SUPPLEMENTARY NOTE

A Burst of Segmental Duplications in the African Great Ape Ancestor

Tomas Marques-Bonet[1], Jeffrey M. Kidd[1], Mario Ventura[2], Tina A. Graves[3], Ze Cheng[1], LaDeanna W. Hillier[3], Zhaoshi Jiang[1], Carl Baker[1], Ray Malfavon-Borja[1], Lucinda A. Fulton[3], Can Alkan[1], Gozde Aksay[1], Santhosh Girirajan[1], Priscillia Siswara[1], Lin Chen[1], Maria Francesca Cardone[2], Arcadi Navarro[4], Elaine R. Mardis[3], Richard K. Wilson[3], Evan E. Eichler[1]

Contents

1. Primate segmental duplication detection

All segmental duplications (SDs) were detected using the whole-genome shotgun sequence detection (WSSD) approach[1]. In brief, we aligned primate whole genome shotgun (WGS) reads using MEGABLAST v.2.2.12[2] (Supplementary Note Table 1) against a repeat-masked version of the human reference assembly (build35) and identified all regions showing an excess of WGS read-depth for 6/7 consecutive 5 kbp windows[1,3]. We considered regions as a potential, high-identity duplication when read-depth exceeded three standard deviations when compared to the mean depth for unique (nonduplicated) regions of the genome. Thresholds were determined separately for each primate WGS sequence dataset using a calibrated set of copy-number defined duplicated and unique sequences obtained from complete sequencing of BAC-based clones[1,3].

We excluded all common repeats with less than 10% sequence divergence from their consensus (RepeatMasker[4]) as well as primate-specific L1P and satellite repeat sequences. Since our goal is to focus only on sequences that arose as a result of segmental duplication, we also excluded regions post-alignment that consisted of >85% common repeats and >75% tandem repetitive DNA[5] such as VNTRs.

We selected only those reads with the following criteria: >200 bp of high quality sequence (20 for human, 27 for chimp, 30 for orangutan and 27 for macaque) and sequence identity >94% for the human and great-ape alignments and up to 88% for macaque-human alignments. Based on neutral estimates of sequence divergence, a threshold of >94% should capture all duplications that arose since the divergence of Old World and hominid lineages. Because of the potential for accelerated rates of single-basepair substitution in duplicated sequences when compared to unique sequence[6], we conservatively lowered the sequence-identity threshold for macaque-human alignments in order to capture more potentially divergent duplications. We also repeated the analysis separately against the macaque genome (using the same 94% sequence identity threshold) and obtained comparable results (see below).

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.

Since all reads were mapped to the human genome assembly, there is a potential bias that might favor an enrichment of human duplicated sequences if some nonhuman primate duplicated sequences were missing from the human genome reference. To correct for this ascertainment bias, we examined sequence contigs from the chimpanzee (panTro2)[7] and orangutan (ponAbe1, http://genome.wustl.edu/genome.cgi?GENOME=Pongo%20abelii) assemblies that did not align to the human genome. This set consisted of 623 chimpanzee sequence contigs (517 kbp) and 506 orangutan sequence contigs (1.4 Mbp). We analyzed these contigs using the WSSD method within the context of their respective assemblies and estimated the amount of duplicated sequence. These contigs contributed negligibly to missed duplications adding an estimated 110 kb of predicted chimpanzee and 254 kb of orangutan duplicated sequence.

Based on our analysis, we identified a total of ~146 Mbp of human duplications, ~90 Mbp in chimpanzee, ~86 Mbp in orangutan and ~54 Mbp in macaque. We grouped all duplicated sequences into shared or lineage-specific categories based on unique or duplicated status within each species (Fig. S1). It is important to note that 'shared', by this definition, applies both to duplications that occurred in the common ancestor of a given group of species as well as duplications affecting the same sequence but occurring independently. We merged duplications into larger duplication regions if two duplications of the same classification mapped within 10 bp of one another. We also limited subsequent analyses to duplications >20 kbp because our previous analysis of

chimpanzee duplications using the WSSD approach showed excellent sensitivity and specificity at this length threshold (with a false positive rate of 1.4% and false negative detection rate of 6.5%). Based on these thresholds and processing, we retained ~58 Mbp, ~60 Mbp, ~28 Mbp and ~15 Mbp of human, chimpanzee, orangutan and macaque duplications for further analysis (Supplementary Note Table 2). Duplication content of the Y chromosome was not considered since female genomes were sequenced for the macaque and orangutan genome projects. For each category, we corrected for copy number by estimating the total number of basepairs that would be found within each genome based on the depth of WGS read alignment (see section 1.2).

Supplementary Note Table 2. Distribution of primate SDs by category.

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**

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**

Duplication content for all categories of SDs is shown in basepairs. "Total bp" refers to the total number of basepairs identified by WSSD analysis using human genome reference (build35). "Total bp (>20 kbp)" subselects those SD intervals greater than 20 kb.

1.1 Macaque SD analysis

The detection of macaque SDs using the comparative WSSD approach (macaque WGS reads mapped against the human genome) was complicated by: i) the greater evolutionary distance between human and macaque (i.e. 93.54% identity in aligned nucleotides between human and macaque[8] and ii) the large fraction of reads (40%) that failed to map to the human genome assembly. Therefore, we re-analyzed the macaque assembly using a self-self WSSD based approach (macaque WGS sequence reads aligned to the macaque genome assembly, rheMac2)[8]. Duplication intervals were identified as previously described[1] and shared versus lineage-specific duplications when compared to the human genome were distinguished using DupMasker[9]. In brief, DupMasker detects nonredundant human duplicons, and hence, all macaque duplications not "masked" by DupMasker were defined as macaque-specific SDs. We detected ~14 Mb of putative macaque SDs by WSSD (~4 Mbp >20 kbp) plus an additional 29 Mbp of duplication (<94%) using a whole-genome assembly comparison[8]. Including duplicated sequences with less than 94% sequence identity that did not intersect with WSSD, we detected a total of 18.2 Mbp. We corrected for copy number of collapsed duplications within the assembly by accounting for the number of copies represented in the assembly and the read-depth of WGS sequence[3].

