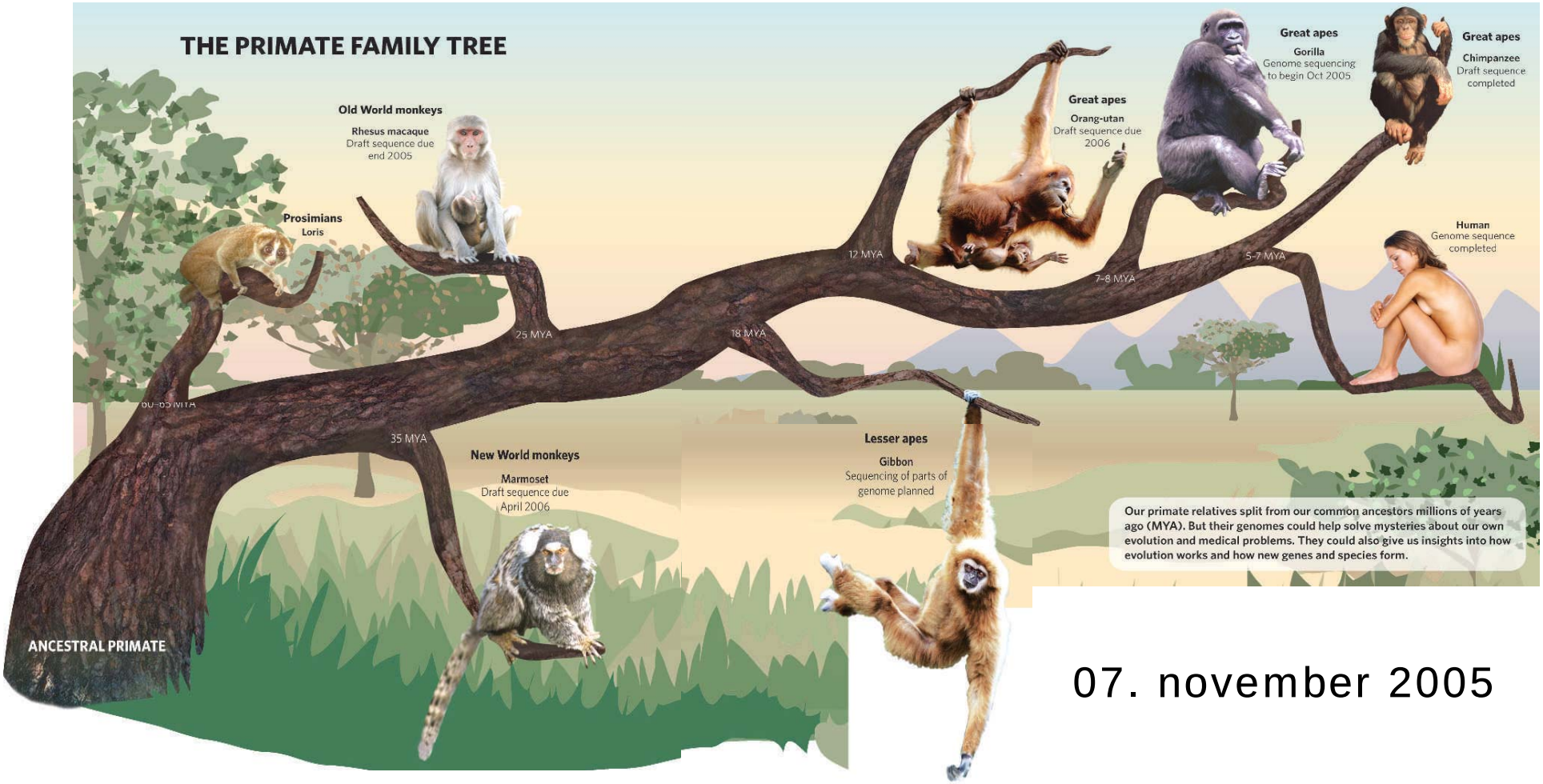
Initial sequence of the chimpanzee genome and comparison with the human genome



07. november 2005

Why chimp?

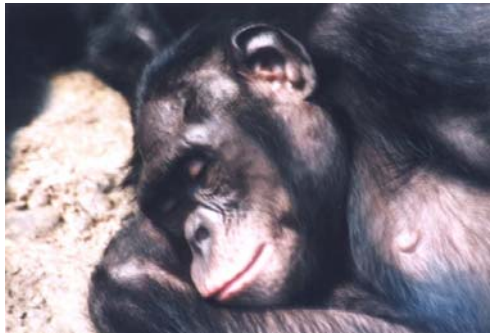
This is not just another genome, where intresting novel genes and pathways are looked for.

