 kb in length) correspond to duplications/gene conversion events that occurred over the last ~30 million years of human genome evolution.

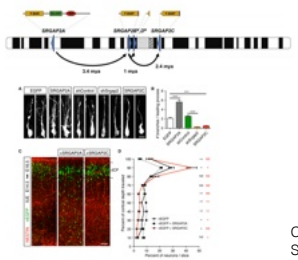
Segmental Duplication: differences in the
Panid lineage: hyperexpansion of a 40kb segment



Bailey & Eichler ***Nature Reviews Genetics*** 2006

By using DNA probes that are specific for segmental duplication in chimpanzees and bonobos one can stain the many places where these chunks of DNA landed after copying themselves… This is an example where the human genome was much less affected.

Evolution of Human-specific Neural SRGAP2 Genes
by Incomplete Segmental Duplication



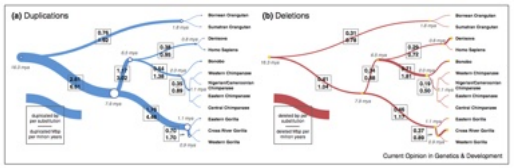
Dennis et al. CELL 2012,
Eichler Group UW

Charrier et al. CELL 2012,Polleux,
Scripps

Top: SLIT-ROBO Rho GTPase-activating protein 2 (srGAP2) also known as formin-binding protein 2 (FNBP2) is a protein that in humans is encoded by the SRGAP2 Schematic depicts location and orientation (blue triangles) of SRGAP2 paralogs on human chromosome 1 with putative protein products indicated above each based on cDNA sequencing. Asterisks indicate a 49 amino acid truncation of the F-BAR domain. Note that the orientation ofSRGAP2D remains uncertain, as the contig containing this paralog has not yet been anchored. Arrows trace the evolutionary history of SRGAP2duplication events. Copy number polymorphism and expression analyses suggest both paralogs at 1q21.1 (SRGAP2B and SRGAP2D) are pseudogenes, whereas the 1q32.1 (SRGAP2A) and 1p12 (SRGAP2C) paralogs are likely to encode functional proteins.Bottom. SRGAP2C Expression in Radially Migrating Mouse Cortical Neurons Phenocopies Srgap2 Knockdown

(A) Confocal images of optically isolated neurons showing representative morphologies of radially migrating cortical neurons in E18.5 embryos following in uteroelectroporation (IUE) at E14.5 of the indicated constructs. sh, short hairpin. Scale bar, 10 mm. (B) Mean number of branches (±SEM) of the leading process of neurons as represented in (A). n = 3 animals/condition, 100–150 neurons/condition. (C) Low magnification confocal images of E18.5 cortical slices showing migration of in utero electroporated neurons expressing nuclear-EGFP (nEGFP) alone or together with SRGAP2A or SRGAP2C. Staining with anti-GFP shows the position of the electroporated neurons, and anti-NESTIN marks the radial glial scaffold.dCP, dense Cortical Plate. (D) Quantification of neuron distribution in cortical slices as illustrated in (C) (mean ±SEM). n = 3 animals/condition, 9–10 slices/condition. In (B) and (D), *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; NS (not significant, $p > 0.05$); Mann-Whitney test. See also Figures S2 and S3.

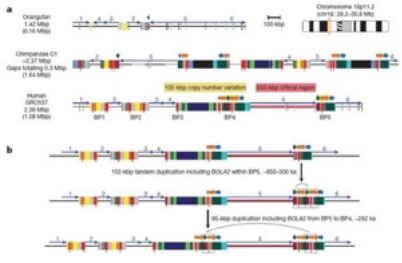
Primate rates of duplication and deletion



Dennis and Eichler 2016 ***Curr. Opinion in genetics and Development***

Primate rates of duplication and deletion. Rates of fixed (a) duplications and (b) deletions are shown as a function of the number of substitutions along each branch of the great ape phylogeny. Branch widths are scaled proportionally to the number of duplicated base pairs per substituted base pair based on analysis of 97 human/ape genomes. A burst of duplicated base pairs appears to have occurred in the common ancestral branch leading to humans and African great apes, where duplicated base pairs were added at 2.6-fold the rate of substitution. In contrast, the rate of deletion in the great ape lineage is more clocklike along all branches (mean of 0.32 deleted base pairs per substitution) with the exception of the chimpanzee–bonobo ancestral lineage, where an approximate twofold increase in the rate of deletion is observed (0.71 deleted base pairs per substitution).