This amount of macaque duplication differs slightly from previously reported[8] since: i) we used a threshold of 20 kbp in length for consistency in both analyses (using human reference assembly and the macaque reference assembly while a 10 kbp threshold was used in the previous work); ii) we based our method of detection primarily on WSSD adding the sequence detected by Whole Genome Assembly Comparison (WGAC) (<94% ID) (the previous study used both WSSD >10 kbp and WGAC >1 kbp; >90% ID); and iii) excluding "DupMasked" regions allowed us to differentiate MMU-specific SDs from

great-ape shared SDs. This, of course, removed ancient shared duplications that most likely had different outcomes in copy number in different lineages (such as macaque or great apes), but on the other hand, we were ensuring the specificity of our MMU-specific SDs dataset.

1.2 Copy number estimation

Mapping the WGS 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 compensate for this, we used the actual depth of coverage to estimate SD copy number (correlation with copy number r^2=0.953)[3]. In short, after eliminating common repeat elements, we computed the number of reads mapping for each position in the human genome sequence and then calculated the average number of reads for each 5 kbp window sliding every 1 kbp across the genome. We then extrapolated the copy number based on the depth of coverage for known single copy BACS. For duplicated sequences represented multiple times within the human genome, we avoided potential redundancy by correcting for the total number of copies of segmental duplication (>94% sequence identity) already within the assembly[10]. This approach was also used to compare the putative copy number differences between lineage-specific and shared SDs (Fig. S6). Using a non-redundant set of duplication subunits[11], we found that "shared" duplications had significantly more copies per genome than lineage-specific SDs (Kolmogorov-Smirnov Test P-value < $2.2e^{-16}$, Fig S6)**.** Interestingly, these more ancient duplications showed a wide distribution of genomic sequence divergences suggesting a continuum of duplication activity over long periods of evolutionary time.

1.3 Duplication map comparison of two individual human genomes

We estimated the duplication content for two other independent human genomes (Watson genome[12] and Venter[13]) in order to determine the extent of variation in duplication regions. We used 74 million (74,198,831) 454 sequence reads with a minimum Phred Q value of 20 for duplication discovery in the Watson genome and 31 million (31,861,638) Sanger capillary sequencing reads (Phred Q>=27) Venter genome. Duplications were

detected using the WSSD method and duplication boundaries and copy number were compared. After removal of all intervals composed largely of common repeats to be consistent with our analyses of the ape genomes, we identified 666 duplication intervals predicted in the Venter/HuRef genome that are greater than 20 kbp in length— of these only 4.2% (28/666) are predicted to correspond to unique sequences within Watson's genome and 6.4% (43/666) are found unique in HuRef Genome based on a similar WSSD analysis. The boundaries of the shared duplication intervals were remarkably consistent (see examples in Fig. S7) suggesting that while copy number may vary considerably within duplicated regions[14,15], the duplication status of most regions remains largely invariant.

Next, we classified each variant as copy-number variant (CNV) or "invariant" based on a comparison of each interval against structural variants from *Database* of Genomic Variants[16] and the results from Kidd et al.[17] (Supplementary Note Table 3). We find that 82% (486/595) of the shared duplications are polymorphic while only 51% (36/71) of the individual-specific duplications show evidence of copy-number variation. One possible explanation may be that many of the Venter- or Watson-specific duplications are relatively rare (specific to the family or the individual) and, as such, screens of copy-number polymorphism that did not include these individuals would fail to discover such sites as CNV (i.e. in most individuals this is unique sequence).

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*	

Although our depth-of-coverage analysis provides no information on the location of the duplication, we performed an additional analysis to assess the relationship between copy number of the duplication and the extent of copy-number polymorphism. First, we split the SDs into quartiles based on their copy number by using WGS sequence data as a surrogate for the actual copy number (Supplementary Note Table 4). Using only those SDs in which both Venter and Watson genomes are in the same quartile, we found no significant difference (Fisher exact test) in the proportion of copy-number variant SDs among the four categories.

Supplementary Note Table 4. Copy-number variation on SD copy number.

SD Copy number quartiles	**Invariant**	***%***	**CNV**	***%***	**Grand Total**
0-25%	25	*21.9*	89	*78.1*	114
25-50%	15	*16.1*	78	*83.9*	93
50-75%	14	*12.3*	100	*87.7*	114
75-100%	23	*17.3*	110	*82.7*	133
Grand Total	*77*		*377*		*454*

2. Validation of human and primate SD maps

We experimentally validated primate SDs (>20 kbp in length) using two orthogonal approaches: FISH and cross-species array comparative genomic hybridization (arrayCGH).

2.1 Fluorescence in situ hybridization (FISH)

Based on our computational predictions among the four primate genomes, we classified intervals as being lineage-specific or "ancestral" duplications based on being duplicated between two or more species. We performed a series of FISH analyses[18] for larger regions (>40 kbp) of duplication using human fosmid (WIBR2) clones as probes against metaphase preparations of transformed primate lymphoblastoid cell lines. The following cell lines were used: Coriell GM15510 (human), Coriell S006006 (chimpanzee "Clint" or Yerkes #C0471), Coriell PR01109 (Sumatran orangutan "Susie" or ISIS #71) and a macaque cell line (MMU #25311). With the exception of human, we analyzed the same individual for which the computational predictions were generated (see Fig. 1, Tables S2-S4 for FISH results). We selected 58 lineage-specific SDs (14 human-specific, 24 chimpanzee and 20 orangutan SDs) and 38 complex regions (that harbored both lineage-specific and shared duplications). We confirmed 86.4% of our lineage-specific (50/58) and shared assignments (32/37). As an orthogonal confirmation of our arrayCGH results, we also tested nine regions that were shared duplications between human and chimpanzee and were also predicted to be a shared duplication with gorilla based on arrayCGH results. (Table S4). All but one were confirmed as duplicated by FISH analysis of the gorilla metaphases (GGO (Coriell AG20600)).