It’s the closest living species to human.

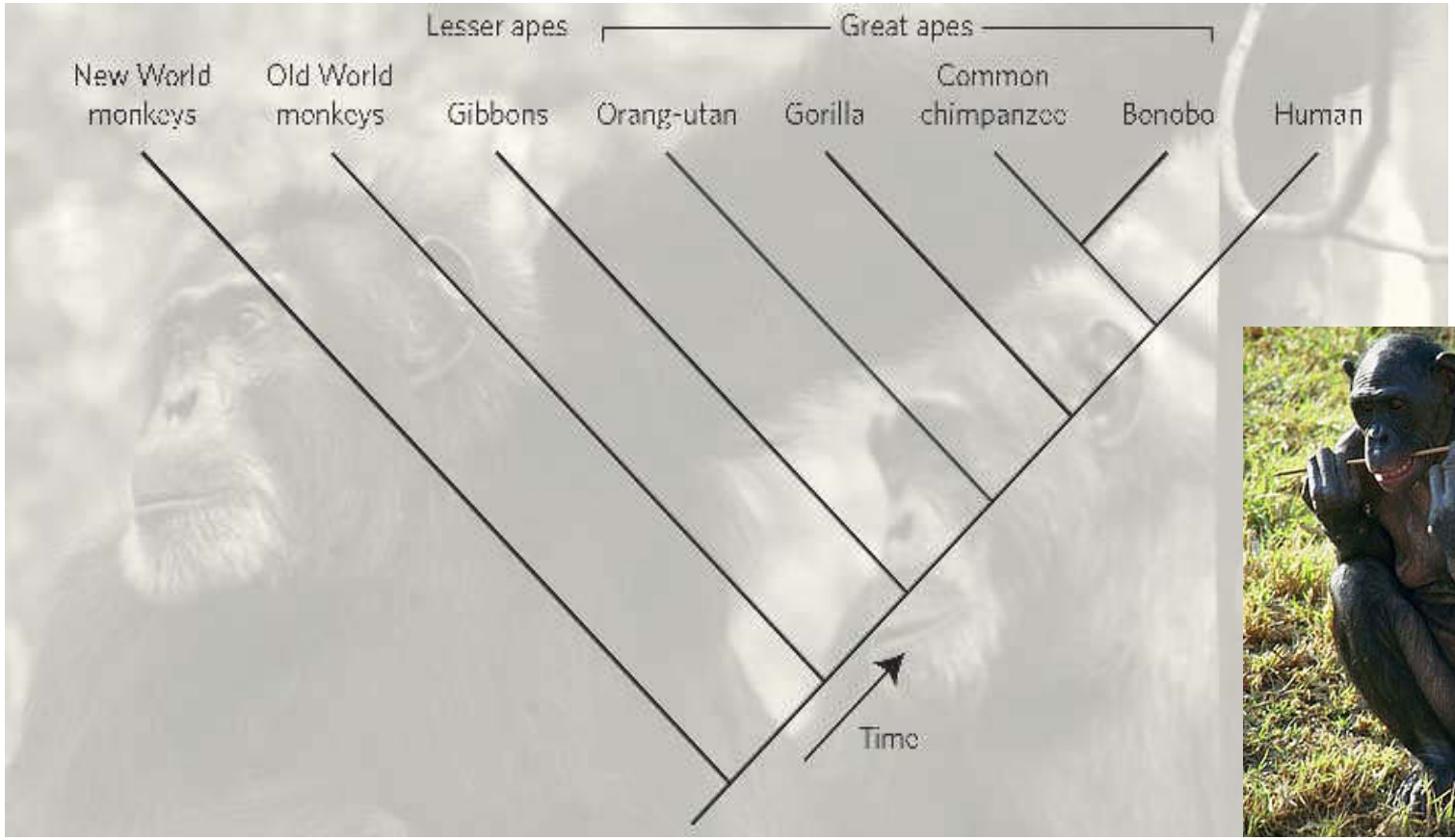
The data provide a treasury of information for understanding human biology and evolution.

What genetic changes make us so different from the chimpanzee, our closest relative?

The chimpanzee *(Pan troglodytes)* has a sister species – pygmy chimpanzee or bonobo *(Pan paniscus)*



bonobo



bonobo

Problems in comparison

- **It's difficult to tell whether a DNA sequence in humans that is missing in chimps was really added during human evolution or has simply been lost in the chimp lineage.**

- **Another problem is that it is hard to be sure straight away that any differences found are significant. Chimps, like humans, differ genetically from each other, although the extent is debatable. More chimps from different subspecies must be sequenced to capture the full extent of sequence diversity.**

- **The chimp genome sequence is still only a draft. To ensure that the differences found are real, the chimp sequence needs to be improved to match the polished 'finished' standard of the human genome.**

More primate genomes

- **Researchers need other primate genomes if they are to address the question of which genetic changes are unique to humans or chimps.**

- **The rhesus macaque, an Old World monkey, will be the first available — a preliminary assembly of its genome sequence was released into the public databases earlier this year and an improved version is expected by the end of the year.**

- **The push to sequence its genome stems from its popularity in biomedical research.**

How was the genome sequenced?

