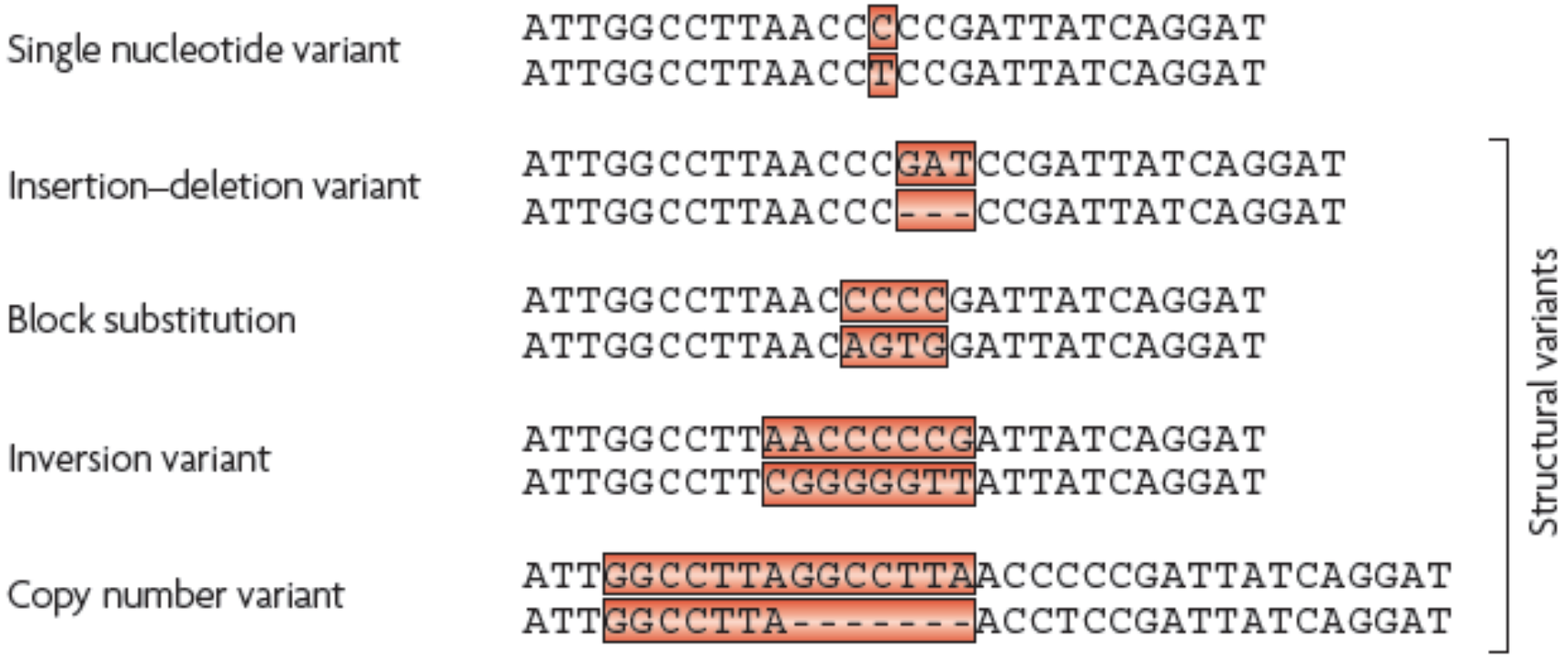
A Burst of Segmental Duplications in the African Great Ape Ancestor

Tomas Marques-Bonet *et al*, Nature 457: 877-881 (12 Feb 2009)