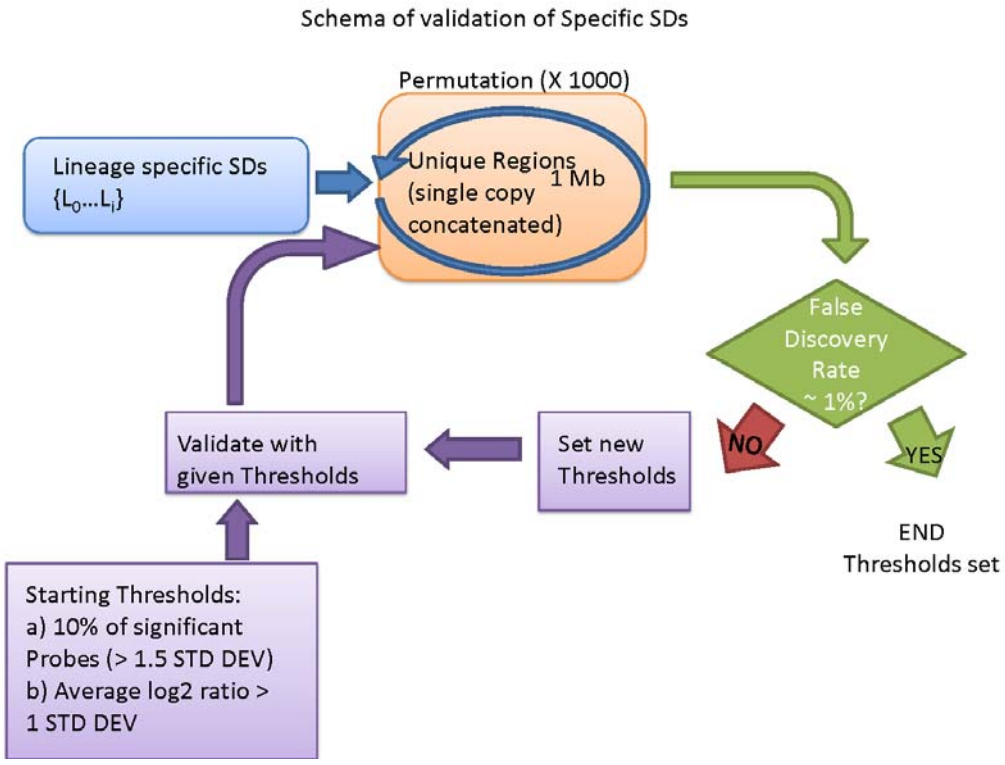
2.2 Interspecific array comparative genomic hybridization

As a second validation approach, we performed cross-species array comparative genomic hybridization to confirm lineage-specific duplications and to detect copy-number differences in shared duplications. We constructed a customized oligonucleotide microarray (NimbleGen, 385,000 isothermal probes) targeted specifically to the primate segmental duplications detected in the four species. This covered 180 Mbp of

corresponding sequence from the human genome at a density of 1 probe every 525 bp. As part of this design, we also selected 15 regions (100 kbp each) of single copy DNA to serve as copy-number invariant control regions for the analysis of the hybridizations (9 autosomal and 1 X chromosome regions). A total of 12 interspecific experiments were performed and the $\log_2$ relative hybridization intensity was calculated for each probe. These experiments included the following genomic DNA comparisons: 3 human (Coriell GM15510) vs. chimpanzee ("Clint"-Coriell S006006, Coriell PR00238 and "Katie"), 1 human (Coriell GM15510) versus orangutan ("Susie" or ISIS #71), 1 human (Coriell GM15510) versus macaque ("ID17573"), 2 human ("G248") versus bonobo ("LB502", "LB501"), 3 human (Coriell GM15510) versus gorilla ("Bahati", "Makari" and "Kobali") as well as an orangutan ("Susie") versus gorilla ("Bahati") and chimpanzee ("Clint") versus gorilla ("Bahati") experiment. With the exception of human, DNA corresponding to the primate genome being sequenced was used as part of these arrayCGH experiments. The main experiments (human DNA vs. chimpanzee, orangutan and macaque DNA) were performed with a standard replicate dye-swap experimental design (reverse labeling of test and reference samples).

To analyze the results of the hybridizations and to validate our predictions, we considered only those probes that showed a consistent result in replicate dye-swap experiments (~73% of probes). We further restricted our analysis to those regions that were greater than 20 kbp in length and contained at least 20 probes (Fig. S3, S4). We used a heuristic approach to calculate $\log_2$ thresholds of significance for each comparison. Based on our invariant control regions, we dynamically adjusted the thresholds for each hybridization to result in a false discovery rate of <1% (see Supplementary Note Fig. 1). First, we binned the $\log_2$ hybridization data for the control regions and permuted the size distribution of the lineage-specific duplications against the control regions for each individual experiment. For every permutation, we estimated two parameters: 1) the percentage of probes with a relative hybridization signal intensity 1.5 standard deviations beyond the mean hybridization signal of the control regions and 2) the average $\log_2$ ratio of the region. We repeated the analysis dynamically adjusting thresholds until a $\log_2$ threshold was discovered that ensured less than or equal to one false positive from the

control region in 1000 permutations. We considered a region validated if the average $\log_2$ signal intensity or percentage of probes exceeded the specified thresholds for that experiment. In most cases (~80%) both metrics were in agreement, but the union criteria ensured the detection of SDs in more complex regions of segmental duplications where both gains and losses in content were occurring (leading to a nonhomogenous distribution of $\log_2$ signal intensity). By these criteria we validated 89–99% of all lineage-specific segmental duplications.