- **The genome of a single male chimpanzee (Clint), a captive-born descendant of chimpanzees from the West Africa was sequenced.**

- **The data were assembled using both the PCAP and ARACHNE programs. The former was a de novo assembly, whereas the latter made limited use of human genome sequence (NCBI build 34) to facilitate and confirm contig linking.**
  The ARACHNE assembly has slightly greater continuity and was used for analysis in this paper.

- **The first draft was assembled Nov. 2003**

Quality of the sequence

- **The draft genome assembly—generated from 3.6-fold sequence redundancy of the autosomes and 1.8-fold redundancy of both sex chromosomes—covers 94% of the chimpanzee genome with 98% of the sequence in high-quality bases.**

- **A total of 50% of the sequence (N50) is contained in contigs of length greater than 15.7 kilobases (kb) and supercontigs of length greater than 8.6 megabases (Mb).**

- **The assembly represents a consensus of two haplotypes, with one allele from each heterozygous position arbitrarily represented in the sequence.**

What was compared between chimp and human genome?

- **SNPs in chimp genome**
- **Number of single nucleotide differences between genomes**
- **Number of modest and large insertions/deletions between genomes**
- **Number of novel repetitive elements in both genomes**
- **Number of novel pseudogenes in both genomes**
- **Nucleotide substitution rate in different gene regions**
- **Nucleotide substitution rate in different genes**

Other individuals were sequenced:

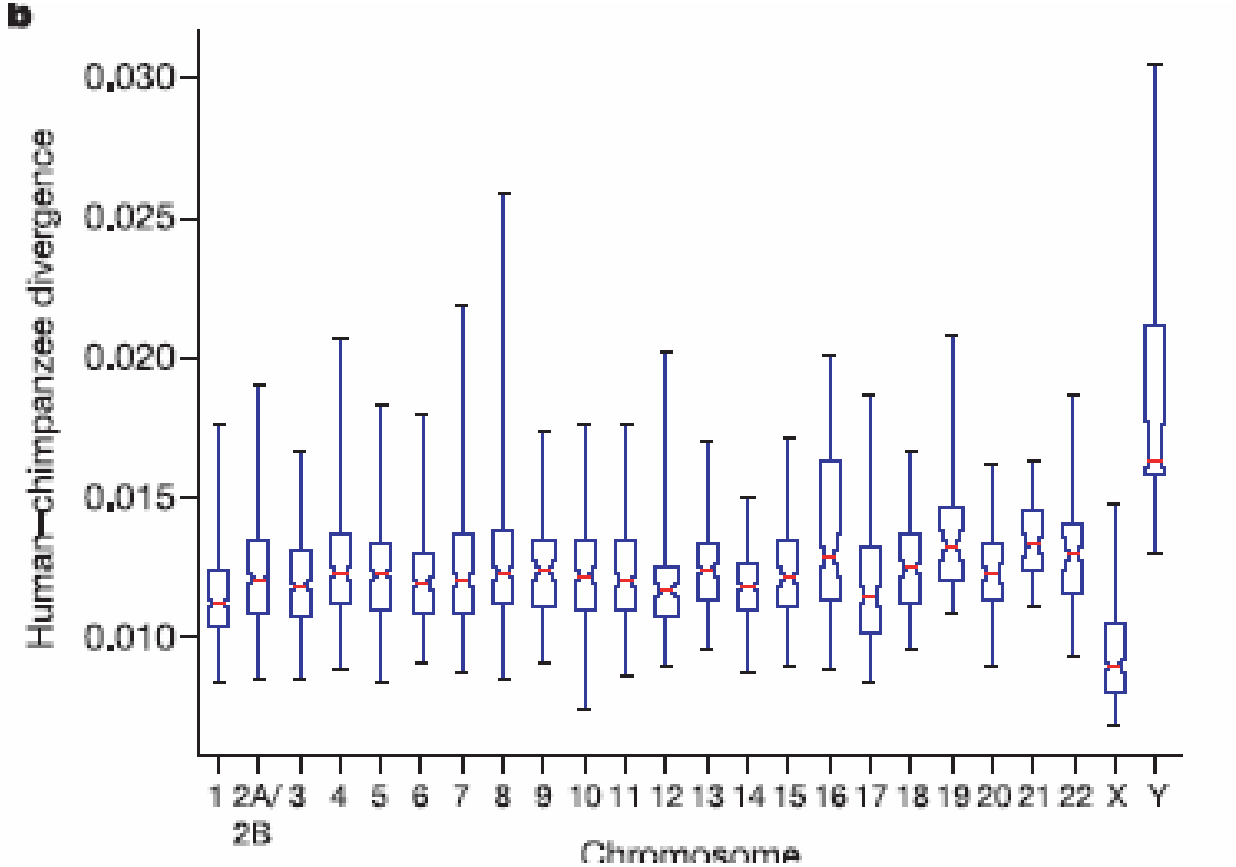
Table S17 Number of Passing SNP Reads from Donor Chimpanzees

