Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee

Supplementary Note

Mario Ventura, Claudia R. Catacchio, Can Alkan, Tomas Marques-Bonet, Saba Sajjadian, Tina A. Graves, Fereydoun Hormozdiari, Arcadi Navarro, Maika Malig, Carl Baker, Choli Lee, Emily H. Turner, Lin Chen, Jeffrey M. Kidd, Nicoletta Archidiacono, Jay Shendure, Richard K. Wilson, and Evan E. Eichler

Table of Contents

I. Gorilla Comparative Cytogenetic Framework

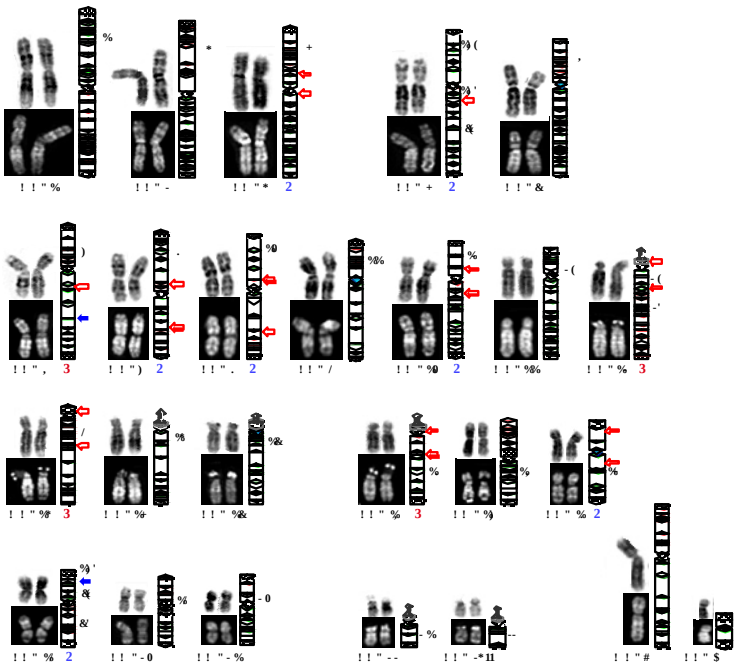
To determine the long-range synteny between the gorilla and human genomes, we began by assaying 788 BAC clones as probes distributed on average every 4 Mbp across the human genome. Bicolor and single-color FISH were performed on gorilla and human metaphase chromosomal preparation to determine synteny in the marker order. This cytogenetic comparative analysis between the two species allowed us to confirm previously reported chromosomal rearrangements and precisely identify corresponding breakpoints(Egozcue and Chiarelli 1967; Miller et al. 1974; Dutrillaux 1980; Yunis and Prakash 1982; Montefalcone et al. 1999; Muller et al. 2000; Carbone et al. 2002; Eder et al. 2003; Locke et al. 2003; Misceo et al. 2003; Ventura et al. 2003; Ventura et al. 2004; Misceo et al. 2005; Cardone et al. 2006; Cardone et al. 2007; Stanyon et al. 2008). Where possible, split signals and breakpoints relative to the human genome were identified (Supplementary Note Table 1 and Table S1). Digital images were obtained using a Leica DMRXA2 epifluorescence microscope equipped with a cooled CCD camera (Princeton Instruments). Cy3-dCTP, FluorXdCTP, Cy5-dCTP, and DAPI fluorescence signals, detected with specific filters, were recorded separately as gray-scale images. Pseudocoloring and merging of images were performed using Adobe Photoshop software.

Supplementary Note Table 1. GGO chromosomal rearrangement breakpoints defined by human clones

	CLONE NAME	Acc. N.	HSA MAP	NCBI35	REMARKS
HSA2q (GGO11)	NA	M73018	2q14.1	chr2:114,075,938-114,077,717	Human FUSION
HSA2p(GGO12)	RP11-316G9	AC009958	2p11.2	chr2:89,619,699-89,830,899	Human/Chimpanzee INVERSION
	BREAKPOINT A				
	CEN				
	RP11-44L16	BES	2q13	chr2:112,159,412-112,326,085	
	BREKPOINT B				
	RP11-243I5	BES	2q14.1	chr2:114,221,969-114,408,721	
HSA4(GGO3)	RP11-317G22	AC020593	4p12	chr4:48,735,458-48,919,660	Gorilla INVERSION
	BREAKPOINT A				
	CEN				
	RP11-669F1	BES	4q13.2	chr4:68,909,436-69,040,975	
	BREAKPOINT B				
	RP11-401E5	AC093829	4q21.23-21.3	chr4:70,538,559-70,708,041	
HSA7(GGO6)	RP11-243I17	BES	7q11.23	chr7:75,428,943-75,590,441	Human/Chimpanzee INVERSION
	BREAKPOINT A				
	RP11-982E3	BES	7q11.23	chr7:76,494,214-76,685,936	
	RP11-163E9	BES	7q22.1	chr7:101,494,176-101,666,161	
	BREAKPOINT B				
	RP11-803J14	BES	7q22.1	chr7:101,791,124-102,097,972	
HSA8(GGO7)	RP11-363L24	AC009563	8p12	chr8:31,053,114-31,243,056	Gorilla INVERSION (identified Breakpoint A)
	RP11-457E21	BES	8q21.2	chr8:85,811,907-85,981,817	
	BREAKPOINT B				
	RP11-219B4	AC011773	8q21.2	chr8:86,197,050-86,357,332	
HSA9(GGO13)	TEL				Human/Chimpanzee INVERSION
	BREAKPOINT A				
	RP11-130C19	AL136979	9p24.3	chr9:615,148-812,246	
	CEN				
	BREAKPOINT B				
	RP11-203L2	BES	9q21.11	chr9:68,528,295-68,682,365	
HSA10(GGO8)	RP11-378M14	BES	10p12.1	chr10:26,965,000-27,144,365	Gorilla INVERSION
	BREAKPOINT A				
	RP11-774G16	BES	10p11.3	chr10:28,304,146-28,502,830	
	RP11-598H8	BES	10q22.3	chr10:80,609,011-80,791,349	
	BREAKPOINT B				
	RP11-715A21	BES	10q22.3	chr10:80,916,595-81,089,979	
HSA12(GGO10)	RP11-737A10	BES	12p12.2	chr12:21,112,105-21,292,548	Gorilla INVERSION (identified Breakpoints A and B)
	RP11-766N7	BES	12q14.3	chr12:63,500,649-63,684,874	
HSA14(GGO18)	TEL				Gorilla INVERSION
	BREAKPOINT A				
	CEN				
	RP11-453F20	BES	14q21.3	chr14:44,679,767-44,872,985	
	BREAKPOINT B				
	RP11-760N14	BES	14q21.3	chr14:45,092,484-45,232,201	
HSA18(GGO16)	TEL				Human INVERSION
	BREAKPOINT A				
	RP11-78H1	BES	18p11.32	chr18:2,136,811-2,307,213	
	CEN				
	BREAKPOINT B				
	RP11-10G8	BES	18q11.2	chr18:17,274,438-17,431,001	
HSA17p-5pq(GGO19)	RP11-1082K15	BES	5q14.1	chr5:80,583,928-80,770,628	Gorilla TRANSLOCATION
HSA5q-17pq(GGO4)	RP11-385D13	AC005838	17p12	chr17:15,367,740-15,435,530	

Other than the fusion between chromosomes 12 and 13 that gave rise to human chromosome 2, the resulting gorilla karyotype can be distinguished from human by eight pericentric inversions (2, 4, 8, 9, 10, 12, 14 and 18), one paracentric inversion on chromosome 7, and one translocation (t5;17) (Supplementary Note Figure 1). Previous data reported a pericentric inversion involving the centromere on human chromosome 1 between human cytogenetic bands 1p11.2 (a 154.2 kbp interval) and 1q21.1 (breakpoint region to a 562.6 kbp interval)(Szamalek et al. 2006a; Szamalek et al. 2006b). Several FISH experiments were performed to verify this inversion. Due to the abundance of segmental duplications (SDs) in the pericentromeric region of chromosome 1, we were not able to distinguish between an inversion or centromere repositioning (data not shown). In order to define the ancestral chromosomal form for the rearranged chromosomes, we performed reiterative FISH experiments utilizing the same panel of probes on orangutan (*Pongo pygmeaus*, PPY) and on macaque (*Macaca mulatta*, MMU) used as an outgroup of the great apes. We showed for chromosomes 4, 7, 8, 10, 12 and 14 that human retained the chromosomal structure most resembling the ancestral form; likewise, gorilla showed the ancestral form for chromosomes 2, 9, and 18

(Supplementary Note Figure 1).



Supplementary Note Figure 1. Gorilla karyotype ideogram and overview of chromosomal rearrangements. A represented-banded, Q-banded and a schematic ideogram is shown for each gorilla chromosome. Red and green arms represent p and q arms, respectively, according to human chromosome organization. Chromosomal rearrangements compared to human are shown with arrows next to the chromosome ideogram. Empty arrows indicate a cytogenetically defined breakpoint, blue arrows (in GGO6 and GGO19) indicate a breakpoint not completely resolved due to the enrichment of large SDs, and filled red arrows indicate a breakpoint fully characterized at the sequence level. A, ancestral chromosomal synteny; D, derivative chromosomal synteny.