Complex set of segmental Duplications on Chromosome 16

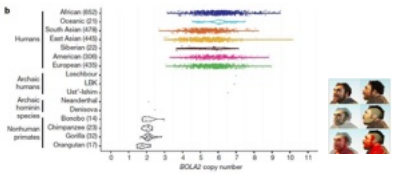


Nuttle *et al.* ***Nature*** 2016

a) Schematic depicts the genomic organization of chromosome 16p11.2 for one orangutan and two chimpanzee haplotypes along with the human reference haplotype (GRCh37 chr16:28195661–30573128; see ideogram for approximate chromosomal location). Blocks of segmental duplications within this locus mediate recurrent rearrangements in humans; thus, these blocks have been defined as breakpoint regions BP1–BP5). The ~550 kbp critical region (pink) and a >1 Mbp chimpanzee-specific inversion polymorphism (orange) are highlighted. Tiling paths of sequenced clones are indicated above each haplotype, with chimpanzee clones that could not be fully resolved marked with asterisks. Colored boxes and thick arrows indicate the extent and orientation of segmental duplications (with different colors denoting duplicons from different ancestral genomic loci, and hashed boxes indicating sequence duplicated in humans but not in the species represented). Thin numbered arrows show orientations of gene-rich regions of unique sequence. Numbers (left) indicate the size of each orthologous haplotype, with the number of segmentally duplicated base pairs shown in parentheses. Note that, for chimpanzee, these sizes are lower bounds due to gaps in the contigs (dotted line sections) and the contigs not reaching unique sequence beyond BP1 (i.e., unique region 1). b) Schematic depicts distinct human structural haplotypes over the chromosome 16p11.2 critical region and flanking sequences (three complete haplotypes extending from unique sequence distal to BP3 to unique sequence proximal to BP5 and one partial haplotype including BP3–BP4 and BP5 sequence contigs). High-quality sequence for each haplotype was generated by sequencing a total of 40 BACs and 15 fosmids from three different human genomic libraries. Regions of copy number variation (highlighted in yellow along the first two haplotypes) occur on both sides of the critical region and involve the same 102 kbp unit in direct orientation, including a 30 kbp block containing BOLA2 and two other genes and a 72 kbp block harboring a partial segmental duplication of SMG1 (SMG1P). Expansion and contraction of this cassette underlie hundreds of kbp of structural diversity between human haplotypes. BOLA2 paralog-specific copy number genotype data suggest that H1 and H3 likely represent the most common haplotype structures in humans.

Complex set of segmental Duplications on Chromosome 16

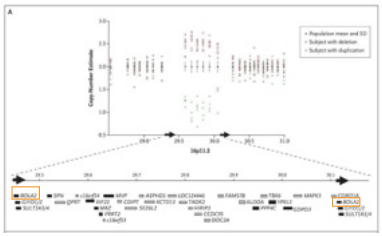
281 kya this event simultaneously increased copy number of gene associated with iron homeostasis and predisposed our species to recurrent rearrangement associated with disease.



Nuttle *et al.* ***Nature*** 2016

Diploid copy number estimates (points) for BOLA2 based on sequence read depth12 are shown for 2,359 humans, three archaic humans13,14, a Neanderthal2, a Denisovan3, and 86 nonhuman primates, with violin plots overlaid. c) Paralog-specific BOLA2 copy number genotypes (points, jittered around their integer values) were inferred from Whole Genome Sequencing read depth over informative markers for 222 individuals sequenced to high coverage. Colors correspond to different populations as in panel b.

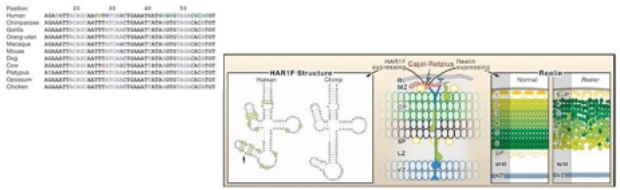
Association between Microdeletion and Microduplication at 16p11.2 and Autism



Weiss LA et al. ***N Engl J Med*** 2008;358:667-675.

Deviation from BOLA2 copy number are associated with psychiatric disease. This is one of the strongest evidence to date, that Neanderthal DNA could have been a liability to neurotypical development in modern humans.

HAR 1 human accelerated region 1



18/118 human specific changes to an RNA coding region

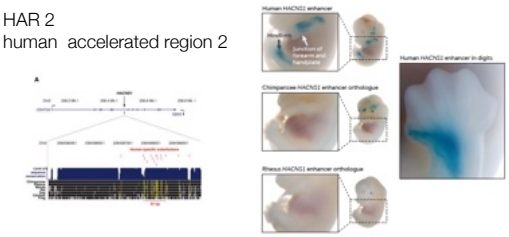
expression during fetal brain development in cortex

Pollard et al. 2006 ***NATURE***
Haeussler Group UCSC

Speed demon sequences: Human accelerated regions HARs:
These DNA sequences are virtually identical from chicken to chimpanzee, but then show many changes in humans
HAR1 is expressed doing embryonal brain development and appear to affect the formation ion the six cortical layers in the brain.
It does not encode a protein but rather two different small functional RNAs, one in each direction of reading….