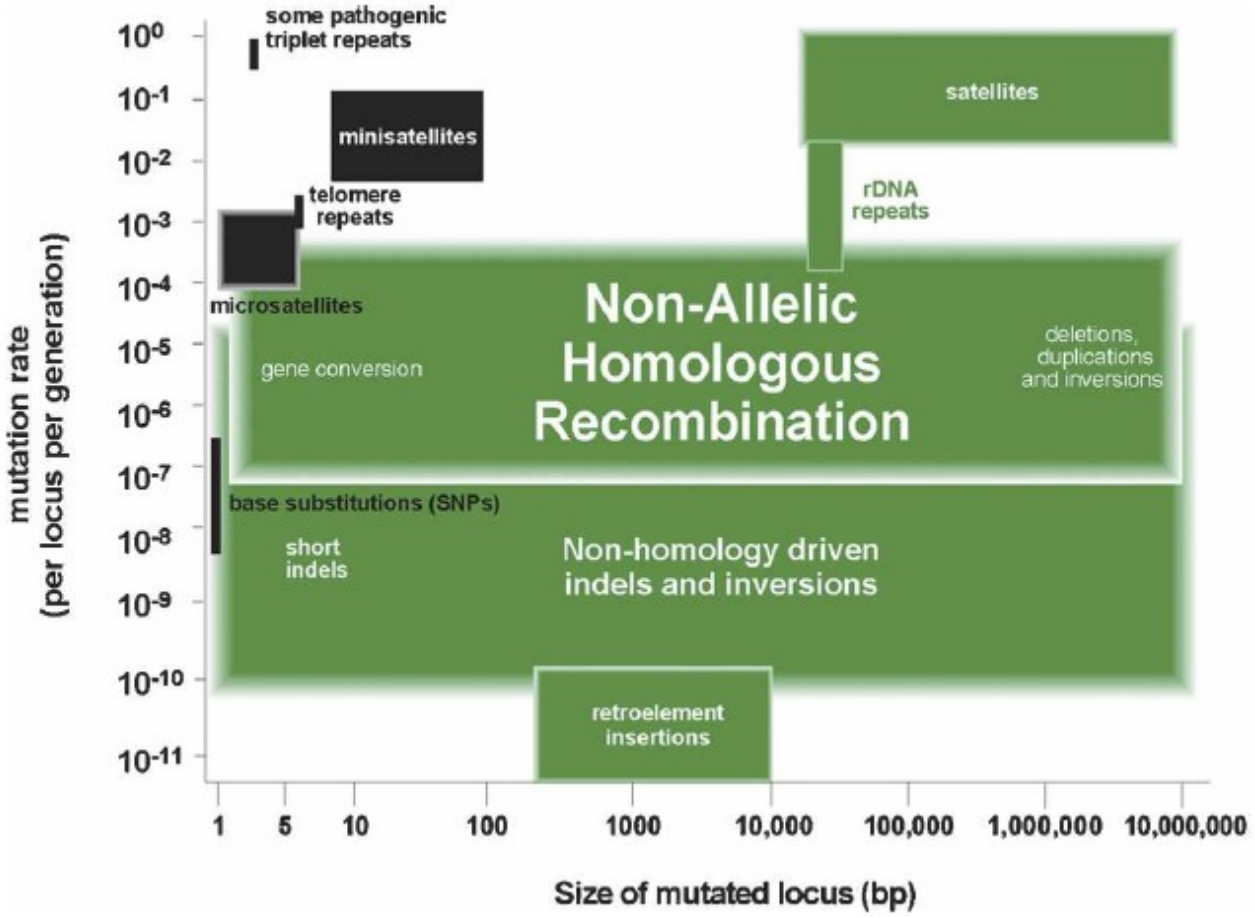
Jclub 31.03.2009

by Tarmo Puurand

Classes of human genetic variants

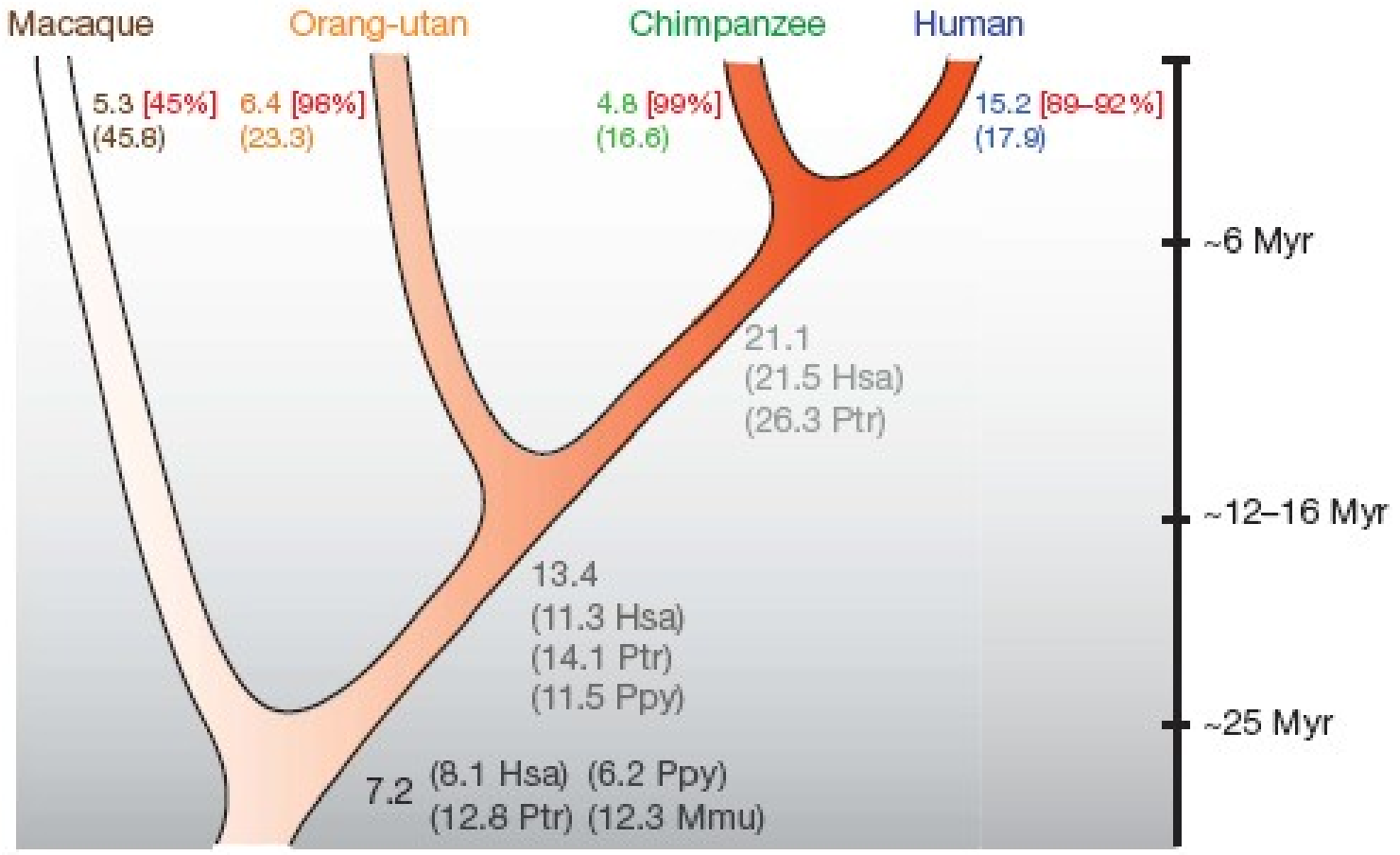


Markers mutation size and rate in humans



Freeman et al, Genome Research, 2006

Generally accepted human/great-ape phylogeny



Detection of segmental duplications

Supplementary Note Table 1. Primate genome datasets.