II. Gorilla Genome Sequencing

Sequencing libraries were constructed from genomic DNA isolated from whole blood obtained from a male silverback gorilla housed at the Lincoln Park Zoo (Kwan, *Gorilla gorilla*, studbook #1107). The blood was drawn and stored in EDTA pretreated vials and DNA isolation was performed using Puregene Core KitA (Qiagen). Library adaptors and oligonucleotides were synthesized by Integrated DNA Technologies and resuspended in nuclease-free water to a stock concentration of 100 mM. Double-stranded library adaptors SLXA_PE_ADAPT_Up ([phos]GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG) and SLXA_PE_ADAPT_Lo (ACACTCTTTCCCTACACGACGCTCTTCCGAATC*T) were prepared to a final concentration of 50 mM by

incubating equimolar amounts at 95ºC for 5 min and then leaving the adaptors to cool to room temperature in the heat block. The annealed adaptors were labeled and stored at -20ºC.

Shotgun libraries were generated from 5 µg of genomic DNA (gDNA) using a modified Illumina protocol(Bentley et al. 2008). gDNA in 300 ml 1xTris-EDTA was first sonicated for 2 cycles of 15 min each using a Bioruptor (Diagenode) set at high, then purified by QIAQuick kit (Qiagen), and finally eluted in 32 µl EB buffer. Shared DNA was end-repaired for 45 min in a 50 µl reaction volume with 1X End-It Buffer, 1X dNTP mix, and 1X ATP from the End-It DNA End-Repair Kit (Epicentre). Further purification and elution in 89 µl of EB buffer were performed after the end-tailing step using a QIAQuick kit (Qiagen). The fragments were then A-tailed for 30 min at 70ºC in a 100 µl reaction volume with 1X PCR buffer (Applied Biosystems) containing 1.5 mM MgCl2, 0.5 mM dATP, and 2.5U AmpliTaq DNA polymerase (Applied Biosystems). Purification and elution were further performed by QIAQuick kit (Qiagen) in 12 µl of EB buffer. Next, library adaptors SLXA_PE_ADAPT_Up and SLXA_PE_ADAPT_Low were ligated to the A-tailed sample in a 30 µl reaction volume with 1x Quick Ligation Buffer (New England Biolabs) with 2.5 µl Quick T4 DNA Ligase (New England Biolabs) and each adaptor in 10X molar excess of sample. Samples were purified on QIAQuick columns (Qiagen) and DNA concentration determined on a Nanodrop-1000 (Thermo Scientific).

Each sample was subsequently size-selected for fragment sizes between 250–350 bp by gel electrophoresis on a 6% TBE-polyacrylamide gel (Invitrogen). A gel slice containing the fragments of interest was excised and transferred to a siliconized 0.5 µl microcentrifuge tube (Ambion) with a 20 G needle-punched hole in the bottom. This tube was placed in a 1.5 µl siliconized microcentrifuge tube (Ambion) and centrifuged in a tabletop microcentrifuge at 14,000 rpm for 5 min to create a gel slurry that was then resuspended in 200 µl 1x Tris-EDTA and incubated at 65ºC for 2 h, with periodic vortexing. This allowed for passive elution of DNA, and the aqueous phase was then separated from gel fragments by centrifugation through 0.2 mm NanoSep columns (Pall Life Sciences) and the DNA recovered using QIAQuick columns (Qiagen) in 30 µl of EB buffer. The eluted DNA was used in 3 aliquots of 10 µl for the following PCR reaction with 1x iProof High-Fidelity Master Mix (Bio-Rad), 0.1x Syber Green, and 200 mM each of primers
SLXA_FOR_AMP
(AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T) and
SLXA_REV_AMP
(CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATTC*T) in a total volume of 50 µl per tube. Amplification conditions were 98ºC for 30 s, 29 cycles at 98ºC for 10 s, 60ºC for 30 s and 72ºC for 50 s, and finally 72ºC for 10 min. All reactions were monitored by an RT-PCR machine and stopped before reaching the plateau phase during the amplification (at 16 cycles). PCR products were purified across three QIAQuick columns (Qiagen) and all eluants pooled. A second size selection on 6% TBE-polyacrylamide gel (Invitrogen) and PCR amplification using SLXA_FOR_AMP and SLXA_REV_AMP primers was performed as above.

All sequencing of postenrichment shotgun libraries was carried out on an Illumina Genome Analyzer II as paired-end 36 bp reads, following the manufacturer's protocols and using the standard sequencing primer. Image analysis and base calling was performed by the Genome Analyzer Pipeline version 1.3 with default parameters, but with no prefiltering of reads by quality.

III. BAC End Mapping

A gorilla large-insert genomic bacterial artificial chromosome (BAC) library, CHORI-255 (http://bacpac.chori.org/gorilla255.htm), consisting of 176,000 clones was end sequenced by Washington University Genome Sequencing Center as part of a white-paper initiative to discover structural variants and facilitate the sequence and assembly of the gorilla genome. We downloaded gorilla BAC end sequences (BES) from the NCBI Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi) using the query "species_code = 'Gorilla Gorilla' and center_name = 'WUGSC' and trace_type_code = 'Cloneend'". We successfully aligned 353,761 BES to the human reference genome NCBI35 using MegaBLAST (parameters: -p 80 -s 90 –v 7 –b 7 –w 12 –t 21) for initial recruitment of map locations. A score threshold (-s 90) allowed for the flexibility to detect shorter alignments with higher similarity or longer alignments with lower sequence identity, such as those due to base-calling errors in poor quality trace. Additionally, an 80% identity threshold (-p 80) was set to avoid recruiting numerous pairwise-representing related transposable/repetitive elements. Following the procedures previously described(Tuzun et al. 2005; Kidd et al. 2008), we optimally realigned all initially recruited BES using an in-house Needleman-Wunsch implementation (match = +10, mismatch = -8, gap opening = -20, gap extension = -1, no penalty for terminal gaps)(Needleman and Wunsch 1970). The percent identity for each global alignment was then recalculated base-by-base to include only those aligned bases where BAC end bp were of high quality (any bases with a phred score <30 were ignored). Each paired-end map location is scored by a previously described, 13-point scoring scheme(Tuzun et al. 2005) to select the "best" or "tied" map locations. Finally, we identified putative rearrangements by requiring at least two independent discordant BAC clones to support the same type of rearrangement at the same genomic locus(Newman et al. 2005; Tuzun et al. 2005; Kidd et al. 2008)(Table S2).

IV. BAC Sequence Analysis

BACs spanning regions of SD or evolutionary rearrangement breakpoints were completely sequenced and assembled using capillary-based sequencing methods. The corresponding genomic sequence of each insert was annotated for genes, common repeats, and SDs. Common repetitive sequences and short tandem repeats were identified using RepeatMasker(Chen 2004), Tandem Repeats Finder(Benson 1999), and WindowMasker(Morgulis et al. 2006). SDs were defined by SegDupMasker(Jiang et al. 2008) and by identifying regions of excess read-depth (whole-genome shotgun sequence detection or WSSD)(Bailey et al. 2002). Whole-genome shotgun (WGS) sequence data were obtained from four hominids (human (NA18507), chimpanzee (Clint), gorilla (Kwan) and

orangutan (Bornean)). WGS sequence data were fragmented and mapped against each masked BAC sequence using mrFast(Alkan et al. 2009). Read-depths were calculated and normalized based on %GC content in 5 kbp (unmasked) windows and duplication thresholds were set based on an analysis of control regions within NCBI35 (as described previously(Alkan et al. 2009; Marques-Bonet et al. 2009)). All putative SD regions greater than 10K were reported and copy number estimates were also calculated in sliding 1 kbp windows. All BAC sequences were compared to the human reference genome (NCBI35) using MegaBLAST (parameters: -D 2 -v 7 -b 7 -e 1e-40 -p 80 -s 90 -W 12 -t 21 -F F). We identified all alignments larger than 1K with identity greater than 90% and concatenated colinear regions >5 kbp distance creating larger pieces. The largest and most identical regions in the human genome were compared to the gorilla sequence using Miropeats(Parsons 1995) and Parasight (Bailey et al., unpublished). Breakpoints were further refined by local alignment (ClustalW).

V. Structural Variation Detection

Next-generation gorilla genome sequence datasets were aligned to the human reference genome using the mrFAST mapping algorithm(Alkan et al. 2009). Deletions and mobile element insertions were detected using VariationHunter (Hormozdiari et al. 2009; Hormozdiari et al. 2010) while SDs (>20 kbp) were detected and copy number quantified using measures of read-depth(Sudmant et al. 2011).

Read Pair Analysis

We mapped 1.6 billion reads generated from three paired-end library preparations and sequenced with Illumina Genome Analyzer IIx to the human reference genome [NCBI build 36 (NCBI36)] using mrFAST(Alkan et al. 2009), a read mapping algorithm that tracks all possible locations of the reads within given edit distance. For this study, we required an edit distance of ≤2 bp for the 36 bp reads. We calculated the average paired-end span and standard deviation statistics (Supplementary Note Table 2) and classified discordant read pairs with mapping span >average+4std. Using VariationHunter(Hormozdiari et al. 2009), we initially predicted 21,431 deletions (56.4 Mbp); however, conversion of the NBCI36 coordinates to NCBI35 coordinates using the LiftOver tool reduced our call set to a total of 21,323 deletions that correspond to 52.6 Mbp, overlapping 4,744 genes (Supplementary Note Table 3 and Supplementary Note Figure 2).