Human Accelerated Non-coding regions
human sequence expressed in mouse embryos dramatic affect on limb development

HAR 2
human accelerated region 2

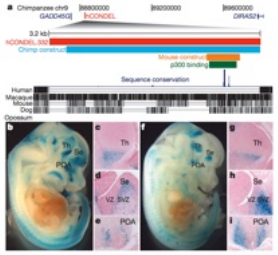


Prabhakar S et al ***Science***, 2009, Noonan Group, Yale

Expression patterns obtained from HACNS1 and its chimpanzee ortholog in E13.5 mouse embryos. Three embryos resulting from independent transgene integration events are shown for each construct. Close-up views of forelimb and hindlimb expression in a representative embryo for each construct are shown at left, and arrows indicate positions where limb expression is present or absent. B. Dorsal view of reporter gene expression in the distal anterior forelimb of a HACNS1 E13.5 transgenic embryo. Arrows indicate the most anterior digit.

Enhancer deletion of tumor suppressor gene GADD45G

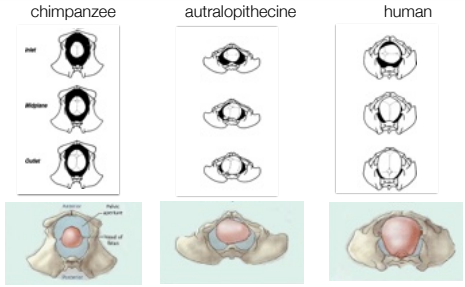
hCONDEL in enhancer of tumor suppressor gene Growth arrest and DNA-damage-inducible, gamma (GADD45G) results in prolonged neuronal cell division which may contribute to thalamic and cortical expansion in humans.



McLean et al. ***Nature***, 2011. Kingsley Group, Stanford

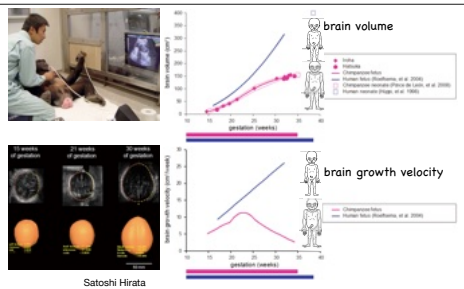
Transgenic analysis of a chimpanzee and mouse forebrain enhancer missing from a tumour suppressor gene in humans. a, Top panel: 1.3Mb region of the chimpanzee chromosome 9. The red bar illustrates a 3,181 bp human-specific deletion removing a conserved chimpanzee enhancer located downstream ofGADD45G. Bottom panel: multiple species comparisonof the deleted region, showing sequences aligned between chimpanzee and other mammals. The green bar represents a mouse forebrain-specific p300
binding site18, and the blue and orange bars represent chimpanzee and mouse sequences tested for enhancer activity in transgenic mice. The chimpanzee (b–e) and mouse sequence (f–i) both drive consistent lacZ expression in E14.5 mouse embryos in the ventral thalamus (c,g), the SVZ of the septum (d,h), and the preoptic area (e, i). Increased production of neuronal subtypes from these regions may contribute to thalamic and cortical expansion in humans27–30. All sections are sagittal with anterior to right. POA, preoptic area; Se, septum; SVZ, subventricular zone; Th, thalamus; VZ, ventricular zone.

Obstetrics: birthing a big-brained baby



Humans have hit the limit: bipedality imposes an upper limit to baby head size.
Human birth has become very risky, as human babies have to rotate their heads in order to clear the pelvic passage of their mothers.
Regular C-sections are now lifting this limit, prediction: in population where c0section rates are very high, there will soon be many more babies with heads that cannot possibly clear the pelvic opening of their mothers (cephalo-pelvic disproportion)

How to grow a big brain

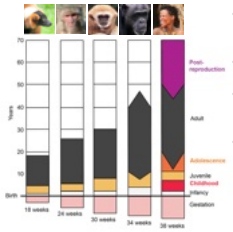


Satoshi Hirata

Sakai et al. **Current Biology**, 2012

Gestational age-related changes in brain volume in chimpanzee (Hatsuka and Iroha) and human fetuses.
Gestational age-related changes in the growth velocity of brain volume in chimpanzee and human fetuses
Chimpanzee brains start slowing down their growth in mid-pregnancy, humans on the other hand continue a high fetal rate for a full year after birth.

Childhood, Adolescence & Post-reproductive Survival



- Adapted to cultural opportunities?
- Nutritional opportunities?
- Facilitated by stronger pair bonds between parents?
- Facilitated by allomothering?

redrawn from Schultz 1963
Boaz & Almquist 1999

Odd life history compared to other primates
Humans have delayed development: humans invented childhood (slow body but rapid brain growth), adolescent growth spurt, and prolonged post-reproductive survival. But evolved shorter inter-birth intervals than apes!