Species	Sample ID	Source	# WGS sequence reads	Phred Quality threshold	# of reads/alignments	# Non-redundant reads placed with quality threshold	# WGS reads required (> 3 standard deviation (Autosomes/X chrom))
Human	#N/A	70% one male human being + 30% pool individuals	27,449,655	20	24,577,141	22,402,464	81/51
Chimpanzee	NS06006	male chimpanzee (Clint)	31,366,275	27	25,493,514	23,393,800	105/59
Orangutan	PR01109	female orangutan (Susie)	25,514,441	30	19,297,789	17,764,564	78
Macaque	ID17573	female Macaque	22,590,543	27	16,769,443	13,380,372	75

The total number of reads, the number of reads mapped against the human reference genome, and the non-redundant number of reads mapped in the reference genome are shown. PHRED quality threshold used for every species and the ID of the samples are also reported.

Detection of duplications (I)

- Repeatmasked sequences (build35)
- Megablast
- WGS read-depth for 6/7 consequtive 5 kbp window
- Thresholds determination for every primate separately. Some kind of calibration: WGS sequence dataset vs. seqences obtained from BAC-based clones.

Detection of duplications (II)

- Reads: > 200 bp (20 for human, 27 for chimp, 30 for orangutan ans 27 for macaque)
- >94% human and great-ape alingments and >88% for human-macaque alignment
- Regions not in human were analyzed with help of sequence contigs from certain primate

Classes of primate segmental

Table 1 | Classes of primate segmental duplication

Category	Segmental duplications	Segmental duplications >20 kb	Validation (%)	Copy-number-corrected duplicated base pairs			
				Hsa	Ptr	Ppy	Mmu
Hsa	51,458,805	15,236,422	89–92	17,847,869	–	–	–
Ptr	11,239,390	4,789,874	99	–	16,583,946	–	–
Ppy	30,553,228	6,417,679	98	–	–	23,327,737	–
Mmu	24,962,092	5,360,646	45	–	–	–	45,810,964
Mmu*	35,493,466	7,715,410	85	–	–	–	18,266,656
Hsa/Ptr	32,392,480	21,061,194	NA	21,524,417	26,304,286	–	–
Hsa/Ptr/Ppy	25,450,827	13,402,545	NA	11,259,061	14,012,351	11,541,148	–
Hsa/Ptr/Ppy/Mmu	14,094,156	7,156,616	NA	8,092,997	12,820,607	6,176,876	12,542,691
Total	190,150,978	73,424,976	–	58,724,344	69,721,190	41,045,761	30,809,347

Duplications were divided into eight categories based on the WSSD analysis of each primate genome (subsequent analyses were restricted to segmental duplications >20 kb in length). Lineage-specific and shared duplication content are indicated. Percentage validation indicates the fraction of species-specific duplications confirmed by cross-species array comparative genomic hybridization. Because the human genome was used, we corrected for copy number and examined sequence contigs not aligned to the human genome (see Methods). Segmental duplications assigned to the Y chromosome were not considered.

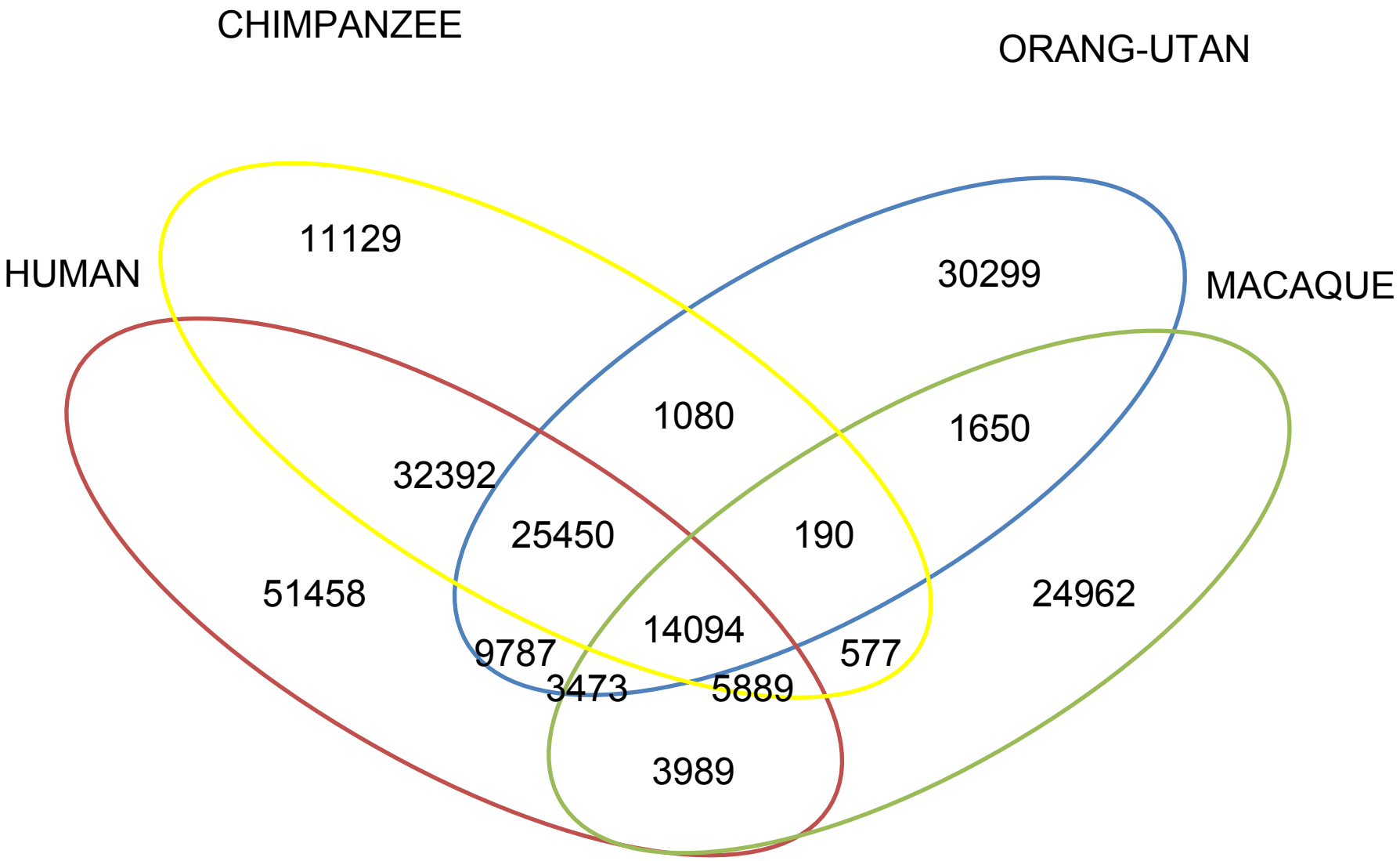
* Macaque segmental duplications detected in the macaque reference genome using WSSD and WGAC (<94% identity) approaches.