Supplementary Note Table 2

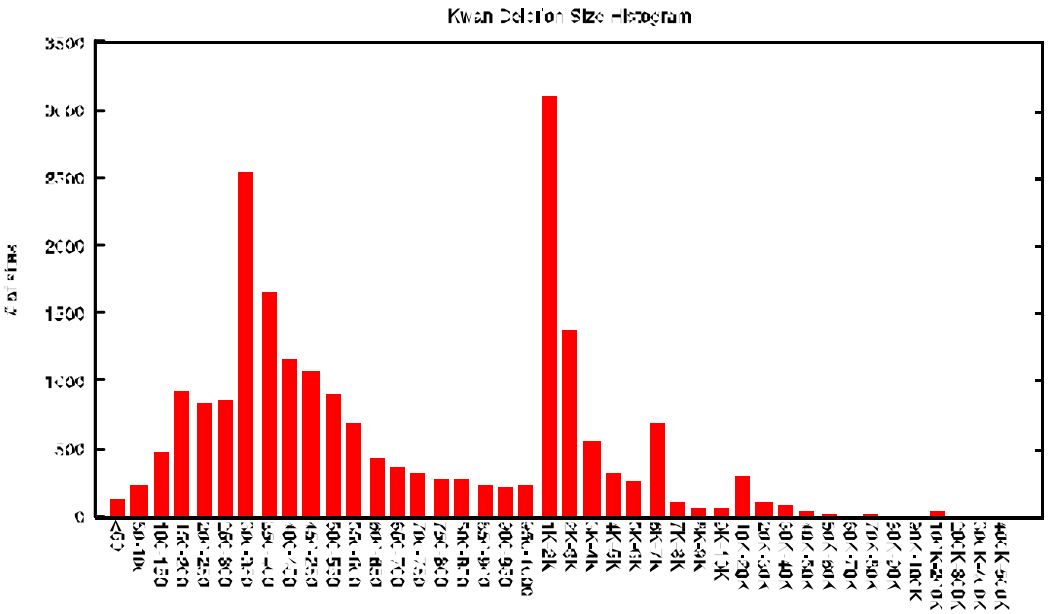
Supplementary Note Table 2. Read length and insert size statistics for the Illumina sequencing libraries

Library	Number of Reads	Read Length	Average Span	Standard Deviation
Library D	91,748,178	36	262.83	43.16
Library G	1,165,314,904	36	319.65	74.95
Library H	362,865,514	36	129.2	33.93

Supplementary Note Table 3. Basic statistics of the deletion predictions using read-pair analysis

		Predicted			Validated		
		Intervals	# bp	Genes	Intervals	# bp	Genes
All		21,323	52,621,636	4,744	*NA*	*NA*	*NA*
Filtered		8,873	25,704,661	2,521	*NA*	*NA*	*NA*
	>=500bp	5,125	24,555,248	1,731	*NA*	*NA*	*NA*
	>=500bp and >=10 probes	2,755	18,051,395	1,142	1,820	6,745,878	593

Filtered intervals have <30% intersection duplications we detected in Kwan, and include <80% repeat content. We required >=1 bp intersection with the RefSeq genes in this table.



Supplementary Note Figure 2. Length distribution of deletions based on paired-end read placements. The deletion sizes are shown in 50 bp, 1 kbp, 10 kbp and 100 kbp bins. Most predicted deletions are small (<1 kbp), and an increased number of deletions of lengths 300 bp and 6 kbp are visible in the histogram, corresponding to Alu and L1 deletions, respectively.

We designed a custom oligonucleotide microarray (Roche NimbleGen, 2.1 million probes) to validate putative deletions (see Section IV for a full description on array design and analyses). For this specific validation, we excluded deletion calls from the sex chromosomes, then filtered the deletion intervals that intersect (>30% over the deletion interval) with gorilla SDs. We excluded intervals with >80% repeat content and intervals <500 bp. A total of 2,755 calls (18 Mbp) were represented in our validation microarray design with ≥10 probes. An interval was considered as *validated* when the median log2 ratio of the region was beyond 1 standard deviation of the

hybridization (log2 ≈0.3). In total, we validated 1,820 deletion intervals (6.7 Mbp) corresponding to 593 genes (Supplementary Note Table 3).

Mobile Element Discovery

VariationHunter(Hormozdiari et al. 2009) was used to cluster gorilla read pairs where one end can be mapped to a repeat element consensus sequence and the other end anchored to a position on the human reference genome not flanked by a common repeat. Using a modified version of mrsFAST (http://mrsfast.sourforge.net), we remapped 91 million one-end anchored and 58.8 million discordant read pairs to both the reference genome (NCBI36) and our consensus repeat library. To facilitate direct comparison with our other results, we then converted the predicted loci to coordinates in NCBI35 using the LiftOver tool. As a postprocessing step, we removed any insertion calls that lie within 50 bp of annotated repeat elements in the human reference genome and any calls supported by less than four read pairs. This process yielded 263 PTERV1, 4272 Alu, 325 SVA, 299 L1, and 716 subterminal tandem repeat insertions. We found no evidence of PTERV2 insertions in the gorilla genome. Experimental validation was carried out on 30 selected new full-length Alu insertions (300 bp). Flanking, 150 bp regions were selected free of duplication and repetitive elements and oligonucleotide primers were designed for a PCR assay. 27/30 sites confirmed a complete novel Alu insertion in the gorilla genome but not in the human genome (450 bp amplification products). The three remaining were dimorphic (alu9, 12 and 27; Supplementary Note Table 4) with both a 150 (null allele) and 450 (insertion allele) bp PCR product being observed (Figure S5). These correspond to heterozygous polymorphisms and may have arisen as a result of an ancient polymorphism or lineage sorting. We assessed our false negative rate by analyzing 20 gorilla BAC clones sequenced in entirety using capillary technology (4.2 Mbp) and performed a direct comparison with the human reference genome. We found 19 new Alu insertions in the gorilla genome using this method (Supplementary Note Table 5). We observed that VariationHunter accurately predicted 8/19 of these Alu elements in the correct location and also assigned the correct subclass (AluY, AluSx, etc.). Close inspection of the new Alu elements missed by VariationHunter (n = 11) revealed that 7/11 Alu elements lie within 50 bp of another Alu repeat annotated in the human genome and, thus, were filtered out. Of the 11, two represent clustered Alu elements located within 100 bp of each other and one is spanned by two new SVA elements, preventing the mapping of read pairs as required by the VariationHunter algorithm. 1/11 Alu element (275 bp) was potentially a false negative prediction within unique sequence. We note, however, that the BAC libraries were generated from the genome of a different gorilla individual, which may account for some of the non-overlapping predictions by the two methods.

Supplementary Note Table 4. GGO alu new insertions validation assay

	Chr	Begin	End	Primer F	Primer R	Ta	HSA Expected size	validation
alu1	chr8	87764900	87764901	GCTGAAACTCAGCAAGAGAACTG	GGGCAGTGACCTAGTCAGTATA/	60-58	142	Yes
alu2	chr2	119131492	119131493	GATTTCAAGAAGTTTCTCAATGTTTT	CACCACACCTACTGGCAAAC	58-59	142	Yes
alu3	chr8	4039502	4039503	TAATGCCAGGAAGCATCTCA	TTGCAAGAAAATGTGGGAGA	59-59	140	Yes
alu4	chr14	57219874	57219875	CTGGACAAGTTAAGAAAAATGCAA	TGCTATGATTGAAGGGGAAAA	59-59	140	Yes
alu5	chr5	89867025	89867026	GCACTCAAATGCATTGCTAAA	GCGCAGACTGCCTTAACTTT	58-59	105	Yes
alu6	chr15	37216517	37216518	GACGTTTCTTTCCTCTCATCTG	GGAAAAGCTTTAGGAAGAAGGA	58-58	140	Yes
alu7	chr2	51078404	51078405	TGATCTCAAGCAACTTTTCTTTTC	GGTACCATGGTGACTAGTTTAAG	59-59	141	Yes
alu8	chr10	20259246	20259247	TGAGGATGATATGTCTCAGTTGG	CCTTATTAGCGGTTTGCAGAG	59-59	143	Yes
alu9	chr8	88872698	88872699	GGAAGAAGTTAGGAATGGAATAAAC	TATTTACTGTCAACAGAAGAAGC	59-60	140	Yes (DBs)
alu10	chr1	174361077	174361078	TGGAGATGATGACCTAGAATCTG	CATGCATCTGCATTGACAGG	58-61	142	Yes
alu11	chr4	20577805	20577806	CAAATAGAACATGATCCCTGTGT	CAACAGATATTTGTAGAATGGAA	58-59	122	Yes
alu12	chr11	104431741	104431742	TGACTTTGATTACCTGAGTCTCTTTT	GTTCTTGCTCTGGGCTCTTG	58-60	141	Yes (DBs)
alu13	chr17	11736134	11736136	TGAACATCACTTTCCACAGCA	GAAATGGTGGGGGCAGAT	60-60	100	Yes
alu14	chr2	215292559	215292561	TATTTAAGTTCCACATACAGCCAGA	ACTAATGTCCCCAGCTGCAC	59-60	140	Yes
alu15	chr14	37959026	37959028	TGAAAGGATTTGAAAGAAACAAAA	GTGGGGGTAAAATCCCACTG	59-61	140	Yes
alu16	chr7	111691854	111691856	CCTCCACTATCATTATTTTATTAGCAA	TCTCAGGTAAAATGAGAAAAAT/	58-58	120	Yes
alu17	chr7	102961679	102961682	TTGTGACAGAAGGGTGAAAAA	TGTCCCTAATACTCTTCTTGTACC	58-60	147	Yes
alu18	chr1	201346252	201346255	GAGAATGGGCTGGGTCAGT	TGTCGACCTAAACCCTACTGTG	60-59	104	Yes
alu19	chr13	94107951	94107954	AAATGGAAGTGGCCTAGAATGA	CAAACAAATGGTGCACTCACA	59-60	121	Yes
alu20	chr5	130639400	130639403	AGCTCGATGGTATCCTGTGC	TTTGAAAGGAGAGTAACTAGGC.	60-59	144	Yes
alu21	chr4	173705388	173705391	TTATTAGCCTCTGTACTGCTTTGTG	TTCATGAGATTAGAGCTATGCAA	59-58	123	Yes
alu22	chr3	133706369	133706373	CCAGATGCCTAAGCAGTCATAA	AAACAAAACCTAAGTCTTTAAG/	59-59	147	Yes
alu23	chr14	77398271	77398281	AGCAGTTGTGTGTGTGTCTGTG	AGAGGCTGGGCTCCTGAT	59-59	144	Yes
alu24	chr4	183998758	183998768	CTGGGGCTTATGAAACCAAA	GGGAGAGTTTAAAAAGGACAAA	59-59	128	Yes
alu25	chr7	111014438	111014448	CACGTCTCTCTTATTATAACCAAAAT	GCTTGTTTTGTTCTCAAAGCTG	58-59	157	Yes
alu26	chr3	4099854	4099864	ACATGCCCAGACTGAAAAGG	ACTGTCCGTGAGGTTCCAAT	60-59	156	Yes
alu27	chr2	227963128	227963138	GATCTCAATGACAGCCTAACTGG	CAGGATCCCCTGGAGGAC	60-60	151	Yes (DBs)
alu28	chr11	113598200	113598211	CACTCCTCTCAATGACTTTTGC	TTAACTTATATTTTGGGGGTTGG	58-58	153	Yes
alu29	chr17	24426955	24426966	CCACCTTTCCCGTACTCCTC	GTGGGGCATCAGTAGGAGAG	60-59	117	Yes
alu30	chr11	100828667	100828678	TTTTCAGGCAGACACTACATGG	TGGTGATTAAACTCCATTGTTCA	60-59	148	Yes