Derived Human Growth Schedule

- Delay allows increased transmission of behavior and concepts.
- Human minds are effective copying devices and idea generators.
- Language is one of the major target of imitation and idea transmission.
- Delayed development: biological assimilation of culture?
- Paradoxically shorter Inter-birth-Interval than apes.

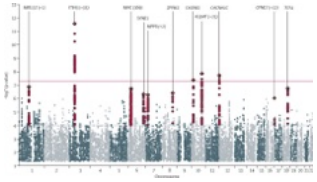
Minds as copying machines and idea generators

Humans over-imitate, focusing as much on the way than on the goal, chimps go for the goal.

Ratcheting culture: build on old ideas with new ideas….

Schizophrenia, a multi-gene disorder

genes contributing risk *ABCA13, AKT1, C4A, COMT, DGCR2, DGCR8, DRD2, MIR137, NOS1AP, NRXN1, OLIG2, RTN4R, SYN2, TOP3B, YWHAE, ZDHHC8*, chromosome 22, & many more!



Cross-Disorder Group of the Psychiatric Genomics Consortium, Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis ***The Lancet*** 2013

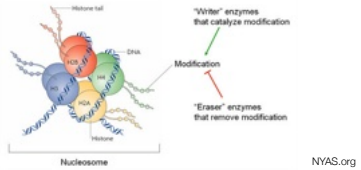
Schizophrenia, is a human-specific condition that appears to have many genetic contributions as well as environmental determinants.

Genome wide association studies (**GWAS**) have identified dozens of loci, most of them with small effects on the risk of schizophrenia.

Identical twins are only 50% congruent with regard to developing schizophrenia, even though they share an identical genome.

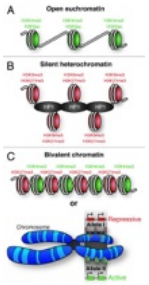
Genetic causes include **non-inherited factors**, such as reactivation of **endogenous retroviruses** (viruses that have inserted themselves in our genomes)!

"Histone Code" & Epigenetics



Wrapping of DNA around histones and modification of DNA and histones affect gene expression (methylation of DNA, acetylation, phosphorylation and methylation of histones)

“Histone Code” & Epigenetics



Histone deacetylase in royal jelly

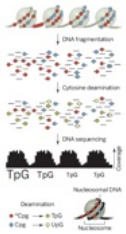
Spannhof et al. EMBO Reports 2011

DNA Methyltransferase inhibitors

Alhosin, M. Epigenetics Insights 2023

One of the best illustrations of an epigenetic effect: roll jelly made by glands in the worker bees can turn any fertilized egg from a worker bee into a queen. The secretions contains a deacetylase enzyme that can modify the histone code (chemical modifications on the histone tails, that can prevent the activity of the DNA wrapped around these…, Royal jelly made by worker bees also contains inhibitors of DSNA methyltrasnferases, which silence certain DNA stretches on eggs of worker bees by methylating the cytosines.

Ancient Epigenetics



DNA damage in ancient genomes may provide insight into past regulatory changes in humans and other species

Ludovic and Willerslev, ***Science*** 2014

There are the first few studies on ancient epigenetics…

“Apportionment” Humans Compared to Chimpanzees

Chimpanzees
Classically divided into 4 subspecies
Recent genomic evidence suggests they are clinal and continuous
2.4 times more genetically diverse than humans

Humans
Less diverse (despite population size & global reach)
Recent African origin
Radiation of small founding populations

Two Contrasting Truths

Human populations are remarkably genetically similar

There also exists real genetic difference between them

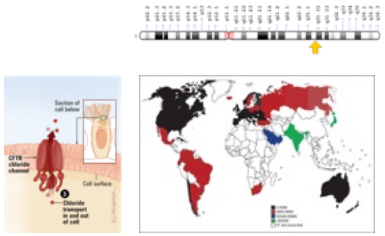
Timbuktu to Tokyo

Where does one "race" stop and another start??

Cystic Fibrosis (CF), *CFTR* gene

(gene for chloride channel)

Frequency of most common mutation: The most common mutation, called delta F508, is a deletion of one amino acid at position 508 in the CFTR protein. The resulting abnormal channel breaks down shortly after it is made, so it never reaches the cell membrane to transport chloride ions. It is present in 2/3 of CF patients.



Mirtajani SB, *et al.* Geographical distribution of cystic fibrosis; The past 70 years of data analyzis. ***Biomed Biotechnol Res J*** 2017;1:105-12

Loss or partial loss-of-function mutation in the **CFTR gene are autosomal recessive causes of disease**: Cystic fibrosis or CF.

In individuals who inherit two mutated copies, challenged chloride ion transport can cause a baby's sweat to taste salty.