[1]Department of Genome Sciences, University of Washington and the Howard Hughes Medical Institute, Seattle, Washington 98195, USA. [2]Institut de Biologia Evolutiva (UPF-CSIC), 08003 Barcelona, Catalonia, Spain. [3]Sezione di Genetica-Dipartimento di Anatomia Patologica e Genetica, University of Bari, 70125 Bari, Italy. [4]Genome Sequencing Center, Washington University School of Medicine, St Louis, Missouri 63108, USA. [5]Institució Catalana de Recerca i Estudis Avançats (ICREA) and Instituto Nacional de Bioinformática (INB), Dr. Aiguader 88, 08003 Barcelona, Spain.

Distribution of SDs in kb (all)

Category	**Total bp**	N	AVG (length)	STD Dev (length)	MAX length	MIN length
HSA specific SDs	**51,458,805**	5,887	8,741	13,318	292,021	49
PTR specific SDs	**11,129,390**	1,169	9,520	18,223	341,154	21
PPY specific SDs	**30,299,228**	3,797	7,980	11,028	275,363	2
MMU specific SDs	**24,962,092**	2,463	10,135	8,698	149,378	41
HSA / PTR	**32,392,480**	2,018	16,052	22,340	345,000	36
HSA / PPY	**9,787,003**	1,586	6,171	6,823	71,000	27
HSA / MMU	**3,989,127**	740	5,391	6,495	93,000	21
PTR / PPY	**1,080,458**	244	4,428	4,938	44,610	51
PTR / MMU	**577,152**	100	5,772	7,094	52,050	545
PPY / MMU	**1,650,595**	321	5,142	4,158	27,000	26
HSA / PTR / PPY	**25,450,827**	1,770	14,379	17,384	234,000	21
HSA / PTR / MMU	**5,889,226**	782	7,531	7,844	63,000	21
HSA / PPY / MMU	**3,473,366**	529	6,566	6,838	47,260	9
PTR / PPY / MMU	**190,558**	69	2,762	2,580	15,330	325
HSA / PTR / PPY / MMU	**14,094,156**	1,011	13,941	13,489	168,780	38
Total	**216,424,463**	**22,486**	**9,625**	**13,692**	**345,000**	**2**

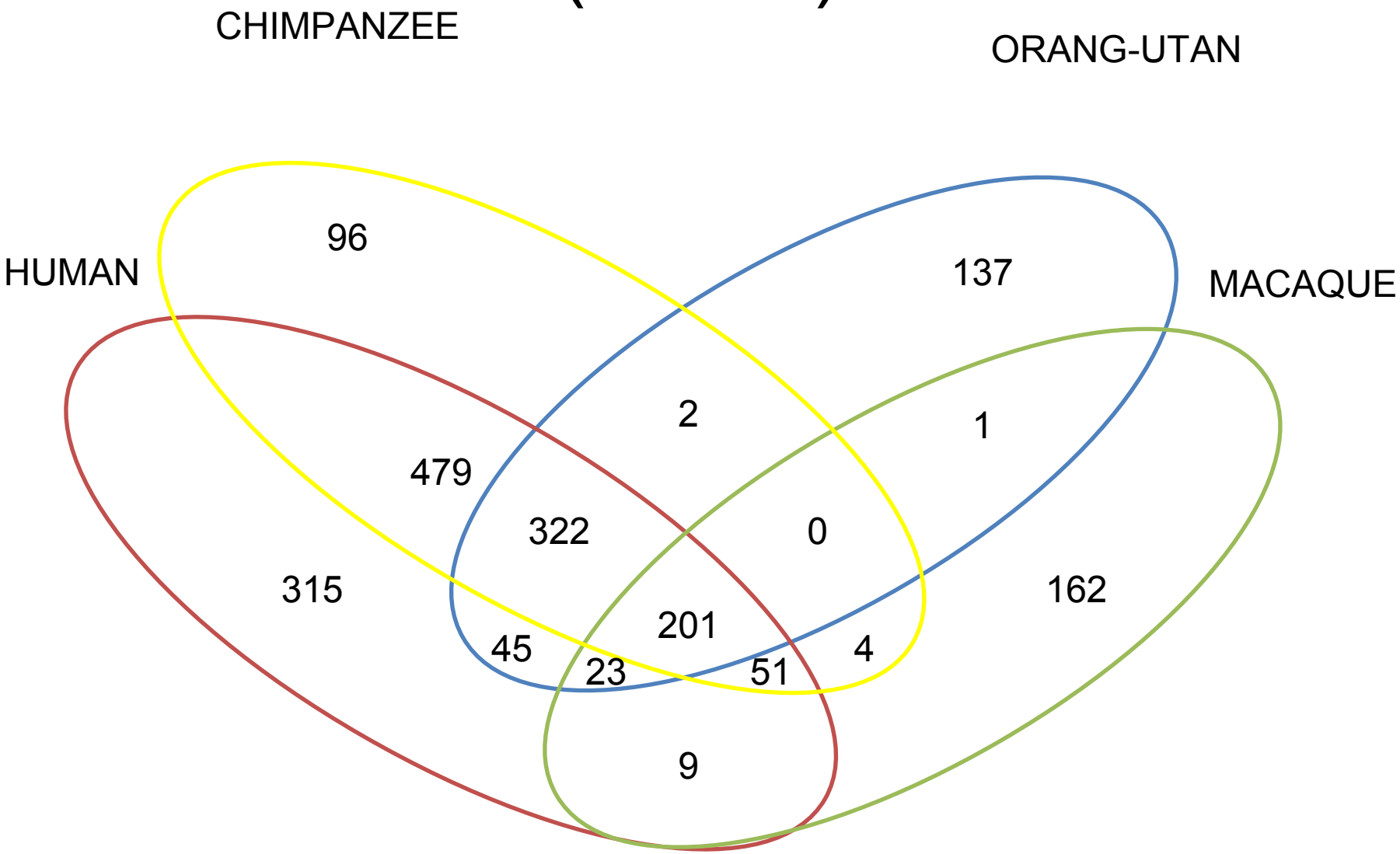
Distribution of SDs in kb (all)