Notes. Ta, Annealing Temperature; DBs, Double Bands

Supplementary Note Table 5. Comparison between human and gorilla alu insertion sequences

	query_seq	query_b	query_e	posit_b	posit_e	Size_HSA	repeat	repeat	begin	end	Size_GGO_BAC	GGO_BAC	GGO_ins_appr_breakpoint	VariationHunter	VariationHunter Exp	Alu
				87007	87144	137	FLAM_C	FLAM_C	68743	68888	145	AC 145852_3				
1	chr7	116074085	116312058					AluY	69023	69298	275		116143108	MISSED	unknown	yes
				87182	87279	97	MLT1J									
				51129	51157	28	(T)n									
2	chr4	77947356	78214186					AluSg	29050	29359	309	AC 144988_2	77976406	FOUND	Sup:4, correct type	yes
				51452	51583	131	Charlie20a	Charlie20a	29626	29754	128					
				67745	68632	887	L1MC1	L1MC1	45692	46199	507	AC 144988_2				
								SVA_D	46212	46618	406					
3	chr4	77947356	78214186					AluY	46619	46919	300		77993975	MISSED	Neighboring SVAs	yes
								SVA_D	46920	47331	411					
								L1MC1	47339	47724	385					
				68650	68694	44	MER4E	MER4E	47742	47786	44					
4	chr16	16852636	17174710	239118	239423	305	AluSx	AluSx	183999	184302	303					
								Alu	184311	184360	49	AC 145176_2	184311	MISSED	Truncation / Close HG17 Alu	yes
				239424	239635	211	THE1B	THE1B	184358	184577	219					
				115577	115708	131	MIRb	MIRb	80895	81206	311					
5	chr16	23169206	23418816					AluY	81698	81823	125	AC 145240_3	23250904	FOUND	Sup5, correct type	yes
				115824	116072	248	AluJb	AluJb	81866	82121	255					
				111804	111954	150	MIR3	MIR3	84007	84140	133					
6	chr7	116731432	117008511					AluY	84162	84470	308	AC 145357_3	116815594	FOUND	Sup=12, correct type	yes
				112151	112276	125	MIRc	MIRc	84677	84802	125					
				159895	159988	93	AT_rich	AT_rich	141250	141328	78	AC 145405_3				
7	chr7	116558786	116774172					AluYd8	141535	141827	292		116700321	FOUND	Sup:8, correct type	yes
				161661	162116	455	L2b									
				137281	137405	124	(TTATA)n	(GAAAA)n	110852	110875	23	AC 145496_3				
8	chr7	116962610	117244522					AluSx	110876	111045	169		117073486	MISSED	Truncation	yes
				138034	138145	111	MIR3	(TA)n	111221	111348	127					
				184990	185318	328	AluY	AluY	161680	162014	334	AC 153309_3				
9	chr20	42985661	43249054	185319	185369	50	MER113									
								AluSq	162145	162365	220		162145	MISSED	Close HG17 Alu	yes
				185370	185671	301	AluSq	AluSq	162416	162715	299					
				87454	87755	301	AluSx	AluSx	58466	58772	306	AC 155197_2				
10	chr19	56136309	56472010					AluSq/x	58784	58915	131		56195093	MISSED	Close HG17 Alu	yes
				88953	89080	127	C-rich	C-rich	60137	60240	103					
				70311	70437	126	FLAM_A	FLAM_A	20029	20155	126	AC 167289_3				
				70512	70549	37	(GA)n									
11	chr2	26956765	27282732					AluJb	20298	20605	307		26977063	FOUND	Sup:11, correct type	yes
				71543	71620	77	L2b	L2c	21522	21644	122					
				71750	71871	121	(TTA)n	(TTA)n	21894	22015	121					
12	chr2	26956765	27282732	91361	91434	73	MIRc	MIRc	41185	41258	73	AC 167289_3				
								AluY	42165	42278	113		42165	MISSED	Truncation	yes
								MIR	43191	43242	51					
				92671	92889	218	L2	L2	43554	43645	91					
13	chr2	26956765	27282732	268198	268507	309	AluSg	AluSg	200427	200736	309					
14								AluSg	200756	201058	302	AC 167289_3	200756	MISSED	Close HG17 Alu	yes
								AluSg	201425	201734	309		201425	MISSED	Close HG17 Alu	yes
				268527	268823	296	AluJo	AluJo	201754	202050	296					
15	chr2	26793841	27060094	123568	123828	260	Tigger3b	Tigger3b	104495	104755	260	AC 167290_2				
								Alu	104853	104910	57		104853	MISSED	Truncation / Close HG17 Alu	yes
				123941	124142	201	AluSc	AluSc	104908	105109	201					
16	chr2	26793841	27060094	162893	163193	300	AluSx	AluSx	144400	144700	300	AC 167290_2				
17								AluY	144989	145105	116		144989	MISSED	Truncation / 2 Alus	yes
								AluY	145207	145350	143		145207	MISSED	Truncation / 2 Alus	yes
				163482	163656	174	FRAM	FRAM	145639	145813	174					
18	chr2	26793841	27060094	233235	233361	126	FLAM_A	FLAM_A	215220	215346	126	AC 167290_2				
				233436	233473	37	(GA)n									
								AluJb	215489	215796	307		215489	FOUND	Sup:11, correct type	yes
				234467	234544	77	L2b	L2c	216713	216835	122					
19	chr20	3715803	3972945	42100	42400	300	AluSx	AluSx	24691	24994	303	AC 169133_2				
				42401	42735	334	L1MC1									
								AluSg/x	24995	25020	25		24995	FOUND	Sup:13, correct type	yes
				42755	43054	299	AluJo	AluJo	25165	25469	304					
20	chr20	3715803	3972945	212318	212630	312	AluSp	AluSp	189804	190142	338	AC 169133_2				
								AluSp	190191	190489	298		190191	MISSED	Close HG17 Alu	yes
				212681	213038	357	L1MB8	L1MB8	190938	191241	303					
21	chr20	3350170	3603571	198748	199021	273	AluSx	AluSg/x	169999	170170	171	AC 171375_2				
								AluY	170464	170769	305		170464	FOUND	Sup:11, correct type	yes
				199356	199576	220	MER30	MER30	170859	171079	220					

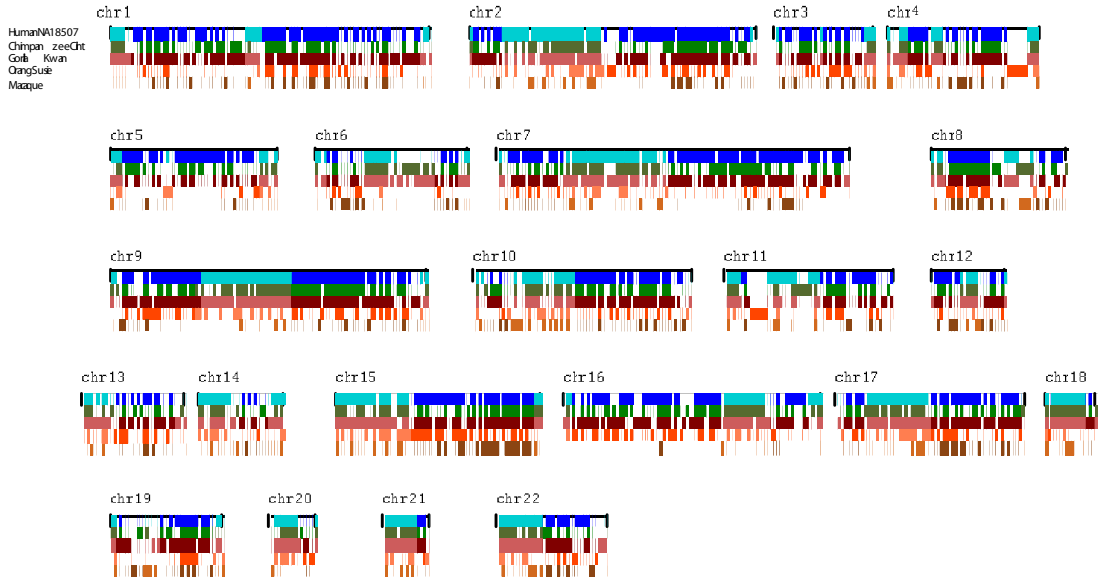
Segmental Duplications (SDs)