Supplementary Note Fig. 1. Schematic of the algorithm used to set the validation thresholds for the arrayCGH hybridizations.

3. Nonrandom distribution of primate segmental duplications

The distribution of human and great-ape SDs is not random (Fig. S5). Previously, we showed that lineage-specific duplications tended to cluster near regions of shared duplication between chimp and human (referred to as a duplication shadowing)[3]. In order to assess the significance of this over a broader evolutionary context, we developed a simple genome randomization model to test the observation that new SDs map in close proximity to more ancient SDs (as determined by comparative primate analysis). We randomly assigned the location of the test category of duplication and measured its distance to the reference duplication set. We computed the number of times that the test SD mapped at a closer or equal distance to that observed in the dataset based on two thresholds of distance (<50 bp or <5 kbp). Gaps, telomeres and centromeres were excluded from the permutations. This test is conservative since randomly assigned test bins were allowed to overlap the reference duplication set; thus, increasing the number of times that a random bin was found closer to an older duplication.

Supplementary Note Table 5a: Duplication shadowing by category (proximity <50 bp).

Test SD1 Category	Reference SD2 Category	#SD1	#SD2	#SD1<**50** bp SD2	%SD1<50 bp SD2	Enrichment	P-Value
HSA specific	HSA/PTR	6537	2205	1342	***20.53%***	10.81	<0.001
PTR specific	HSA/PTR	1266	2205	304	***24.01%***	12.57	<0.001
PPY specific	HSA/PTR	4110	2205	0	***0.00%***	0.00	1
MMU specific	HSA/PTR	2818	2205	2	***0.07%***	0.03	1
HSA specific	HSA/PTR/PPY	6537	2008	162	***2.48%***	1.53	<0.001
PTR specific	HSA/PTR/PPY	1266	2008	29	***2.29%***	1.40	0.035
PPY specific	HSA/PTR/PPY	4110	2008	141	***3.43%***	2.16	<0.001
MMU specific	HSA/PTR/PPY	2818	2008	0	***0.00%***	0.00	1
HSA specific	HSA/PTR/PPY/MMU	6537	1044	9	***0.14%***	0.16	1
PTR specific	HSA/PTR/PPY/MMU	1266	1044	4	***0.32%***	0.36	0.99
PPY specific	HSA/PTR/PPY/MMU	4110	1044	10	***0.24%***	0.29	1
MMU specific	HSA/PTR/PPY/MMU	2818	1044	19	***0.67%***	0.73	0.931
HSA/PTR	HSA/PTR/PPY	2205	2008	995	***45.12%***	23.26	<0.001

HSA/PTR	HSA/PTR/PPY/MMU	2205	1044	46	***2.09%***	1.70	<0.001
HSA/PTR/PPY	HSA/PTR/PPY/MMU	2008	1044	605	***30.13%***	29.54	<0.001

Supplementary Note Table 5b: Duplication shadowing by category (proximity <5 kbp).

Test SD1 Category	Reference SD2 Category	#SD1	#SD2	#SD1<**5 kbp** SD2	%SD1<5 kbp SD2	Enrichment	P-Value
HSA specific	HSA/PTR	6537	2205	1374	***21.02%***	8.69	<0.001
PTR specific	HSA/PTR	1266	2205	311	***24.57%***	10.11	<0.001
PPY specific	HSA/PTR	4110	2205	26	***0.63%***	0.26	1
MMU specific	HSA/PTR	2818	2205	21	***0.75%***	0.30	1
HSA specific	HSA/PTR/PPY	6537	2008	557	***8.52%***	4.00	<0.001
PTR specific	HSA/PTR/PPY	1266	2008	94	***7.42%***	3.47	<0.001
PPY specific	HSA/PTR/PPY	4110	2008	322	***7.83%***	3.73	<0.001
MMU specific	HSA/PTR/PPY	2818	2008	23	***0.82%***	0.37	1
HSA	HSA/PTR/PPY/MMU	6537	1044	129	***1.97%***	1.76	<0.001
PTR	HSA/PTR/PPY/MMU	1266	1044	18	***1.42%***	1.26	0.2
PPY	HSA/PTR/PPY/MMU	4110	1044	55	***1.34%***	1.22	0.09
MMU	HSA/PTR/PPY/MMU	2818	1044	106	***3.76%***	3.19	<0.001
HSA/PTR	HSA/PTR/PPY	2205	2008	1032	***46.80%***	19.34	<0.001
HSA/PTR	HSA/PTR/PPY/MMU	2205	1044	252	***11.43%***	9.00	<0.001
HSA/PTR/PPY	HSA/PTR/PPY/MMU	2008	1044	633	***31.52%***	25.02	<0.001

The proximity of the test SD category (column 1) is measured with respect to the reference SD category (column 2). The total number of duplications for each category and the percent mapping at the specified distance between test and reference category are indicated. The enrichment and empirical P-value were determined based on 1000 random simulations as described above. A gradient duplication shadowing effect is observed based on the phylogenetic age of the duplication. For example, human and chimpanzee lineage-specific segmental duplications are biased near human/chimpanzee duplications, while human-chimpanzee and human-chimpanzee-orangutan duplications map

preferentially near duplications shared among all four species. These observations dispel the notion that all sequences within the primate genomes have an equal probability of duplication.