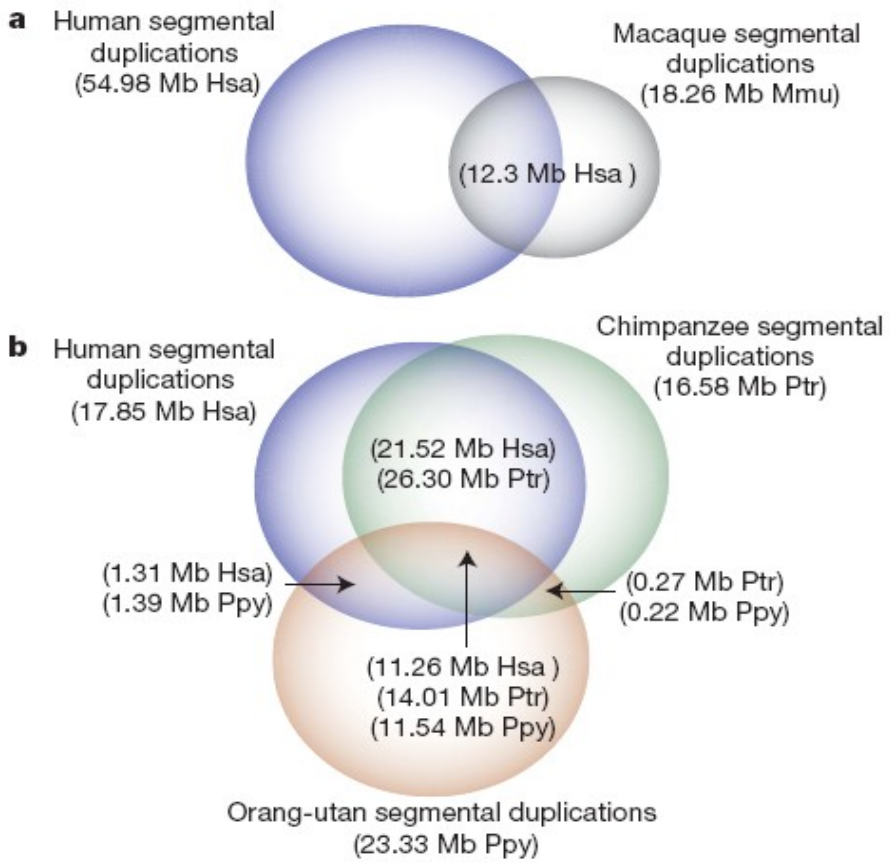
Distribution of SDs in events (>20kb)

Category	**Total bp (>20 kb)**	N	AVG (length)	STD Dev (length)	MAX length	MIN length
HSA specific SDs	**15,236,422**	315	48,370	37,561	292,021	20,035
PTR specific SDs	**4,789,874**	96	49,895	46,215	341,154	20,024
PPY specific SDs	**6,417,679**	137	46,844	39,283	275,363	20,076
MMU specific SDs	**5,360,646**	162	33,090	16,531	149,378	20,047
HSA / PTR	**21,061,194**	479	43,969	31,495	345,000	20,023
HSA / PPY	**1,452,735**	45	32,283	14,097	71,000	20,161
HSA / MMU	**392,712**	9	43,635	27,989	93,000	21,698
PTR / PPY	**86,700**	2	43,350	1,782	44,610	42,090
PTR / MMU	**135,794**	4	33,949	12,436	52,050	23,748
PPY / MMU	**27,000**	1	27,000		27,000	27,000
HSA / PTR / PPY	**13,402,545**	322	41,623	25,497	234,000	20,012
HSA / PTR / MMU	**1,545,552**	51	30,305	9,943	63,000	20,026
HSA / PPY / MMU	**704,864**	23	30,646	9,438	47,260	20,065
PTR / PPY / MMU						
HSA / PTR / PPY / MMU	**7,156,616**	201	35,605	15,546	168,780	20,025
Total	**77,770,333**	**1,847**	**42,106**	**30,462**	**345,000**	**20,012**

Distribution of SDs in events (>20kb)



Shared vs. lineage-specific duplications



Validation

- FISH- 58 lineage specific (14 human, 24 chimp, 20 orangu) and 38 complex regions > 40 kbp in length
- Interspecific array comparative genomic hybridisation

Supplementary Note Table 9. Segmental duplications and copy-number polymorphism.