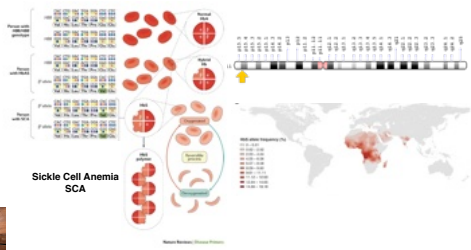
The disorder affects the lungs, pancreas, liver, kidneys, and intestine, it also causes infertility in males.

The disease occurs in 1 in 2,500 to 3,500 European American newborns. Cystic fibrosis is less common in other ethnic groups, affecting about 1 in 17,000 African Americans and 1 in 31,000 Asian Americans.

Humans carry an average of one to two recessive lethal alleles! Recessive disease genes are much more common than dominant genes.

Sickle Cell Anemia, *HBB* gene

(gene for hemoglobin B subunit, structural and gas transport protein, **autosomal recessive**)



Piel, F. *et al.* Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. ***Nat Commun*** 1, 104 (2010).

Sickle cell disease affects millions of people worldwide.

The disease is caused by inheritance of two copies (autosomal recessive) of mutated hemoglobin beta subunit genes.

It is most common among people whose ancestors come from Africa; Mediterranean countries such as Greece, Turkey, and Italy; the Arabian Peninsula; India; and Spanish-speaking regions in South America, Central America, and parts of the Caribbean.

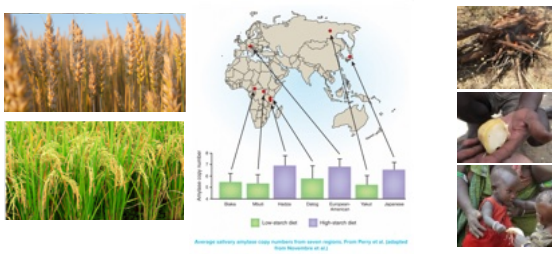
The mutation appears to be selected by the relative resistance to malaria in carriers of just a single mutation (heterozygote carrier).

One copy of a inactivating mutation protects from malaria, but two copies cause

disease!
Sickle cell disease is the most common inherited blood disorder in the United States, affecting 70,000 to 80,000 Americans.
The disease is estimated to occur in 1 in 500 African Americans and 1 in 1,000 to 1,400 Hispanic Americans.
Sickle cell anemia is a **pleiotropic** disease because the expression of a single mutated HBB gene produces numerous consequences throughout the body. The mutated hemoglobin forms polymers and clumps together causing the deoxygenated sickle red blood cells to assume the disfigured sickle shape. As a result, the cells are inflexible and cannot easily flow through blood vessels,

---

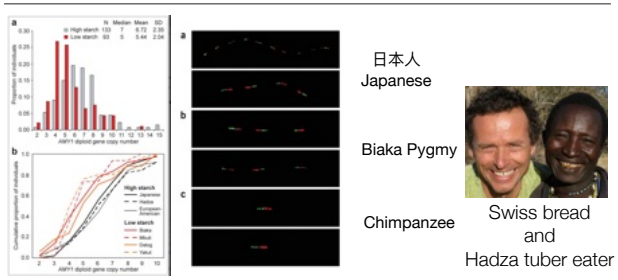
Amylase CNV in humans and apes



Perry et al. ***Nature Genetics*** 2011

Differences in salivary amylase gene copy number exist between present day humans populations. They correlate with the history of grain or tuber use.

Amylase CNV in humans and apes



日本人
Japanese

Biaka Pygmy

Chimpanzee

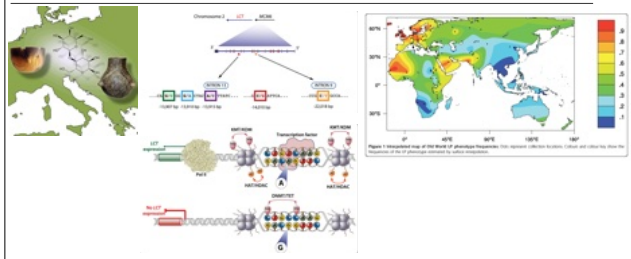
Swiss bread
and
Hadza tuber eater

Perry et al. ***Nature Genetics*** 2011

Different number of copies of the salivary amylase gene in the genomes of humans and chimpanzees. These copies can be fluorescently stained on prepared chromosomes.

Descendent from European farmers (Pascal) and foraging Hadza (my friend Mbugoshi), both have higher salivary amylase gene numbers —no correlation with “races”.

Domestication: Culture changing genes



Carlber, C. 2023. ***Redox Biology***

Itan et al. . 2010. ***Evolutionary Biology***

7.5 kya process began to select for lactase persistence, the only populations with many individuals who can digest lactose as adults have histories of stealing milk from other species….
The practice was so advantageous, that lactase persistence mutations were selected at least twice (in Europe and in East Africa). A cultural practice that has changed the human genome in less than 10,000 years.
This happened at least twice independently in Europe and in East Africa respectively.