Lineage	Designation	Reads
Western	Clint	23,021,928
Western	Donald	38,633
Western	Gon	77,239
Western	Yvonne	451,782
Western	Karlien	464,633
Central	Clara	302,768
Central	Masuku	372,185
Central	Noemie	502,682
Total		25,231,850

A total of 1.66 million high-quality single-nucleotide polymorphisms (SNPs) were identified, of which 1.01 million are heterozygous within the primary donor, Clint.

None can be found in ENSEMBL or UCSC genome browsers
Downloadable from dbSNP from June 2005

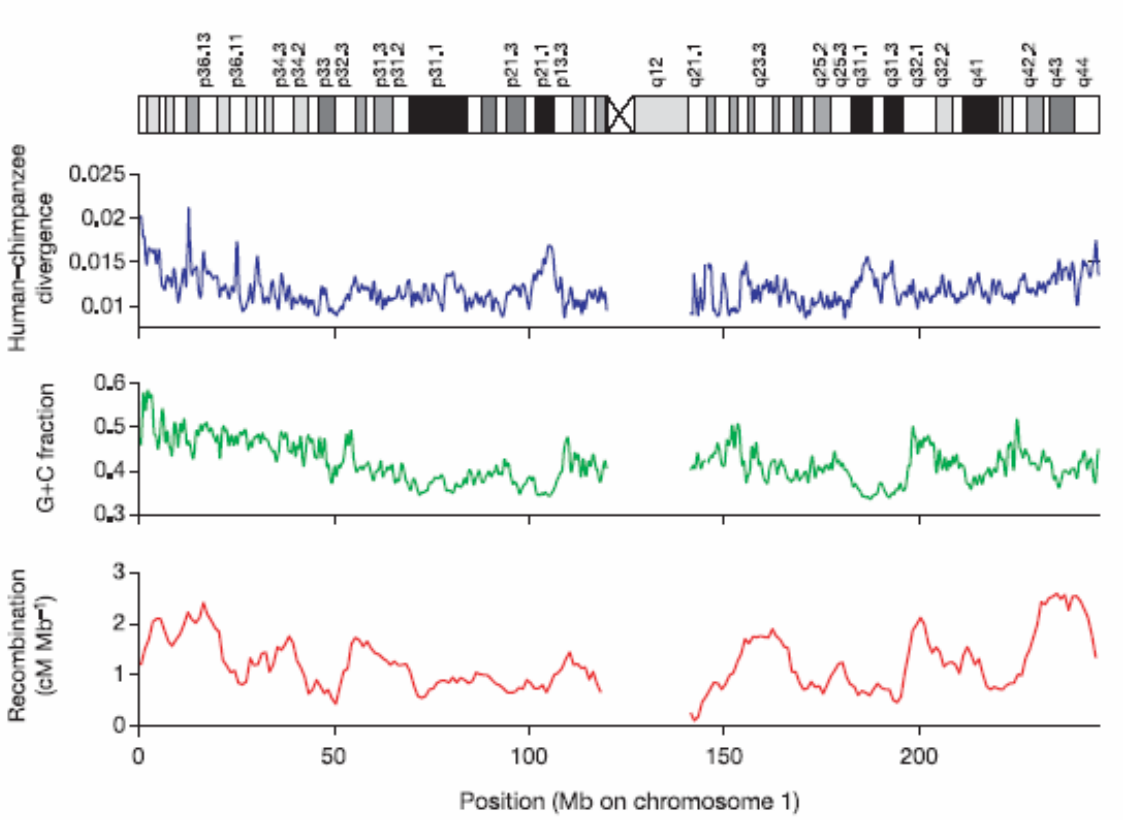
Human-chimp divergence



1.23% difference in aligned DNA sequence (2.4 Gb)

Distribution of variation by chromosome, shown as a box plot. The edges of the box correspond to quartiles; the notches to the standard error of the median; and the vertical bars to the range. The X and Y chromosomes are clear outliers, but there is also high local variation within each of the autosomes.