We used the WSSD method to identify regions >20 kbp in length with a significant excess of read-depth within 5 kbp overlapping windows(Bailey et al. 2002). We applied different correction methods specific to next-generation sequencing data as previously described(Alkan et al. 2009). In brief, we mapped the gorilla genome sequences to a repeatmasked version of the human genome (NCBI35) to detect regions with excess of depth-of-coverage(Marques-Bonet et al. 2009). After eliminating sequence duplicates, we constructed an SD map based on 1.5 billion sequences (effective coverage 9.6X). We initially detected 112 Mbp (99 Mbp >20 kbp) of duplication in the Kwan genome. This is slightly larger than what has been detected in previous genomes (Supplementary Note Table 6). We detected 100 Mbp (>20 kbp) in NA18507(Alkan et al. 2009), 77 Mbp (>20 kbp) in chimpanzee (Sanger sequencing,

sequences were cropped in 76 bp to obtain comparable sequences to Illumina reads (Illuminazation)), and 33 Mbp (>20 kbp) in orangutan (Sanger sequencing) (Supplementary Note Figure 3).

Supplementary Note Table 6. Duplication map of human (NA18507), chimpanzee (Clint) and Gorilla (Kwan)

	All	>20 kb
NA18507 WSSD (no SEX chr)	109,704 Kb	100,772 Kb
PTR_Ill WSSD (no SEX chr)	83,248 Kb	77,264 Kb
Kwan WSSD (no SEX chr)	112,408 Kb	99,540 Kb



Supplementary Note Figure 3. Primate comparative SD map. SDs (>95% sequence identity; >10 kbp) from each chromosome were extracted and concatenated based on human chromosome coordinates. Each line represents a different primate species where interstitial SDs (dark) are distinguished from pericentromeric and subtelomeric SDs (light color). The species are color coded: human, individual NA18507 (blue); common chimpanzee, Clint (green); gorilla, Kwan, (dark red); orangutan, Susie (orange); and macaque (brown). The specific intersections for different primate genomes are shown (Supplementary Note Table 7).

Supplementary Note Table 7. Intersection of fragments of African ape and human dupmaps (WSSD based)

Shared duplications

	NA18507_Clint_Kwan	NA18507_Clint	NA18507_Kwan	Clint_Kwan
ALL	68,085	5,671	20,331	3,053
>20 kb	63,556	3,788	13,657	2,520

Specific duplications

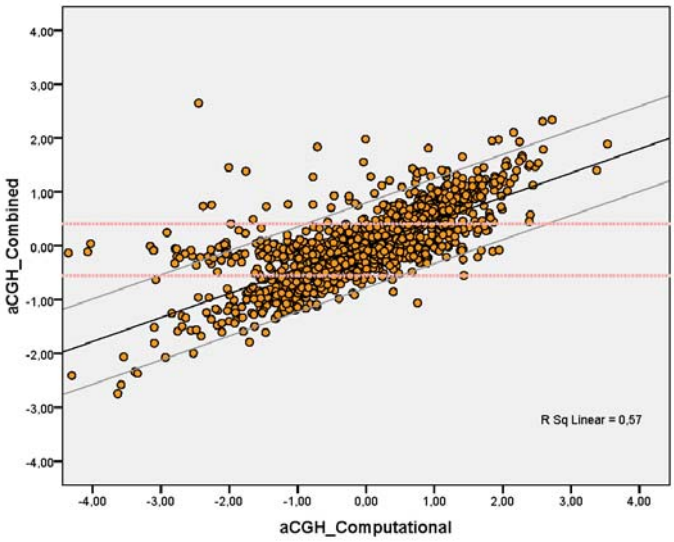
	NA18507 only	Clint only	Kwan only
ALL	15,616	6,438	20,937
>20 kb	10,639	3,916	13,482

all values in Kb

VI. ArrayCGH Validation

We performed a series of interspecific array comparative genomic hybridizations (arrayCGH) to confirm gorilla specific deletions and duplications. Two designs were employed. First, a customized oligonucleotide microarray (Roche NimbleGen, 2.1 M isothermal probes) targeted to predicted gorilla duplications and deletions. As part of this design, we also selected four regions (600 kbp) of single-copy DNA to serve as copy number not variable control regions for the analysis of the hybridizations. We initially interrogated 160 Mbp of sequence (GGOchip) with a density on average of 1.3 probes every 100 bp. Second, a standard 2.1 million standard Roche NimbleGen arrayCGH microarray with the probes evenly distributed throughout the human genome (~1 probe per kbp). Human DNA from sample NA18507 and Kwan gorilla blood were co-hybridized and dye-swap replicates were performed between the test and reference. After normalization we selected only those probes that performed reciprocally and reproducibly within dye-swaps (87% of the probes in the standard 2.1 array and 81% of the probes in the custom designed array).

To validate gorilla-specific duplications and deletions, we used a combination of two methods given the complexity of the regions: 1) a previously described segmentation algorithm(Day et al. 2007) applied to the average log2 of each probe. An interval was considered as validated if there was >50% overlap with the HMM (Hidden Markov Model) calls (based on 1 standard deviation) and 2) if the median log2 of all the probes of the region was beyond 1 standard deviation for all signals across the entire experiment (~log2 threshold = ~0.3). This threshold was selected to result in a false discovery rate of <1%(Marques-Bonet et al. 2009) based on our invariant control regions. Only sites with at least 10 probes (custom design) or 5 probes (standard 2.1 design) were analyzed. In most cases (~85%), both metrics were in agreement, but the union criteria ensured the detection of copy number differences in more complex regions where both gains and losses were occurring in close proximity. (Note, these regions are particularly problematic because they lead to a nonuniform distribution of log2 signal intensity by segmentation.) We obtained a good correlation coefficient between a computational log2 ratio (based on the estimated copy number inferred from read-depth) and the experimental log2 from the standard 2.1 arrayCGH (Supplementary Note Figure 4; R2 = 0.57).



Supplementary Note Figure 4. Correlation of computational copy number and arrayCGH log2 per site. The computational log2 was calculated with the estimated copy numbers based on read depth-of-coverage, and the arrayCGH values were obtained with the median of experimental log2 of consistent probes in both experiments, the 2.1 dye swap experiment and the 2.1 custom designed array. We denote (in pink) the median log2 threshold used for defining a region as validated.

To assess the evolution of great ape SDs (Section V), we made use of previously published inter-specific hybridization microarrays (Apechip1 and 2), which specifically included great ape SDs discovered in chimpanzee and orangutan(Marques-Bonet et al. 2009) (Supplementary Notes Table 8). Use of these microarrays combined with the gorilla customized design, facilitated designation of lineage-specific and shared duplications based among human and great ape species.

Supplementary Note Table 8. Summary of the arrayCGH hybridizations used for re-classification of segmental duplications

Specie1	Specie2	Array	Category
HSA G248	PTR Clint	Apechip	HSA/PTR
HSA G248	GGO Bahati	Apechip	HSA/GGO
HSA G248	PTR PR00238	Apechip	HSA/PTR
PTR Clint	GGO Bahati	Apechip2	PTR/GGO
HSA G248	PTR Katie	Apechip	HSA/PTR
HSA G248	GGO Kowali	Apechip	HSA/GGO
HSA G248	GGO Makari	Apechip	HSA/GGO
HSA G248	GGO Bahati	2.1	HSA/GGO
HSA ABC8	GGO Kwan	GGOchip	HSA/GGO
PTR Clint	GGO Kwan	GGOchip	PTR/GGO
HSA ABC8	GGO Kwan	2.1	HSA/GGO