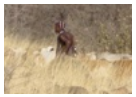
Molecular basis of lactase persistence. The genomic region of thegenes LCT and MCM6 is shown
(A). SNPs (single nucleotide polymorphism) located approximately 14 and 22 kb upstream of the TSS of the LCT gene, which are located within introns 13 and 9 of the MCM6 gene, respectively, are associated with lactase persistence. The function of regulatory SNPs is schematically depicted
(B). The SNP is part of a transcription factor binding site and provides in oneallele (top) high affinity and in the otherallele (bottom) low affinity for the transcription factors. In case of rs4988235 at position –13,910 relative to the LCT genethis is POU2F1 (POU class 2 homeobox1). Moreover, epigenetic effects, such as histone acetylation and methylation as well as DNA methylation can affect the expression of the LCT gene and mediate lactase persistence.

Why would mammals evolve to stop the production of the enzyme lactase around weaning age? Is it economy of investment or does it reflect parent-offspring conflict?

Animal milk use allows you to out-compete your neighbors!

Hadza foragers

Datoga pastoralists



Swiss pastoralist

Southern Italian

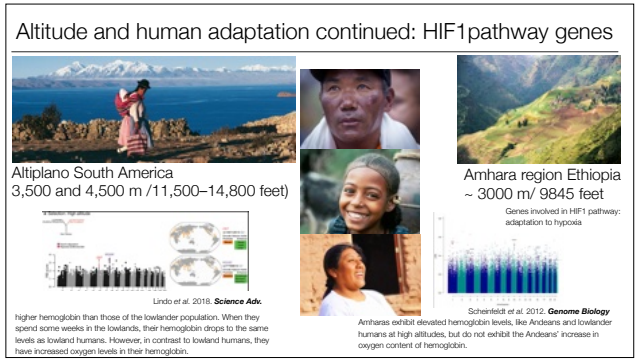
Swiss pastoralist

Chaga (Bantu) pastoralist

Hunter gatherers have no access of milk from other species, pastoralists thrive on milk from their domesticated animals. The latter can wean very early ~ one year or less and proceed to have many more babies.
Farming people and pastoralists simply “out-baby” their neighboring foraging societies!

Lactase persistence is not a “racial trait”! Swiss and Bantu pastoralist , my fiend Douglas Simbeye can both still digest lactose as adults! —no correlation with “races”.
Two Europeans, a Swiss pastoralist and a Southern Italian differ, my friend and former Aikido teacher Michele Quaranta is lactose intolerant -no correlation with

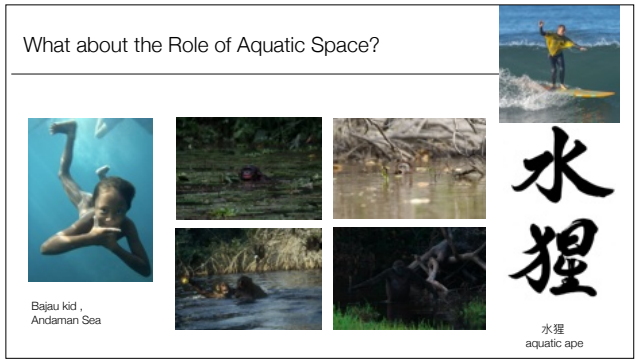
“races”.



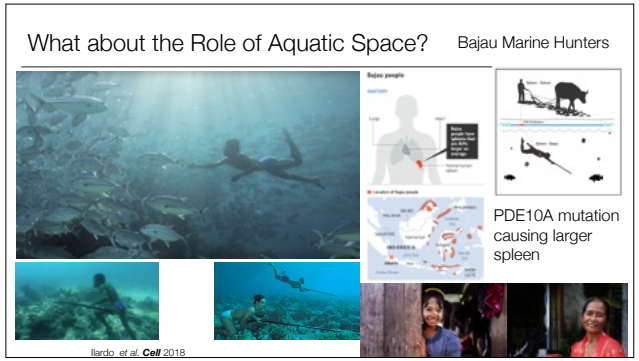
In contrast to the Tibetans, Andean highlanders show different patterns of hemoglobin adaptation. Their hemoglobin concentration is higher than those of the lowlander population, which also happens to lowlanders who move to high altitudes. When they spend some weeks in the lowlands, their hemoglobin drops to the same levels as lowland humans. However, in contrast to lowland humans, they have increased oxygen levels in their hemoglobin; that is, more oxygen per blood volume. This confers an ability to carry more oxygen in each red blood cell, meaning a more effective transport of oxygen throughout their bodies.
PBS scans testing for differentiation along the Ancient Andean branch, with respect to the lowland population of the Huilliche-Pehuenche. (B) PBS scan testing for differentiation along the modern Aymara branch (post-European contact), with respect to the precontact ancient population of the Rio Uncallane. Selected global SNP frequencies (Frq) (64) and histone modifications (38) are shown on the right, which are relevant to the strongest selection signals. “*” indicates closest gene. GI, gastrointestinal tract; NK, natural killer.