Divergence is not randomly distributed



Regional variation in divergence rates. Human–chimpanzee divergence (blue), GC content (green) and human recombination rates (red) in sliding 1-Mb windows for human and chimpanzee chromosome 1.
Divergence and GC content are noticeably elevated near the 1p telomere, a trend that holds for most subtelomeric regions (see text). Internally on the chromosome, regions of low GC content and high divergence often correspond to the dark G bands

Insertions/deletions

Reliable estimation for insertions between 1-15 000 bp

Two classes
- insertion in chimp (which in fact could also be deletion in human)
- insertion in human (which in fact could also be deletion in chimp)

35 Mbp insertions in chimp
32 Mbp insertions in human

163 large insertions in human genome, from these 34 contain exons
Together with estimated larger insertions 90Mbp DNA divergence

90 Mbp difference with 5 million events (3% of the genome)
35 Mbp difference with 35 M events (1.23% of the genome)

Insertions/deletions

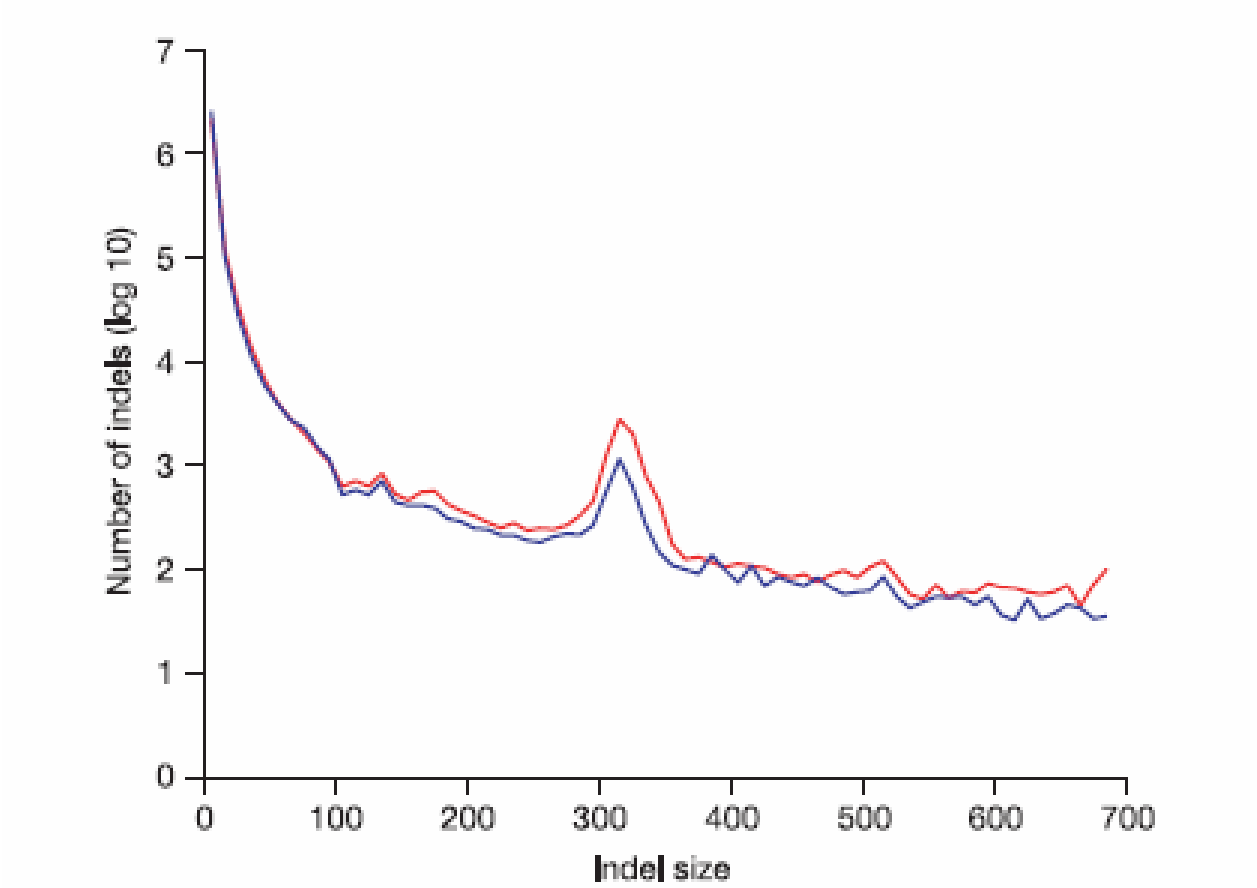


Figure 5 | Length distribution of small indel events, as determined using bounded sequence gaps. Sequences present in chimpanzee but not in human (blue) or present in human but not in chimpanzee (red) are shown. The prominent spike around 300 nucleotides corresponds to SINE insertion events. Most of the indels are smaller than 20 bp, but larger indels account for the bulk of lineage-specific sequence in the two genomes.