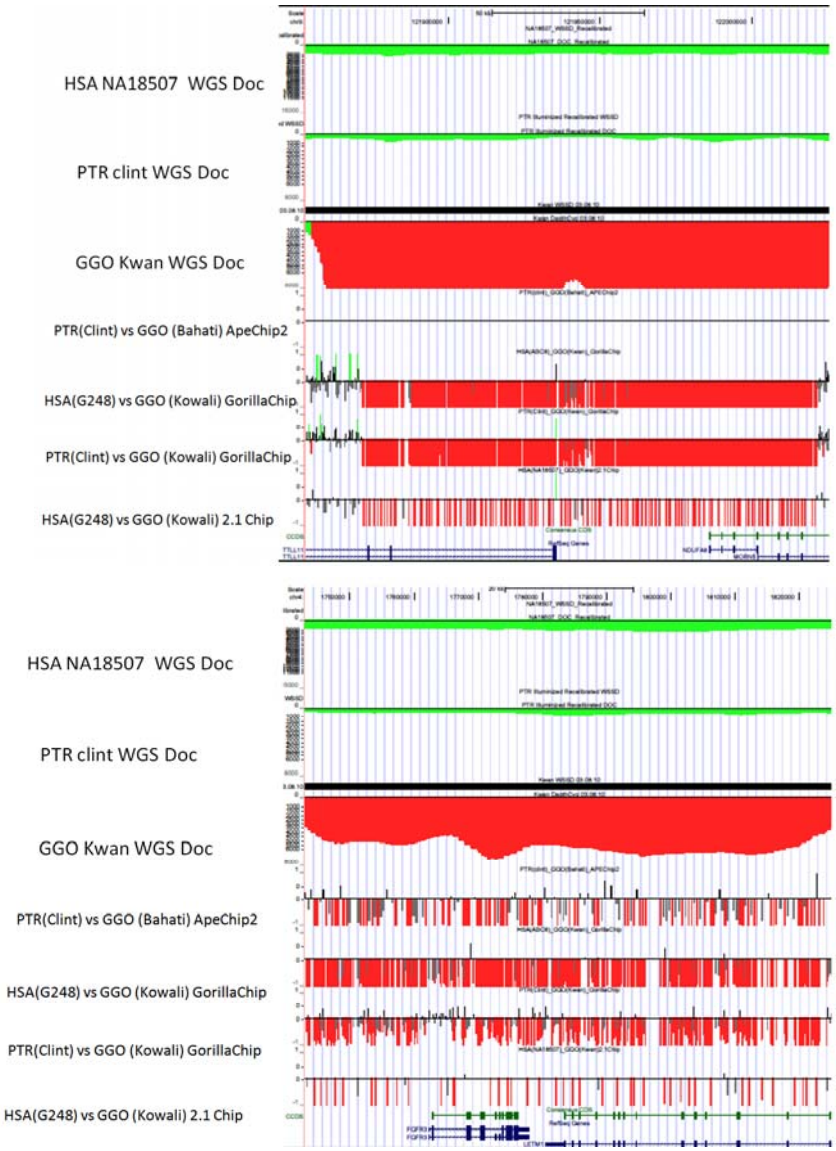
Note. Four different designs have been used to analyze the data. Two of them (Apechip and Apechip2) were previously published (Marques-Bonet et al. 2009). These two designs cover most of the ape segmental duplications (Apechip) and ape copy number variant regions greater than 20 Kbps (Apechip2). The 2.1 array is a standard oligonulcleotide design, with proves evenly distributed throughout the genome. Finally, the GGOChip is a custom design targeted to duplications and deletions in the gorilla genome.

For each pairwise species comparison (human/chimp, human/gorilla and chimp/gorilla), a global median log2 of all hybridizations involving the very same species was considered. To validate species-specific regions, we required reciprocal significance (using the aforementioned criteria). For example, to validate a duplication as gorilla-specific, it had to be independently validated in the human/gorilla hybridizations and chimp/gorilla hybridizations. The remaining categories (shared duplications) were later reclassified according to the results of arrayCGH (467 sites (28%) were reclassified). Final results are summarized in Supplementary Note Table 9.

Supplementary Note Table 9. Intersection of fragments of African ape and human dupmaps (WSSD based plus arrayCGH correction)

Class	Total length	#
HSA_only	5,807,120	97
PTR_only	1,795,699	30
GGO_only	6,813,344	88
HSA_GGO_only	6,306,531	123
HSA_PTR_only	4,201,509	95
PTR_GGO_only	3,213,472	47
HSA_PTR_GGO	73,983,219	1,047

Of the initially predicted 13.4 Mbp of gorilla-specific SDs (>20 kbp), we validated 6.81 Mbp (50.8%) with 68 genes located within the duplicated regions (23 completed and 45 partial) (Supplementary Note Figure 5 and Supplementary Note Table 10).



Supplementary Note Figure 5. Two examples of genes containing gorilla-specific duplications.

Supplementary Note Table 10. Gene list of validated gorilla-specific duplications (>20 kb)

GENE NAME	Complete/Partial	GENE ID	Protein ID	Description
NM_019899	Partial	ABCC1	NP_063954	ATPbinding cassette, subfamily C, member 1
NM_000692	Complete	ALDH1B1	NP_000683	aldehyde dehydrogenase 1B1 precursor
NM_203382	Partial	AMACR	NP_976316	alphamethylacylCoA racemase isoform 2
NM_003899	Partial	ARHGEF7	NP_003890	PAKinteracting exchange factor beta isoform a
NM_004326	Partial	BCL9	NP_004317	Bcell CLL/lymphoma 9
NM_016561	Partial	BFAR	NP_057645	bifunctional apoptosis regulator
NM_033201	Partial	C16orf45	NP_149978	hypothetical protein LOC89927 isoform 1
NM_001113434	Partial	C17orf51	NP_001106905	hypothetical protein LOC339263
NM_030945	Complete	C1QTNF3	NP_112207	C1q and tumor necrosis factor related protein 3
NM_023073	Partial	C5orf42	NP_075561	hypothetical protein LOC65250
NM_001039803	Complete	CCRK	NP_001034892	cell cycle related kinase isoform 3
NM_007053	Complete	CD160	NP_008984	CD160 antigen
NM_172101	Partial	CD8B	NP_742099	CD8b antigen isoform 3 precursor
NM_033225	Partial	CSMD1	NP_150094	CUB and Sushi multiple domains 1
NM_145918	Complete	CTSL1	NP_666023	cathepsin L1 preproprotein
NM_001023564	Complete	CTSL3	NP_001018858	cathepsin Llike protein
NM_004938	Partial	DAPK1	NP_004929	death-associated protein kinase 1
NM_173660	Complete	DOK7	NP_775931	downstream of tyrosine kinase 7
NM_022965	Complete	FGFR3	NP_075254	fibroblast growth factor receptor 3 isoform 2
NM_002015	Partial	FOXO1	NP_002006	forkhead box O1
NM_012201	Partial	GLG1	NP_036333	golgi apparatus protein 1 isoform 1
NM_001500	Partial	GMDS	NP_001491	GDPmannose 4,6dehydratase
NM_004667	Partial	HERC2	NP_004658	hect domain and RLD 2
NM_001528	Complete	HGFAC	NP_001519	HGF activator preproprotein
NM_001077443	Partial	HNRNPC	NP_001070911	heterogeneous nuclear ribonucleoprotein C
NM_001007563	Complete	IGFBPL1	NP_001007564	insulin-like growth factor binding protein-like
NM_024773	Complete	JMJD5	NP_079049	jumonji domain containing 5 isoform 2
NM_015443	Partial	KIAA1267	NP_056258	hypothetical protein LOC284058
NM_012318	Complete	LETM1	NP_036450	leucine zipperEFhand containing transmembrane
NM_022458	Partial	LMBR1	NP_071903	limb region 1 protein
NM_002337	Complete	LRPAP1	NP_002328	low density lipoprotein receptor-related protein
NM_152900	Partial	MAGI3	NP_690864	membrane-associated guanylate kinase-related 3
NM_198469	Partial	MORN5	NP_940871	MORN repeat containing 5
NM_017520	Partial	MPHOSPH8	NP_059990	Mphase phosphoprotein 8
NM_031902	Partial	MRPS5	NP_114108	mitochondrial ribosomal protein S5
NM_022844	Partial	MYH11	NP_074035	smooth muscle myosin heavy chain 11 isoform
NM_014222	Complete	NDUFA8	NP_055037	NADH dehydrogenase (ubiquinone) 1 alpha
NM_000267	Partial	NF1	NP_000258	neurofibromin isoform 2
NM_138400	Complete	NOM1	NP_612409	nucleolar protein with MIF4G domain 1
NM_145080	Complete	NSMCE1	NP_659547	nonSMC element 1 homolog
NM_000275	Partial	OCA2	NP_000266	oculocutaneous albinism II
NM_001004693	Complete	OR2T10	NP_001004693	olfactory receptor, family 2, subfamily T,
NM_001001964	Partial	OR2T11	NP_001001964	olfactory receptor, family 2, subfamily T,
NM_002582	Partial	PARN	NP_002573	poly(A)specific ribonuclease (deadenylation
NM_002614	Partial	PDZK1	NP_002605	PDZ domain containing 1
NM_152309	Partial	PIK3AP1	NP_689522	phosphoinositide3kinase adaptor protein 1
NM_152666	Partial	PLD5	NP_689879	phospholipase D family, member 5
NM_005729	Complete	PPIF	NP_005720	peptidylprolyl isomerase F precursor
NM_002926	Partial	RGS12	NP_002917	regulator of Gprotein signaling 12 isoform 2
NM_014455	Partial	RNF115	NP_055270	Rabring 7
NM_014649	Partial	SAFB2	NP_055464	scaffold attachment factor B2
NM_147156	Partial	SGMS1	NP_671512	sphingomyelin synthase 1
NM_025181	Partial	SLC35F5	NP_079457	solute carrier family 35, member F5
NM_003498	Partial	SNN	NP_003489	Stannin
NM_003105	Partial	SORL1	NP_003096	sortilinrelated receptor containing LDLR class
NM_031272	Partial	TEX14	NP_112562	testis expressed sequence 14 isoform b
NM_194252	Partial	TTLL11	NP_919228	tubulin tyrosine ligase-like family, member 11
NM_015914	Complete	TXNDC11	NP_056998	thioredoxin domain containing 11
NM_017811	Partial	UBE2R2	NP_060281	ubiquitin-conjugating enzyme UBC3B
NM_032582	Partial	USP32	NP_115971	ubiquitin specific protease 32
NM_018034	Partial	WDR70	NP_060504	WD repeat domain 70
NM_033131	Complete	WNT3A	NP_149122	wingless-type MMTV integration site family,
NM_015171	Partial	XPO6	NP_055986	exportin 6
NM_005431	Complete	XRCC2	NP_005422	Xray repair cross complementing protein 2
NM_014153	Partial	ZC3H7A	NP_054872	zinc finger CCCHtype containing 7A
NM_152604	Complete	ZNF383	NP_689817	zinc finger protein 383
NM_152484	Partial	ZNF569	NP_689697	zinc finger protein 569
NM_003426	Complete	ZNF74	NP_003417	zinc finger protein 74

On NA18507, we tested 10.64 Mbp and validated 5.81 of them (54.6%). They encompassed 42 genes (19 complete and 23 partially duplicated) (Supplementary Note Table 11).