Highland Amharas exhibit elevated hemoglobin levels, like Andeans and lowlander humans at high altitudes, but do not exhibit the Andeans’ increase in oxygen content of hemoglobin.[49] Among healthy individuals, the average hemoglobin concentrations are 15.9 and 15.0 g/dl for males and females, respectively (which is lower than normal, similar to the Tibetans), and an average oxygen saturation of hemoglobin is 95.3% (which is higher than average, like the Andeans).

Again: three independent populations on three continents share similar high altitude adaptations - no correlation with “races”!

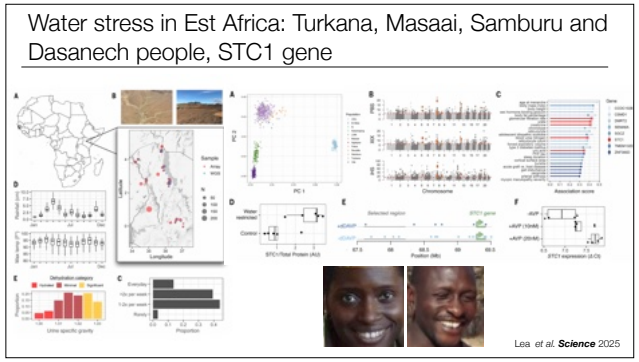


The aquatic ape hypothesis tried to explain a suite of human traits by an ancient aquatic phase.
From subcutaneous fat, loss of hair, to diving reflex, to swimming neonates, floating long hair, breasts in female, bipedality etc. Many of these features evolved millions of years apart and cannot simply be explained by an overarching hypothesis.



Genetic adaptation to diving in the Bajau. Private genetic mutations in the gene PDE10A in this population have been selected by conferring the ability to free dive for much longer.

Bajau and non-Baja Indonesian neighbor: —no correlation with “races”



Ecology and lifestyle of the Turkana people. (A) Map of Africa with an inset showing northwest Kenya (the present-day homelands of the Turkana people). Dots indicate where samples were collected
for this study, for both WGS and array genotyping. (B) Representative photographs of the arid ecology of the Turkana region (photos taken by the authors). (C) Proportion of Turkana pastoralists in this
study who consume meat at differentself- reported frequencies (N = 346). (D) Average rainfall and maximum temperature for the Turkana country region by month. Data were sourced from WorldClim
(56). (E) Proportion of Turkana pastoralists with different urine specific gravity values (N = 175), colored by whether each value meets the criteria for dehydration provided in (11).

Scans for selection and evidence for the STC1 gene. (A) Principal components (PC) analysis

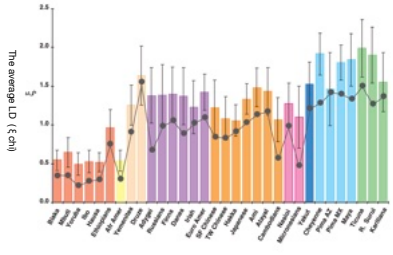
comparing WGS data from the Turkana as well as other study communities to African populations included in the 1000 Genomes Study (15). CEU, Northern Europeans from Utah; LWK, Luhya in Kenya; YRI, Yoruba in Nigeria. (B) Number of PBS, XtX, and iHS outliers per 50-kb window. Red dotted lines represent the cutoff for the 99th percentile of each empirical distribution for each selection statistic. Windows that exceeded the 99th percentile for all three statistics are highlighted in red. The candidate region that falls near the STC1 gene is highlighted with a red arrow. (C) Phenotypes associated with genes that fall in or near candidate regions. Phenotypic associations were sourced from the Open Targets Platform (57), and the association score (y axis)
represents the aggregate evidence across all published studies for a gene-phenotype link. Only association scores >0.25 are plotted, and association scores involving the STC1 gene are highlighted with an asterisk. (D) Normalized STC1 protein concentrations in mouse kidneys after 24 hours of water restriction. (E) Open chromatin regions in mouse collecting duct cells

---

The average LD for 83 SNPs across 21 haplotypes for 32 populations

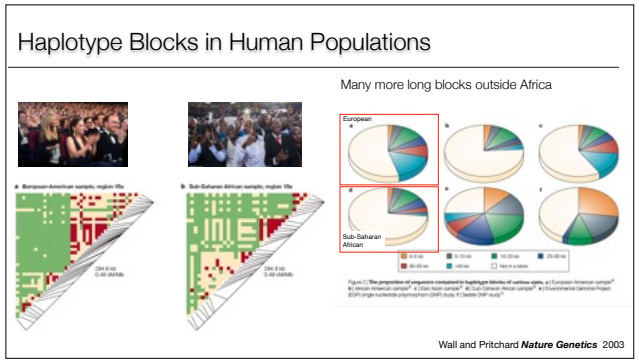
based on published data on CD4, DM1, DRD2, DRD4, PAH and COMT plus unpublished data.