Transposable elements

Endogeneous retroviruses
Line1 repeat elements

Alu repeat elements
Pseudogenes

Transposable elements

Endogeneous retroviruses

- 1 active in both lineages (HERV-K)
  7 human + 1 chimp full-length copy
- 2 chimp-specific ERVs, older PtERV1 (ca 200 copies) and novel PtERV2 (ca 20 copies)

- several others have died out in human lineage (only LTR repeats remaining)

Line1 repeat elements

Both genomes show ca 2000 lineage specific L1 elements

Transposable elements

Alu repeat elements
3x more active in human lineage
human 7000 novel insertions
chimp 2300 lineage specific insertions

Chimp copy of Alu is close to the source gene.

Most human specific insertions belong to 2 new subfamilies (AluYa5 and AluYb8), which differ substantially from the original source Alu sequence.

Baboon has even higher Alu activity, so some activity loss may be specific to chimp lineage.

Pseudogenes
163 lineage-specific insertions in human and 168 in chimp genome. Largest gene class is ribosomal protein genes.

History of Alu elements

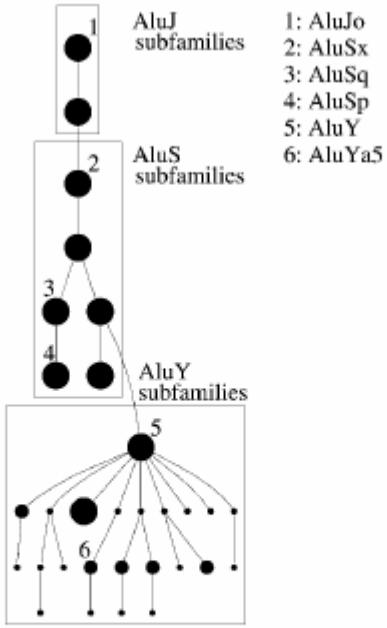
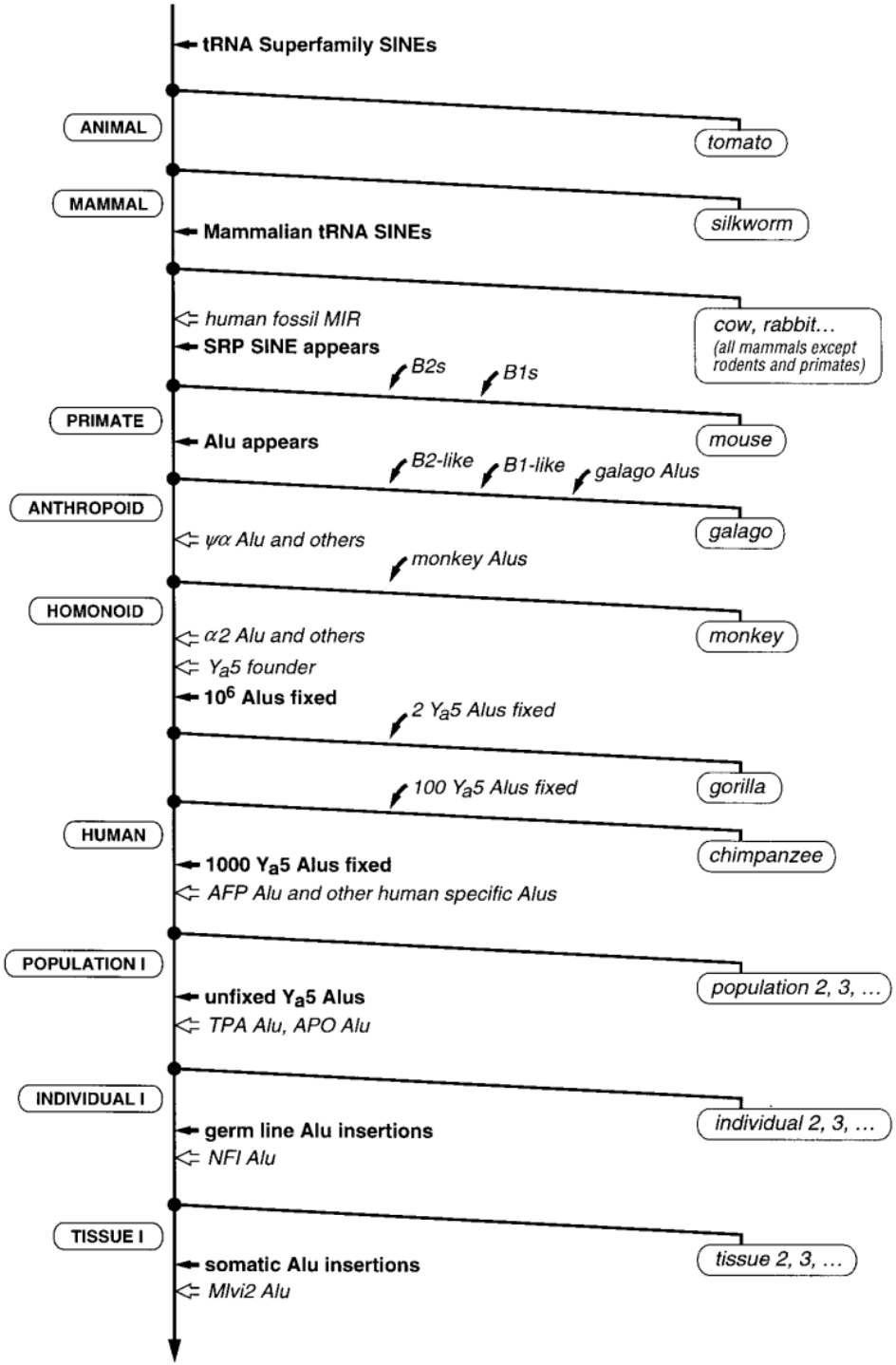