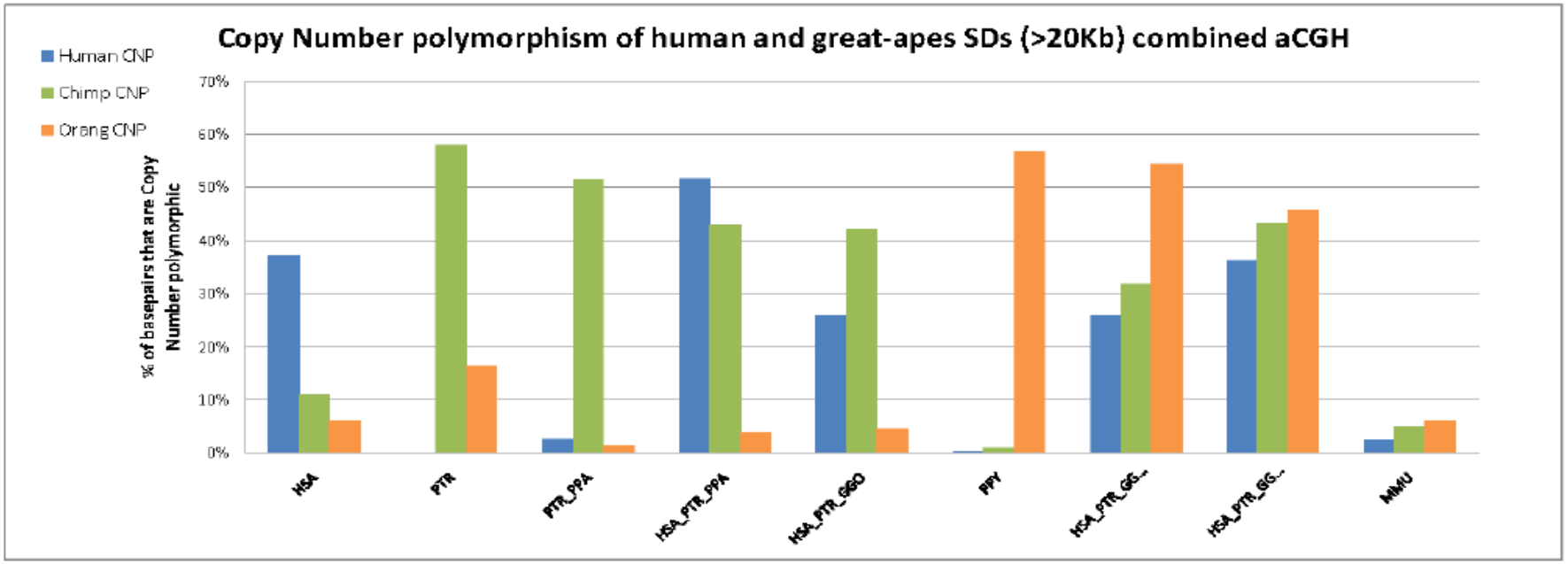
Human copy-number polymorphisms (n = 8 individuals)

SD Category	# CNV SD intervals	Total Length (bp)	CNV SD Intervals	Total Length CNV SD	% CNV SD
Human specific SDs	199	9,809,268	106	5,018,693	**33.9%**
Human/chimpanzee shared SDs	300	12,222,058	179	8,839,136	**42.0%**
Human/chimp/orang shared SDs	235	10,303,447	87	3,099,098	**23.1%**
Human/chimp/orang/macaque shared SDs	145	5,114,155	56	2,042,461	**28.5%**
Chimpanzee-specific SDs	91	4,684,302	2	42,140	**0.9%**
Orangutan-specific SDs	134	6,344,870	2	51,655	**0.8%**
Macaque-specific SDs	148	4,894,873	12	414,331	**7.8%**
Total	1,252	53,372,973	444	19,507,514	**26.8%**

Chimpanzee copy-number polymorphisms (n = 8 individuals)					
SD Category	**# CNV SD intervals**	**Total Length (bp)**	**CNV SD Intervals**	**Total Length CNV SD**	**% CNV SD**
Human-specific SDs	255	12,842,592	50	1,985,369	**13.4%**
Human/chimpanzee shared SDs	312	12,224,102	167	8,837,092	**42.0 %**
Human/chimp/orang shared SDs	204	8,591,738	118	4,810,807	**35.9%**
Human/chimp/orang/macaque shared SDs	110	3,761,322	91	3,395,294	**47.4%**
Chimpanzee-specific SDs	35	1,443,956	58	3,282,486	**69.4%**
Orangutan-specific SDs	135	6,343,149	1	53,376	**0.8%**
Macaque-specific SDs	149	4,829,875	11	479,329	**9.0%**
Total	1,200	50,036,734	496	22,843,753	**31.3%**

Orangutan copy-number polymorphisms (n = 8 individuals)					
SD Category	**# CNV SD intervals**	**Total Length (bp)**	**CNV SD Intervals**	**Total Length CNV SD**	**% CNV SD**
Human-specific SDs	276	13,794,990	29	1,032,971	**7.0%**
Human/chimpanzee shared SDs	452	20,060,117	27	1,001,077	**4.8%**
Human/chimp/orang shared SDs	146	5,667,445	176	7,735,100	**57.7%**
Human/chimp/orang/macaque shared SDs	107	3,688,249	94	3,468,367	**48.5%**
Chimpanzee specific SDs	88	4,560,303	5	166,139	**3.5%**
Orangutan specific SDs	72	2,934,820	64	3,461,705	**54.1%**
Macaque specific SDs	146	4,826,751	14	482,453	**9.1%**
Total	1,287	55,532,675	409	17,347,812	**23.8%**