4. Gene Duplication Analyses

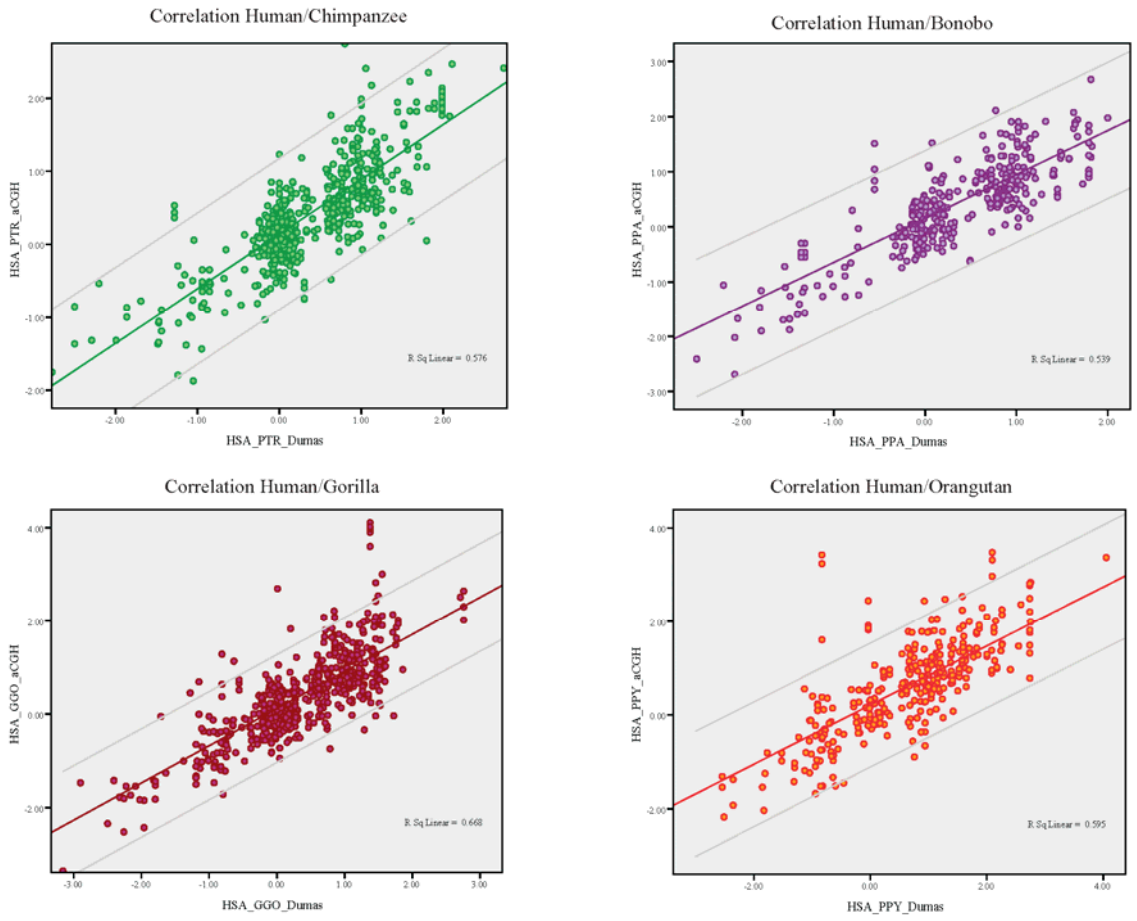
We examined the gene content of primate segmental duplications using the RefSeq gene (May08) annotation. This annotation is biased toward human genes and there is a possibility that other primate gene models will be missed (see section 8). We classified gene duplications as complete if the entire transcript mapped within a SD interval and partial if only a portion of the transcript was contained within the duplication. For a complete list of genes by category see Table S7.

4.1 Gene Ontology Analysis

We used PANTHER (http://www.pantherdb.org/) to assess over- and under-representation of duplicated genes by biological process and molecular function. For each category of segmental duplication, we computed an expected number of genes for different biological processes based on their curated representation in the human reference genome (build35). The P-value was obtained from a binomial distribution and represents the probability that the difference between the observed and expected number of genes is random (Table S8).

4.2 Overlap with Dumas et al.

cDNA arrayCGH experiments have identified candidate lineage-specific expansions of genes and gene families within various primate lineages[19,20]. The comparison of human-great ape arrayCGH using the cDNA[20] with our own genomic arrayCGH results showed a good correlation. In short, we obtained the underlying raw log2 relative hybridization data (courtesy of Dr. Sikela) for each of the genes and correlated it with our genomic arrayCGH results (at the exonic level). The graphs below show the correspondence between the arrayCGH results of the two studies for each species comparison (Supplementary Note Fig. 2). As can be seen, there is generally a good correlation for intersecting genes (R2=0.576-0.668) (Supplementary Note Table 6).



Supplementary Note Fig. 2. Correlations between Dumas et al. (X axis). and our comparative arrayCGH results (Y axis).Only intersecting genic intervals are compared.

Supplementary Note Table 6. Correspondence between EST cDNA arrayCGH (Dumas et al.) vs. arrayCGH results.

	R	Rsq	Spearman's Rho coefficient
Human/Chimpanzee	0.759	0.576	0.782
Human/Bonobo	0.734	0.539	0.763
Human/Gorilla	0.818	0.668	0.792
Human/Orangutan	0.771	0.595	0.763

Despite this correlation, we have identified a large number of duplicated gene fragments that were not identified by Dumas and colleagues (39 human, 15 chimpanzee and 25

orangutan specific genes). We also intersected our gene sets with those identified in previous studies and found minimal overlap (25–40%) (Supplementary Note Table 7). There are important methodological differences that may account for this. In our study, we limited our analysis to duplications >20 kbp in length; therefore, genes mapping within duplications less than 20 kbp would not be identified. In the study of Dumas and colleagues, the authors could not distinguish processed pseudogenes from *bona fide* duplicated genes since their arrays were based on human cDNA. This would inflate the number of duplicated genes, while potentially divergent genes (due to positive selection) would exclude certain genes from characterization by cDNA microarrays. In several instances, we also found classification differences. For example, Dumas et al. used cDNA arrayCGH to classify lineage-specific versus shared duplications, while we used depth-of-coverage of WGS to assign duplications as lineage-specific or shared duplications. ArrayCGH, however, is relative while the latter WSSD approach is absolute. For example, a duplication that has increased in copy in human by arrayCGH but is still duplicated in other non-human primates could be assigned as a lineage-specific duplication in human.