Figure 3. Evolutionary tree of the 31 subfamilies currently reported in Repbase Update. (Large nodes) Subfamilies with more than 10,000 elements; (medium nodes) 1000 to 10,000 elements; (small nodes) less than 1000 elements. Each of the 6 Repbase Update subfamilies listed in Figure 2 is labeled. The *AluJ*, *AluS*, and *AluY* classes of subfamilies are contained in boxes.

Whole-genome analysis of Alu repeat elements reveals complex evolutionary history
Alkes L. Price,1 Eleazar Eskin, and Pavel A. Pevzner
Genome Res. 14:2245–2252 (2004)

Schmid CW (1998)
Does SINE evolution preclude Alu function?
Nucl. Acids Res. **26**: 4541.



Gene evolution

13 454 genes with clear alignment and 1:1 orthology

29% identical DNA level?
5% with in-frame indels
median 2 amino acid substitutions per gene

Ka - fraction of substituted nonsynonumous nucleotides
Ks - fraction of substituted synonumous nucleotides

In human-chimp comparison average Ka/Ks = 0.23
This indicates that 77% of amino acid changes are deleterious and are eliminated

Ka/Ks = 0.13 in rat-mouse lineage

Gene content

53 known human genes not found in chimp genome

Some more inactivated?

Known differences in gene content:

caspase-12
mediator of apoptosis, triggers apoptosis in response to perturbed calcium homeostasis
gene inactivated in human lineage – may contribute to Alzheimer's disease

Segmental duplications

Several gene and genomic-based analyses suggest that the human genome is particularly enriched for genes that have emerged as a result of recent duplication.

We used two independent approaches to estimate the size and extent of chimpanzee (*Pan troglodytes*) duplications.

We first performed a self-comparison of the chimpanzee genome assembly using the whole-genome assembly comparison method (WGAC)

The second duplication detection method11 that uses the depth of coverage of random sequence read data against a reference sequence to identify duplicated sequence.We applied the whole-genome shotgun sequence detection (WSSD) strategy by mapping 23.7 million reads from chimpanzee against the human genome reference.