Supplementary Note Table 11. Gene list of human-specific (not chimp, not gorilla) duplications

GENE NAME	Complete/Partial	GENE ID	Protein ID	Description
NM_002285	Partial	AFF3	NP_002276	AF4/FMR2 family, member 3 isoform 1
NM_005435	Partial	ARHGEF5	NP_005426	rho guanine nucleotide exchange factor 5
NM_172101	Partial	CD8B	NP_742099	CD8b antigen isoform 3 precursor
NM_004061	Partial	CDH12	NP_004052	cadherin 12, type 2 preproprotein
NM_032545	Complete	CFC1	NP_115934	cryptic
NM_001079530	Complete	CFC1B	NP_001072998	cripto, FRL1, cryptic family 1B
NM_139320	Partial	CHRFAM7A	NP_647536	CHRNA7FAM7A fusion isoform 1
NM_000746	Partial	CHRNA7	NP_000737	cholinergic receptor, nicotinic, alpha 7
NM_018180	Partial	DHX32	NP_060650	DEAD/H (AspGluAlaAsp/His) box polypeptide 32
NM_001039350	Partial	DPP6	NP_001034439	dipeptidylpeptidase 6 isoform 3
NM_020185	Complete	DUSP22	NP_064570	dual specificity phosphatase 22
NM_014719	Partial	FAM115A	NP_055534	hypothetical protein LOC9747
NM_173678	Partial	FAM115C	NP_775949	hypothetical protein LOC285966 A
NM_001100910	Partial	FAM72B	NP_001094380	hypothetical protein LOC653820
NM_000566	Complete	FCGR1A	NP_000557	Fc fragment of IgG, high affinity Ia, receptor
NM_001017986	Complete	FCGR1B	NP_001017986	Fc fragment of IgG, high affinity Ib, receptor
NM_001003702	Complete	FLJ43692	NP_001003702	hypothetical protein LOC445328
NM_152428	Partial	FRMPD2	NP_689641	FERM and PDZ domain containing 2 isoform 1
NM_001042524	Complete	FRMPD2L1	NP_001035989	FERM and PDZ domain containing 2 like 1 isoform
NM_014696	Partial	GPRIN2	NP_055511	G protein regulated inducer of neurite outgrowth
NM_001515	Complete	GTF2H2	NP_001506	general transcription factor IIH, polypeptide 2,
NM_001098729	Complete	GTF2H2B	NP_001092199	general transcription factor IIH, polypeptide
NM_001098728	Complete	GTF2H2C	NP_001092198	general transcription factor IIH, polypeptide
NM_033000	Partial	GTF2I	NP_127493	general transcription factor II, i isoform 2
NM_173537	Complete	GTF2IRD2	NP_775808	GTF2I repeat domain containing 2
NM_001003795	Complete	GTF2IRD2B	NP_001003795	GTF2I repeat domain containing 2B
NM_032821	Partial	HYDIN	NP_116210	hydrocephalus inducing isoform a
NM_022892	Partial	NAIP	NP_075043	NLR family, apoptosis inhibitory protein isoform
NM_000265	Complete	NCF1	NP_000256	neutrophil cytosolic factor 1
NM_006310	Partial	NPEPPS	NP_006301	aminopeptidase puromycin sensitive
NM_002538	Partial	OCLN	NP_002529	occludin
NM_001001802	Complete	OR2A42	NP_001001802	olfactory receptor, family 2, subfamily A,
NM_001005328	Complete	OR2A7	NP_001005328	olfactory receptor, family 2, subfamily A,
NM_130901	Partial	OTUD7A	NP_570971	OTU domain containing 7A
NM_001002811	Partial	PDE4DIP	NP_001002811	phosphodiesterase 4D interacting protein isoform
NM_001042363	Partial	PTPN20B	NP_001035822	protein tyrosine phosphatase, nonreceptor type
NM_022978	Complete	SERF1B	NP_075267	small EDRKrich factor 1B, centromeric
NM_000344	Complete	SMN1	NP_000335	survival of motor neuron 1, telomeric isoform d
NM_022875	Complete	SMN2	NP_075013	survival of motor neuron 2, centromeric isoform
NM_001042758	Partial	SRGAP2	NP_001036223	SLITROBO Rho GTPase activating protein 2
NM_181519	Partial	SYT15	NP_852660	synaptotagmin XV isoform b
NM_001039397	Complete	TBC1D28	NP_001034486	TBC1 domain family, member 28

Copy Number Correction

Mapping the Illumina WGS (Whole Genome Shotgun) reads against the human reference genome to detect and estimate the amount of duplications introduces a potential bias since nonhuman duplications are represented as unique loci in the genome. To correct for this, we used the actual depth-of-coverage to estimate nonhuman SD copy number as described previously(Marques-Bonet et al., 2009) (Supplementary Note Table 12).

Supplementary Note Table 12. Copy number correction on specific segmental duplications

Human (NA18507) specific			PTR (Clint) specific		GGO (Kwan) specific	
chr	**rgtotal**	**cpyTotal**	**rgtotal**	**cpyTotal**	**rgtotal**	**cpyTotal**
chr1	778,062	1,147,002	265,540	695,076	1,071,935	2,268,188
chr2	1,393,015	1,326,364	208,159	582,573	274,320	1,801,311
chr3					41,473	84,970
chr4					280,738	1,408,369
chr5	860,000	683,583	26,603	72,038	216,924	553,616
chr6	129,814	275,491			200,526	545,812
chr7	899,515	942,416	263,203	786,409	407,441	1,016,024
chr8			240,007	590,085	123,970	535,835
chr9	26,640	20,222	78,696	345,948	860,462	4,549,642
chr10	499,575	600,872	58,912	611,738	114,811	254,635
chr11	69,574	64,866	64,524	188,390	220,368	689,862
chr12						
chr13					631,708	1,683,282
chr14	20,528	21,686	234,553	672,674	59,242	147,820
chr15	495,930	463,545	21,028	96,442	337,593	1,116,148
chr16	512,464	902,812	225,113	734,161	1,156,078	3,830,440
chr17	121,997	109,261	63,575	444,007	339,267	927,781
chr18						
chr19					152,962	553,016
chr20						
chr21						
chr22			45,786	132,577	323,526	1,681,790
TOTAL	**5,807,114**	**6,558,120**	**1,795,699**	**5,952,118**	**6,813,344**	**23,648,541**

We parsimoniously assigned duplicated basepairs to each branch of the human-ape phylogeny based on shared and lineage-specific duplicated basepairs from a five-way primate genome comparison. Among the 63 Mbp of duplications shared among human/chimpanzee and gorilla, we determined that 21 Mbp are also shared with orangutan and, of these, only 6 Mbp are shared with MMU (Main Text Figure 2a and Supplementary Note Table 13).

Supplementary Note Table 13. Estimation of rates of duplication in great ape evolution

	Duplications*	Million years	Rate of dups (Mb/Myr)
HSA	**6,558,120**	6,000,000	*1.09*
PTR	**5,952,118**	6,000,000	*0.99*
GGO	**23,648,541**	8,000,000	*2.96*
HSA/PTR only	4,201,509	2,000,000	*2.1*
HSA/PTR/GGO only	41,689,362	6,000,000	*6.95*
HSA/PTR/GGO/PPY only	16,134,998	12,000,000	*1.34*
HSA/PTR/GGO/PPY/MMU	5,640,750		

*bold = copy number corrected

We applied maximum likelihood methods developed in Marques-Bonet et al.(Marques-Bonet et al. 2009) to estimate the rates of duplication in each branch of the African great ape phylogeny and test whether the rate of

accumulation of SD in the gorilla branch was significantly different than in (1) the branches of humans and chimpanzees, (2) the common ancestor of these two species, and (3) the common ancestor of all the African great apes.

A simple maximum likelihood model based on a Poisson rate of accumulation of duplications per time unit and on a 20% homoplasy was used. To perform every test, we first obtained maximum-likelihood estimates for two different models. The simplest one assumes a single rate of accumulation in all tested branches, while the other assumes that the gorilla branch has its own rate. Afterwards, we performed a likelihood-ratio test between the two models. Every test was performed four times, considering two units of duplication accumulation (number of SD regions and number of SD Mbp) and two different time units (Myr and number of substitutions per kbp in the corresponding branch). Supplementary Note Table 14 shows all of the rate estimates and the p-values of every test we performed.