Supplementary Note Table 7. Overlap of genes detected in this study and Dumas et al.

Genes in Dumas et al.	**Overlaps with corresponding APE SDs**	**Total**	**% Support**
Human increase in copy	67	166	**40.36%**
Chimpanzee increase in copy	15	62	**24.19%**
Orangutan increase in copy	78	321	**24.30%**

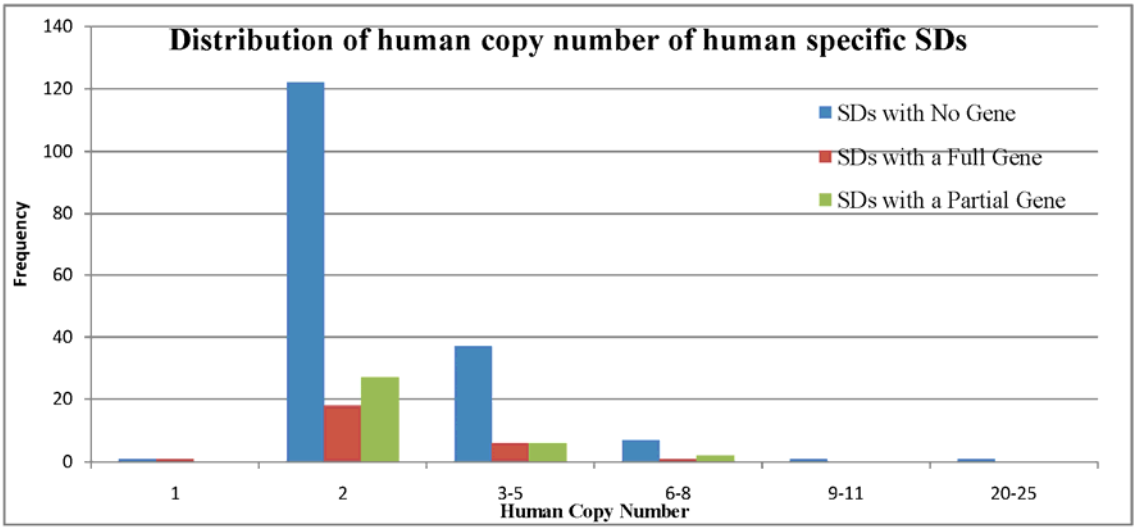
Genes in our APE SDs	**Overlaps with Dumas et al. 2007**	**Total**	**% Support**
Human-specific SDs (>20 kbp)	17	56	**30.36%**
Chimpanzee-specific SDs (>20 kbp)	8	23	**34.78%**
Orangutan-specific SDs (>20 kbp)	16	41	**39.02%**

We cross-referenced our genes with the genes reported in Dumas et al. 2007. 60% of lineage-specific genes detected in this study have not previously been reported.

4.3 Positive selection in gene families

Systematic characterization of positive selection has been difficult to assess due to the draft quality of non-human primate genome assemblies, the relatively incomplete gene annotation of these regions and the difficulties in constructing accurate gene models. Not surprisingly, most genome-wide analyses of positive selection have excluded these regions. While the focus of this manuscript has not been on determining and contrasting the evolutionary rates of segmental duplication with other forms of variation, we have performed a number of analyses to provide additional insight into the evolution of those gene families.

We first tested whether there was any difference between full-length and partial genes with respect to copy number of the duplicates. Here, we are specifically testing whether copy number itself has been selected (i.e. full-length genes are more likely to associate with higher copy duplicates). To avoid potential counting redundancy in our dataset, we used the non-redundant set of duplication subunits described by our group (Jiang et al. 2007) to categorize the genes—such that the analysis was performed at the level of the gene family. We limited our analysis to the most recent gene duplicates (i.e. emerged in the human lineage).We found no differences in the duplication copy number distribution between full-length and partial gene duplicates (Mann-Whitney U-test P-value=0.851) (Supplementary Note Fig. 3).



Supplementary Note Fig. 3. Distribution of human copy number of human specific SDs.

Next, we examined whether there was any difference in the extent of copy-number polymorphism for partial and complete gene duplicates between the two broad gene categories that we distinguished by our ontology analysis. We partitioned the genes into two groups: group 1 (n=347) consisted of young human-specific gene associated with neuronal activities, signal transduction or synaptic transmission, while group 2 (n=41) consisted of older duplications (shared with orangutan and macaque) associated with oncogenesis and amino acid metabolism/catabolism. We used the dataset from Redon et al.[14] to assess the extent of copy-number polymorphism as the number of examined individuals was sufficiently large. We found that group 1 human-specific genes were less polymorphic (66/347 or 19% are CNP) when compared to the older group 2 genes (20/41 or 49% are CNP). Given that we showed in the paper that shared SDs are as polymorphic as lineage-specific SDs, this significant difference (p=0.003, Fisher's exact test) is intriguing and may suggest that selection has been operating. We note, however, that the vast majority of duplicate genes are partial and therefore, by our assessment, incapable of producing full-length proteins although there is ample evidence of ESTs as well as fusion "gene" products from these duplicates. One possibility may be that this higher rate of