Increase in LD with distance from Africa provides one of the strongest evidence of our common African ancestry.



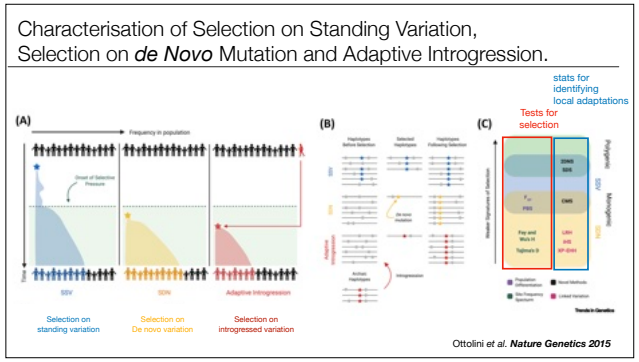
Tishkoff & Kidd VOLUME 36 | NUMBER 11 | NOVEMBER 2004 NATURE GENETICS SUPPLEMENT

The average LD (ξ chi) for 83 SNPs across 21 haplotypes for 32 populations. LD is measured as the ξ coefficient, a **standardized measure of overall nonrandomness of alleles at the sites in the haplotypes**. The bars are the mean values of ξ across the same 21 independent haplotype systems in all populations. The standard errors of the means are given as the error bars and the median values are plotted as dots connected by the line. Bars are color-coded by geographic region of origin of the populations, from left to right as sub- Saharan Africa, African Americans, Southwest Asia, Europe, East Asia, Pacific, Siberia, North America and South America. Population and sample descriptions are in ALFRED69.
Different samples of populations with the same name are distinguished by initials: SF, San Francisco; TW, Taiwan; AZ, Arizona; MX, Mexico; R, Rondonian. The haplotyped loci were chosen with no prior knowledge of LD values at the locus. The number of sites per haplotyped locus varied from 2 to 7 for a total of 83 SNPs. The graph is based on published data on CD4, DM1, DRD2, DRD4, PAH and COMT plus unpublished data. These results show less LD in African populations than elsewhere and greater LD in the Native American populations than in other regions, as well as variation in LD within geographic regions.



A range of methods have been proposed for defining haplotype blocks. Broadly speaking, they can be classified into two main groups:
1. those that define blocks as regions with limited haplotype diversity
2.those that make use of pairwise disequilibrium (for example, based on |D′|) to identify transition zones in which there is evidence for extensive historical recombination.

Pairwise |D′| plots for representative regions from different studies. Each square in the triangle plots the level of
linkage disequilibrium (LD) between a pair of sites in a region; comparisons between neighbouring sites lie along the diagonal. Red
colouring indicates strong LD, green indicates weak LD and light brown indicates intermediate or uninformative LD (see BOX 2 and
REF. 32 for details). The long diagonal line indicates the physical length of the region, and the short black lines plot the position of
each marker in this region. We include the physical length and estimated recombination rate53 for each region. EGP, Environmental
Genome Project; SNP, single nucleotide polymorphism.



Characterisation of **Selection on Standing Variation, Selection on de Novo Mutation** and **Adaptive Introgression**.
(A) Cartoon showing the rise in frequency of an allele through a population according to its origin. The stars represent mutations (blue: mutation present in the population prior to selection; yellow: mutation occurring after the onset of selection; and red: mutation present in an archaic population which spreads through a receiving population following an **admixture event**). The frequency of each mutation in a population following a selection event is represented by the number of people icons of their respective colours under each panel. The red walking person icon in the top right represents an archaic human.
(B) Cartoon depicting the haplotypes arising from selection on **standing variation** (SSV), selection on **de novo mutation** (SDN), and **adaptive introgression**. Stars represent the

beneficial allele and grey circles represent neutral linked polymorphic alleles. (C) Some of the most common statistics used to identify local adaptation (not a comprehensive list) Abbreviations: iHS, integrated haplotype score; LRH, long range haplotype; PBS, population branch statistic; SDS, Singleton density score; XP-EHH, cross-population extended haplotype homozygosity.

Summary

- Individual genomes are unique genetic mosaics.
- Each piece of DNA has its distinct history.
- Complete genome sequences for many apes allow:
  - Reconstruction of past population histories
  - Finding changes that define the human species.
- Humans are enriched for changes affecting brain development, incl. genes involved in uniquely human diseases.
- Fossil DNA data, experiments in model animals and cell culture allow testing for biological effects.
- Experimental approaches to neurodevelopment are severely limited for ethical reasons.