CNV-s in SD regions



Supplementary Note Fig. 5. Copy-number polymorphism of human and great-ape SDs. *In this figure, SDs were further categorized (see Fig. 2c) using arrayCGH information from gorilla and bonobo. The same trends reported in the text are observed.*

Extreme duplications

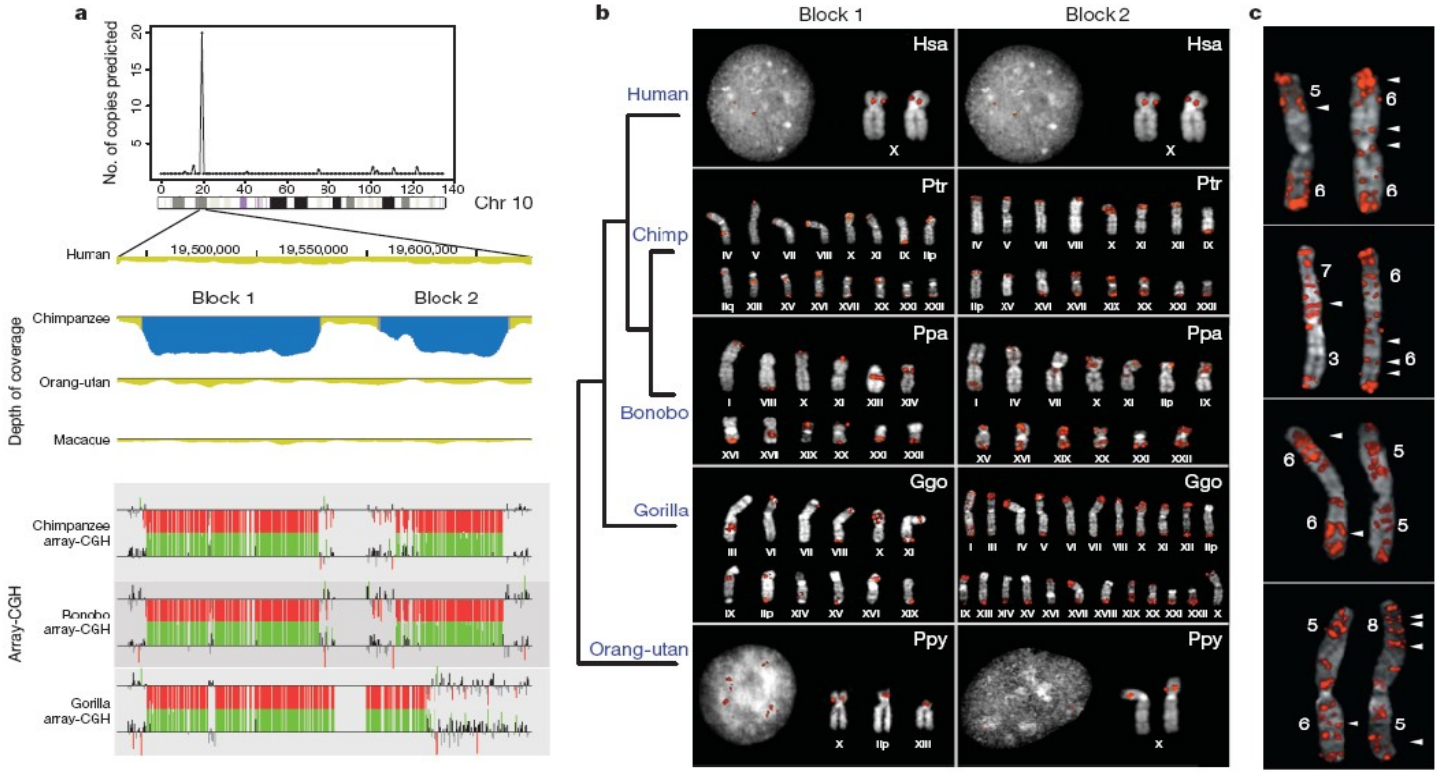


Figure 3 | Convergent gene duplication expansion in African great apes but not humans. **a**, Two regions on chromosome 10 have expanded in chimpanzee, gorilla and bonobo when compared to human based on computational and interspecific array-CGH experiments (see Fig. 1 legend). **b**, FISH confirms 23–50 copies in chimpanzee and bonobo (Ppa, *Pan paniscus*), and >100 copies in gorilla (Ggo, *Gorilla gorilla*) (Methods). End-sequence pair analysis using gorilla and chimpanzee WGS sequences reveals that all but the ancestral location are non-orthologous, indicating independent expansions in both lineages. **c**, Detailed analysis of eight homologues of gorilla chromosome 1 reveals interstitial locations of the block 2 duplication that show variation both in copy number and in terms of location.

Rates of duplications (Mbp)

Supplementary Note Table 14. Great-ape comparisons.

	subt/1000bp	SDs (Mb)	Rate Mb/subt per 1000 bp	Million Year per branch	SDs (Mb)	Rate Mb/Mya
Human terminal branch	5.4	13.6	**2.519**	6	13.6	**2.267**
Chimpanzee terminal branch	5.56	6.1	**1.097**	6	6.1	**1.017**
Human/chimpanzee shared branch	1.07	9.32	**8.710**	2	9.32	**4.660**
Gorilla terminal branch	7.19			8		
Human/chimpanzee/gorilla shared branch	7.62	16.82	**2.207**	4	16.82	**4.205**
Orangutan terminal branch	15.03	20.33	**1.353**	12	20.33	**1.694**
Human/chimpanzee/orangutan shared branch	14.7213	15.97	**1.085**	13	15.97	**1.228**

The rates of segmental duplication (>20 kbp) accumulation on different branches were compared as a function of millions of years since divergence and as a function of the genetic distance (single basepair substitutions)[38] between the species.