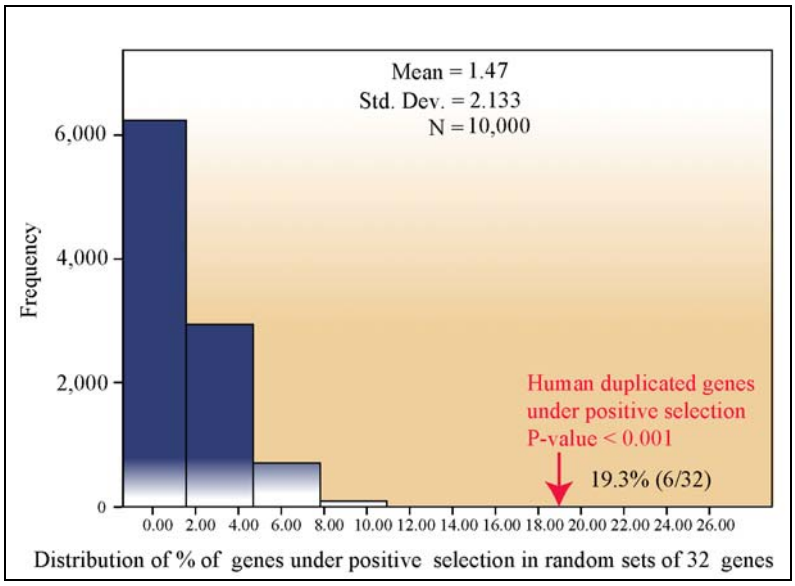
fixed duplicates is playing a role in gene regulation as has been recently proposed by other groups.

Finally, we tested more generally whether genes within expanded human great-ape gene families were more apt to show evidence of positive selection than a control set of unique (non-duplicated) gene families. We limited our analysis to 45 gene families where there was evidence of full-length gene models among the duplicates and experimental validation by array comparative genomic hybridization. This set of 45 included seven that were shared by human, chimpanzee and orangutan; three that were shared by human and chimpanzee; and 35 that were human specific. We used ENSEMBL to retrieve gene model information based solely on the human genome assembly and from non-human primate species if the gene could be used as an outgroup. We constructed multiple sequence alignments based on amino acid composition and backtranslating to DNA (DIALIGN). We manually inspected each alignment and removed paralogs with potential misalignments, conservatively retaining a set with highly similar sequences. We assessed positive selection using three maximum likelihood tests: two tests (Tests 1 and 2) are designed to detect positive selection at individual sites (across the alignment; site codon-substitution models)[21,22] while the third test (Test 3) is designed to assign positive selection to branches along the phylogenetic tree (branch codon-substitution model)[23]. Of the 45 initial hominid gene families, we excluded 14 gene families with insufficient number of gene models (reduced power) to perform tests of positive selection. 38% (12/31) of the gene families showed evidence of positive selection by at least one of the three tests, while 19.3% (6/31) showed evidence of positive selection by all three tests (Supplementary Note Table 8). These included genes assigned to signal transduction (FCGR1A, LOC440607), protease inhibition (BIRC1), basal transcription (GTF2IRD2B) and chromatin-binding (DKFZP434A0131). Several of the genes in this list have no known function or ontological gene classification (nearly 25% (535/2200) of all hominid gene family expansions fall into this category).

Supplementary Note Table 8. Summary of genes under positive selection.

M1(neutral) vs. M2a (positive selection)	Codons with positive selection (Test1)	Codons with positive selection (Test2)	Branches with excess of dN/dS (Test3)	Gene Families
Significant	Yes	Yes	Yes	NM_000566_FCGR1A NM_001004340_LOC440607 NM_004536_BIRC1 NM_001003795_GTF2IRD2B NM_001002840_DKFZP434A0131 NM_022661_SPANXC
Significant	Yes	Yes	No	NM_001008218_NULL
No Significant	Yes	Yes	Yes	NM_207418_MGC57827 NM_032579_RETNLB NM_173537_GTF2IRD2 NM_033514_LIMS3
No Significant	Yes	Yes	No	NM_001025202_NULL

In order to test the significance of positive selection with respect to non-duplicated genes, we simulated all our statistics on the alignments for 13,721 orthologous primate gene alignments[24] (courtesy of Mark Adams). In this dataset, we identified only 209 orthologous genes where positive selection Models (M2a) were significantly better (P-values<0.01) than the neutral model (M1a). This represents 1.49% of all the genes and is consistent with previous estimates of the primate genome average based on studies of human (2%)[25,26] and macaque (1.7%)[8]. We then permutated 10,000 random samples of 32 genes in the orthologous samples and recorded how many times we found at least six genes under positive selection. The result of the permutation shows that hominid gene families are clearly enriched by positive selection (P-value<0.0001) (Supplementary Note Fig. 4).



Supplementary Note Fig. 4. Histogram of the percentage of positively selected genes in a random distribution of unique genes versus the real observed value.

In summary, we can conclude that genes within segmental duplications are enriched for positive selection, supporting the idea that adaptive processes are playing an important role. Our analysis, however, suggests that only a small fraction of duplicated gene products show evidence of positive selection by classical tests or by indirect tests of copy-number variation. Interestingly, we find a significant difference in the extent of human copy-number polymorphism between neuronal, signal transduction and synaptic transmission (group 1) when compared to amino acid metabolism/oncogenesis gene duplicates (group 2). This may suggest selective pressure but not at the amino-acid level. There are two important caveats to this analysis. First, not all adaptive or selected events may relate to protein-encoding genes. Segmental duplications have the potential to generate non-coding mRNA transcripts that may be important in the regulation of ancestral gene products. Second, there is considerable bias against annotation of protein encoding genes that lack orthologs in more distant outgroup species due to technical and analytical regions. Over the last six years, numerous great-ape and human gene families have been described including recent fusion genes that were initially unannotated and for which there is now evidence of positive selection. Characterization of each of these gene

families, however, is a slow process requiring high-quality BAC or cDNA sequencing in various outgroup.