Supplementary Note Table 14. Rate estimates and p-values

Time Unit: Myrs			
Duplication unit: SD regions. Rates expressed in SD Regions/Myrs			
	Model 1 **Identical rate in all rates**	**Model 2** **Two rates: λ_{GGO} for gorilla and λ_{Rest} rest of branches**	**p-values of LRT**
(Test 1) GGO against HSA and PTR	λ = 13.10 SDs/Myrs	λ_{Rest} = 9.75 λ_{GGO} = 18.12	$5.6*10^{-7}$
(Test 2) GGO against common ancestor of HSA_PPT	λ = 22.70	λ_{Rest} = 41.00 λ_{GGO} = 18.12	$1.8*10^{-8}$
(Test 3) GGO against common ancestor of HSA_PPT_GGO	λ = 27.07	λ_{Rest} = 39.00 λ_{GGO} = 18.12	$1.5*10^{-13}$
Duplication unit: SD Mbs. Rates expressed in Mbs/Myrs			
(Test 1) GGO against HSA and PTR	λ = 2.82 Mbs/Myrs	λ_{Rest} = 1.48 λ_{GGO} = 4.84	$<10^{-15}$
(Test 2) GGO against common ancestor of HSA_PPT	λ = 4.94	λ_{Rest} = 5.30 λ_{GGO} = 4.84	0.409
(Test 3) GGO against common ancestor of HSA_PPT_GGO	λ = 5.34	λ_{Rest} = 6.01 λ_{GGO} = 4.84	$4.25*10^{-5}$
Time Unit: Substitutions per kb			
Duplication unit: SD regions. Rates expressed in SD Regions/Substitution			
(Test 1) GGO against HSA and PTR	λ = 14.43 SDs/SubstKb	λ_{Rest} = 10.68 λ_{GGO} = 20.17	$2.8*10^{-7}$
Test 2) GGO against common ancestor of HSA_PPT	λ = 27.48	λ_{Rest} = 76.93 λ_{GGO} = 20.17	$<10^{-15}$
(Test 2) GGO against common ancestor of HSA_PPT_GGO	λ = 23.87	λ_{Rest} = 26.93 λ_{GGO} = 20.17	$5.7*10^{-3}$
Duplication unit: SD Mbs			
(Test 1) GGO against HSA and PTR	λ = 3.11 Mbs/SubstKb	λ_{Rest} = 1.62 λ_{GGO} = 5.38	$<10^{-15}$
(Test 2) GGO against common ancestor of HSA_PPT	λ = 5.97	λ_{Rest} = 9.91 λ_{GGO} = 5.38	$1.63*10^{-7}$
(Test 3) GGO against common ancestor of HSA_PPT_GGO	λ = 4.71	λ_{Rest} = 4.15 λ_{GGO} = 5.38	$3.99*10^{-4}$

These results support a “burst” of SDs near the time of the common ancestor of human and African great apes (shared with chimpanzee and gorilla), which continued along the gorilla lineage after divergence. We estimate that the rate of duplications at this time (after separation from orangutan) is 6- to 7-fold higher compared to the human and chimpanzee branches. The gorilla-specific branch shows a significant SD excess compared to the human (~2 to 4X, depending on whether time or single nucleotide divergence are used to calibrate). These data suggest a burst of SD activity before and after speciation of humans and African great apes followed by a strong deceleration in humans and chimpanzees and a milder deceleration in gorillas. Interestingly, the point-mutation slowdown is stronger in the gorilla lineage than in humans or chimpanzees(Elango et al. 2006).

References

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O et al. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41**(10): 1061-1067.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. *Science* **297**(5583): 1003-1007.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**(2): 573-580.

Bentley DR Balasubramanian S Swerdlow HP Smith GP Milton J Brown CG Hall KP Evers DJ Barnes CL Bignell HR et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**(7218): 53-59.

Carbone L, Ventura M, Tempesta S, Rocchi M, Archidiacono N. 2002. Evolutionary history of chromosome 10 in primates. *Chromosoma* **111**(4): 267-272.

Cardone MF, Alonso A, Pazienza M, Ventura M, Montemurro G, Carbone L, de Jong PJ, Stanyon R, D'Addabbo P, Archidiacono N et al. 2006. Independent centromere formation in a capricious, gene-free domain of chromosome 13q21 in Old World monkeys and pigs. *Genome Biol* **7**(10): R91.

Cardone MF, Lomiento M, Teti MG, Misceo D, Roberto R, Capozzi O, D'Addabbo P, Ventura M, Rocchi M, Archidiacono N. 2007. Evolutionary history of chromosome 11 featuring four distinct centromere repositioning events in Catarrhini. *Genomics* **90**(1): 35-43.

Chen N. 2004. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **Chapter 4**: Unit 4 10.

Day N, Hemmaplardh A, Thurman RE, Stamatoyannopoulos JA, Noble WS. 2007. Unsupervised segmentation of continuous genomic data. *Bioinformatics* **23**(11): 1424-1426.

Dutrillaux B. 1980. Chromosomal evolution of the great apes and man. *J Reprod Fertil Suppl* **Suppl 28**: 105-111.

Eder V, Ventura M, Ianigro M, Teti M, Rocchi M, Archidiacono N. 2003. Chromosome 6 phylogeny in primates and centromere repositioning. *Mol Biol Evol* **20**(9): 1506-1512.

Egozcue J, Chiarelli B. 1967. The idiogram of the lowland gorilla (Gorilla gorilla gorilla). *Folia Primatol (Basel)* **5**(3): 237-240.

Elango N, Thomas JW, Yi SV. 2006. Variable molecular clocks in hominoids. *Proc Natl Acad Sci U S A* **103**(5): 1370-1375.

Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC. 2009. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res* **19**(7): 1270-1278.

Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, Alkan C, Eichler EE, Sahinalp SC. 2010. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* **26**(12): i350-357.

Jiang Z, Hubley R, Smit A, Eichler EE. 2008. DupMasker: a tool for annotating primate segmental duplications. *Genome Res* **18**(8): 1362-1368.

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**(7191): 56-64.

Locke DP, Archidiacono N, Misceo D, Cardone MF, Deschamps S, Roe B, Rocchi M, Eichler EE. 2003. Refinement of a chimpanzee pericentric inversion breakpoint to a segmental duplication cluster. *Genome Biol* **4**(8): R50.

Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, Hillier LW, Jiang Z, Baker C, Malfavon-Borja R, Fulton LA et al. 2009. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**(7231): 877-881.

Miller DA, Firschein IL, Dev VG, Tantravahi R, Miller OJ. 1974. The gorilla karyotype: chromosome lengths and polymorphisms. *Cytogenet Cell Genet* **13**(6): 536-550.

Misceo D, Cardone MF, Carbone L, D'Addabbo P, de Jong PJ, Rocchi M, Archidiacono N. 2005. Evolutionary history of chromosome 20. *Mol Biol Evol* **22**(2): 360-366.

Misceo D, Ventura M, Eder V, Rocchi M, Archidiacono N. 2003. Human chromosome 16 conservation in primates. *Chromosome Res* **11**(4): 323-326.

Montefalcone G, Tempesta S, Rocchi M, Archidiacono N. 1999. Centromere repositioning. *Genome Res* **9**(12): 1184-1188.

Morgulis A, Gertz EM, Schaffer AA, Agarwala R. 2006. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics* **22**(2): 134-141.

Muller S, Stanyon R, Finelli P, Archidiacono N, Wienberg J. 2000. Molecular cytogenetic dissection of human chromosomes 3 and 21 evolution. *Proc Natl Acad Sci U S A* **97**(1): 206-211.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**(3): 443-453.

Newman TL, Tuzun E, Morrison VA, Hayden KE, Ventura M, McGrath SD, Rocchi M, Eichler EE. 2005. A genome-wide survey of structural variation between human and chimpanzee. *Genome Res* **15**(10): 1344-1356.

Parsons JD. 1995. Miropeats: graphical DNA sequence comparisons. *Comput Appl Biosci* **11**(6): 615-619.

Stanyon R, Rocchi M, Capozzi O, Roberto R, Misceo D, Ventura M, Cardone MF, Bigoni F, Archidiacono N. 2008. Primate chromosome evolution: ancestral karyotypes, marker order and neocentromeres. *Chromosome Res* **16**(1): 17-39.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Eichler EE. 2011. Diversity of human copy number variation and multicopy genes. *Science* **330**(6004): 641-646.

Szamalek JM, Goidts V, Cooper DN, Hameister H, Kehrer-Sawatzki H. 2006a. Characterization of the human lineage-specific pericentric inversion that distinguishes human chromosome 1 from the homologous chromosomes of the great apes. *Hum Genet* **120**(1): 126-138.

Szamalek JM, Goidts V, Searle JB, Cooper DN, Hameister H, Kehrer-Sawatzki H. 2006b. The chimpanzee-specific pericentric inversions that distinguish humans and chimpanzees have identical breakpoints in Pan troglodytes and Pan paniscus. *Genomics* **87**(1): 39-45.

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D et al. 2005. Fine-scale structural variation of the human genome. *Nat Genet* **37**(7): 727-732.

Ventura M, Mudge JM, Palumbo V, Burn S, Blennow E, Pierluigi M, Giorda R, Zuffardi O, Archidiacono N, Jackson MS et al. 2003. Neocentromeres in 15q24-26 map to duplicons which flanked an ancestral centromere in 15q25. *Genome Res* **13**(9): 2059-2068.

Ventura M, Weigl S, Carbone L, Cardone MF, Misceo D, Teti M, D'Addabbo P, Wandall A, Bjorck E, de Jong PJ et al. 2004. Recurrent sites for new centromere seeding. *Genome Res* **14**(9): 1696-1703.

Yunis JJ, Prakash O. 1982. The origin of man: a chromosomal pictorial legacy. *Science* **215**(4539): 1525-1530.