Segmental duplications

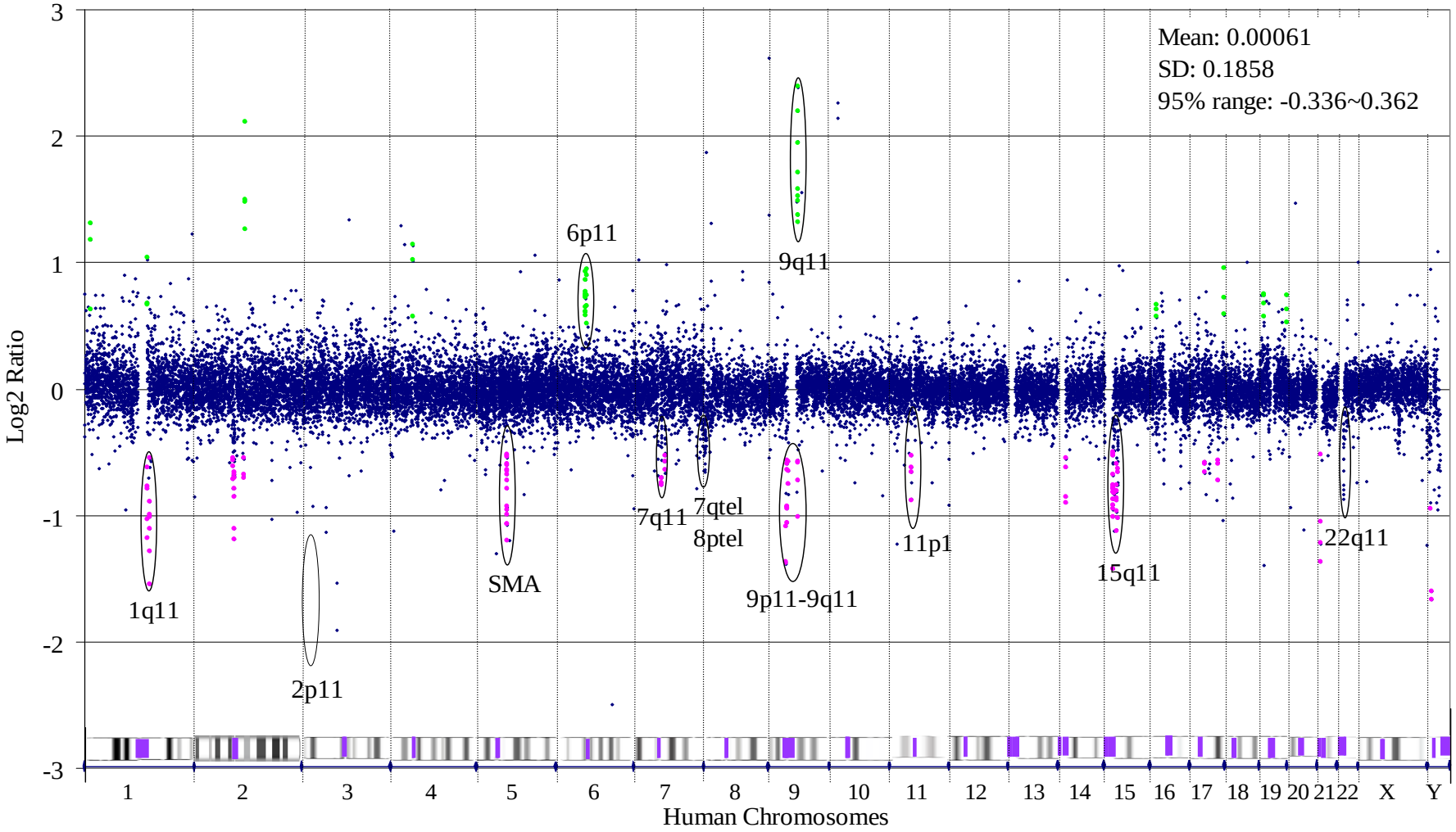
26.5 Mbp duplications are specific to human lineage

11.4 Mbp of human sequence that was duplicated in chimp (overall 32 Mbp of sequence)

Average duplication ca 50 kbp

177 complete or partial gene duplications in human
94 gene duplications in chimp

Fig. S6



Supplementary Figure S6. Array comparative genomic hybridization between human and chimpanzee (Clint) genomes. A full-tiling path microarray of 32,855 of human BAC clones 3 was analyzed by array comparative genomic hybridization using an anonymous human blood donor DNA as reference and Clint chimpanzee DNA as test DNA. A reverse-label replicate experiment was performed and the average log2 ratios (T/R) were plotted for the corresponding coordinate for each BAC. Regions with three or more consecutive BACs with log2 >0.5 (green) or three or more consecutive BACs < -0.5 (pink) are highlighted. Note the global decrease in chimpanzee signal intensity for pericentromeric regions when compared to human. The experiment was repeated with three unrelated chimpanzees with essentially the same result.

Segmental duplications

Supplementary Figure S6. Array comparative genomic hybridization between human and chimpanzee (Clint) genomes. A full-tiling path microarray of 32,855 of human BAC clones 3 was analyzed by array comparative genomic hybridization using an anonymous human blood donor DNA as reference and Clint chimpanzee DNA as test DNA. A reverse-label replicate experiment was performed and the average log2 ratios (T/R) were plotted for the corresponding coordinate for each BAC. Regions with three or more consecutive BACs with log2 >0.5 (green) or three or more consecutive BACs < -0.5 (pink) are highlighted. Note the global decrease in chimpanzee signal intensity for pericentromeric regions when compared to human. The experiment was repeated with three unrelated chimpanzees with essentially the same result (data not shown).