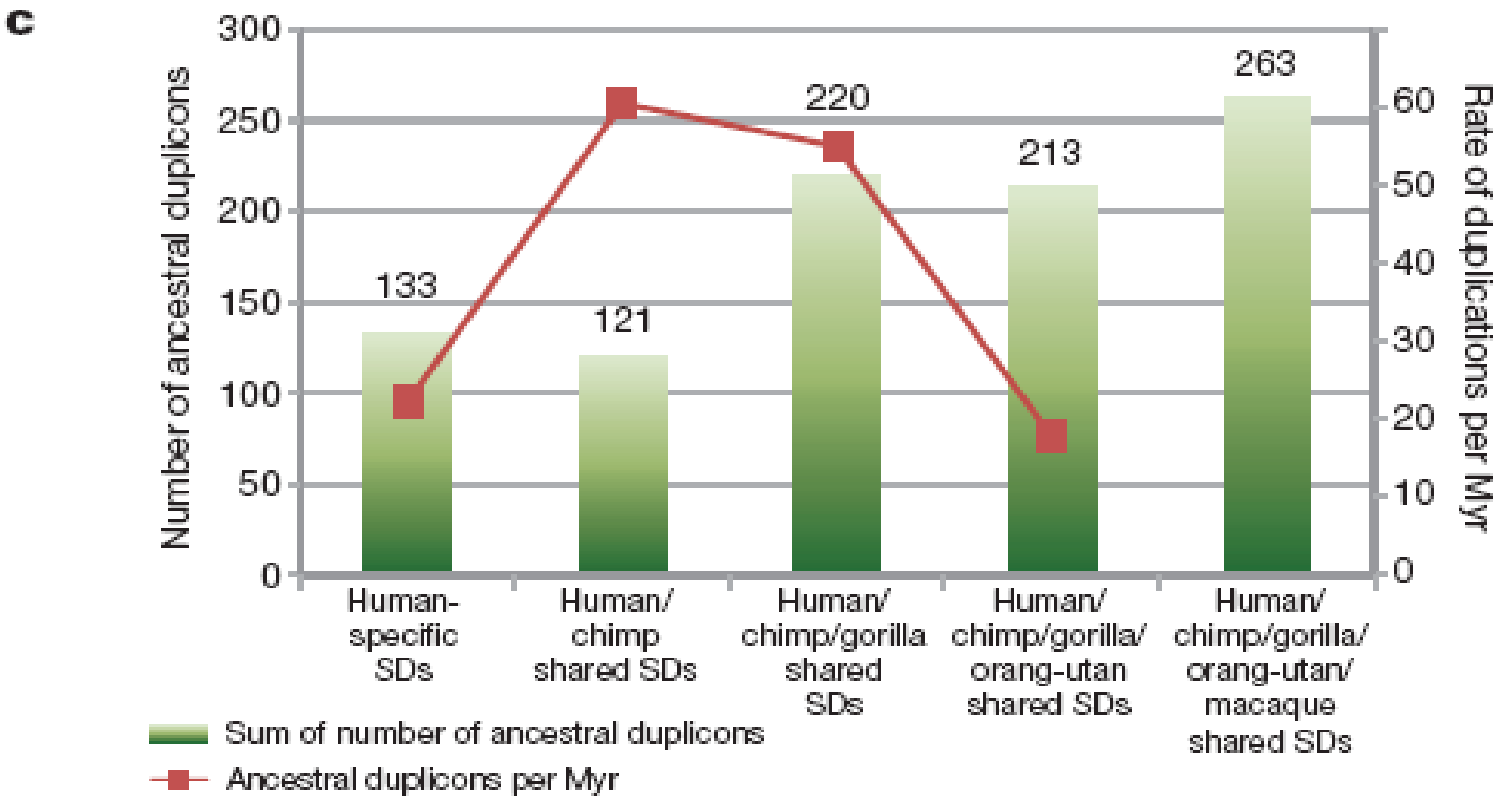
Rates of duplications (events)

Supplementary Note Table 15. Hominid rates of duplication (events >20 kbp).

	Number of chained sub-units human Duplicons	**Rate of Duplications (Events /Myr)**	**subt/1000bp**	**Rate of Duplications (Events /substitutions)**
HSA specific	133	**22.17**	5.40	**24.63**
PTR specific			5.56	
HSA/PTR shared	121	**60.50**	1.07	**113.08**
HSA/PTR/GGO shared	220	**55.00**	7.62	**28.87**
PPY specific			15.03	
HSA/PTR/PPY Shared	213	**16.38**	14.72	**14.47**

950 duplicons detected previously[11] were used as a surrogate for duplication events. Two measures of time were applied to calculate the rates: a) million years of divergence and b) genetic distance estimates[38].

Rates of SD


c
Number of ancestral duplicons
Rate of duplications per Myr
300
250
200
150
100
50
0
60
50
40
30
20
10
0
133
121
220
213
263
Human-specific SDs
Human/chimp shared SDs
Human/chimp/gorilla shared SDs
Human/chimp/gorilla/orang-utan shared SDs
Human/chimp/gorilla/orang-utan/macaque shared SDs
Sum of number of ancestral duplicons
Ancestral duplicons per Myr

Other analysis

- Nonrandom distribution of primate SD
- Gene duplication analyses
- SD and disease susceptibility loci
- Duplication status vs. copy number
- Lineage specific deletions

Venter vs. Watson (CNV-s)

Supplementary Note Table 3. Copy-number variation of shared and individual-specific human SDs.

Cat_SD	Invariant	%	CNVs	%	Grand Total	Fisher exact test P-value (vs. Shared)
SHARED	109	*18.3*	486	*81.6*	595	
VENTER	18	*41.9*	25	*58.1*	43	0.0001989417
WATSON	17	*60.7*	11	*39.3*	28	0.0000002341
Grand Total	*144*		*522*		*666*	