4.4 Fixed human-specific gene duplications

Among the candidate set of lineage-specific expansions (Table S9), we sought to identify gene duplications that emerged specifically within the human lineage and had become fixed in copy number (i.e. showed no evidence of copy-number polymorphism (CNP)). First, we identified all RefSeq genes (both complete and partial) mapping to human-specific segmental duplications. Next, we excluded any gene where the underlying genomic region showed any evidence of copy-number variation (CNV) (442 complete genes and 3699 partial overlapping genes) based on three sources: the Database of Genomic Variants v.4 (http://projects.tcag.ca/variation/)[16], a human structural variation map created using fosmid ESPs[17] and our own intraspecific human arrayCGH results. All human SDs that did not overlap with a CNV region were considered fixed (see Table S10). We identified only three duplicated genes that show a complete gene structure and show no evidence of copy-number polymorphism: GCUD2, OR1D5 and SLC29A4. Remarkably, SLC29A4—solute carrier family 29 (monoamine transporter), member 4—is responsible for the reuptake of monoamines into presynaptic neurons[27].

4.5 Comparative Analysis of Potential Human Adaptation Gene Families

We explored the individual cases associated with human adaptation as suggested in recent reviews[28]. In a few cases the genes of interest are below the limit of our 20-kbp threshold (although there is still information on each of these).

AMY1-> A human-specific duplication with a duplication unit less than 8 kbp in size that was found to be polymorphic in humans. Increased copy number was confirmed by arrayCGH for all human to non-human primate comparisons (i.e. average log2 hybridization intensity for human/chimpanzee comparison).

AQP7-> According to our analyses, 80% of the gene structure overlaps a shared duplication with human, chimpanzee and bonobo. Once again humans have more copies

than chimpanzees and bonobos, although in humans there is no evidence of copy-number polymorphisms. Notably, in chimpanzee there are three chimpanzee individuals that have fewer copies than the reference chimpanzee genome.

DUF1220-> Our duplication analysis shows that this duplication is shared among human, chimpanzee and macaque (although we predict that the duplication is a single copy in orangutan). Consistent with the publication by Popesco et al.[29], our arrayCGH results for this shared duplication indicate that there are more copies in humans when compared to most other primates. Interestingly, we note copy-number variation among humans (ABC13, for example, has fewer copies than G248), but we see no variability among chimpanzees suggesting that it is fixed in these lineages.

5. Copy-number polymorphism (CNP)

5.1 Segmental duplications and copy-number polymorphism

Using our primate SD microarray, we assessed all regions classified as segmental duplication among human, chimpanzee, orangutan and macaque for copy-number polymorphisms within the human, chimp and orangutan species. We tested DNA samples from 8 HapMap individuals (4 African DNA samples, namely: NA18517, NA18507, NA19240 and NA12878 and 4 non-African DNA samples NA18956, NA18555, NA19129 and NA12156), 8 chimpanzees ("Logan", PR00238, PR00226, PR00496, PR00738, PR01097, PR1105 and PR01009) and 8 orangutans ("PPY9", AG05252, AG06105, "Hati", "PPY6", "Tengku", AG12256 and "Puti") against a common reference sample (NA15510 for humans, "Clint" for chimpanzees and "Susie" for orangutan) for copy-number variation by arrayCGH (as previously described in section 2.3). In order to be classified as copy-number variant, we required two or more individuals to show a significant departure of $\log_2$ ratio signal intensity (estimated from single copy regions, section 2.3). The percentage of copy-number variant basepairs were compared for segmental duplications classified by type (from the computational approach; Supplementary Note Table9) and for the segmental duplications combining information from gorilla (GGO) and bonobo (PPA) arrayCGH (Supplementary Note Fig. 5).

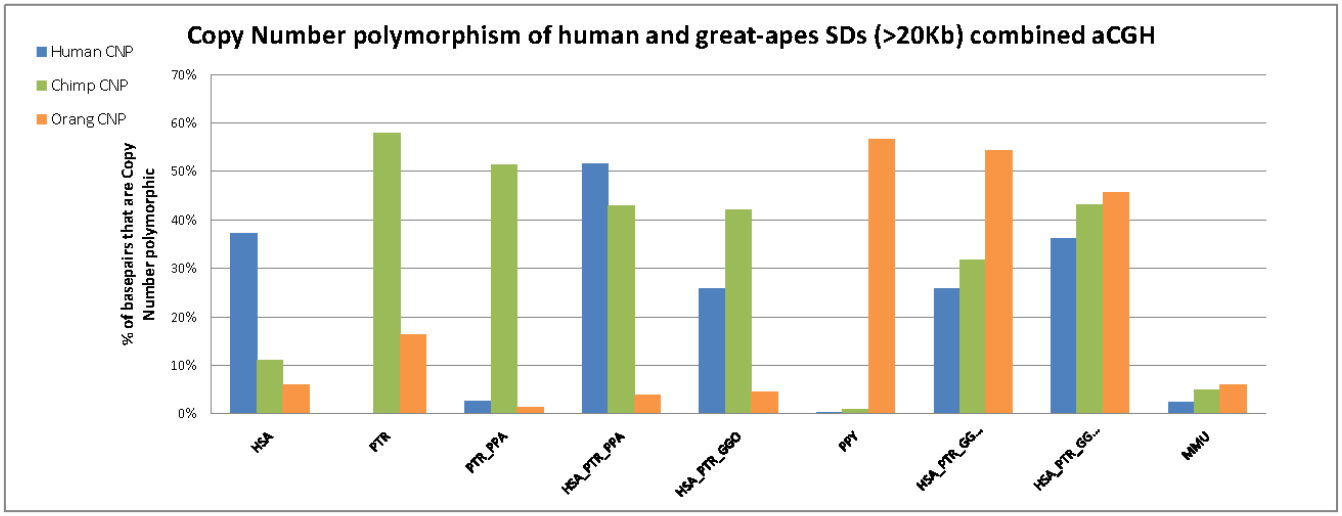
Supplementary Note Table 9. Segmental duplications and copy-number polymorphism.

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%**

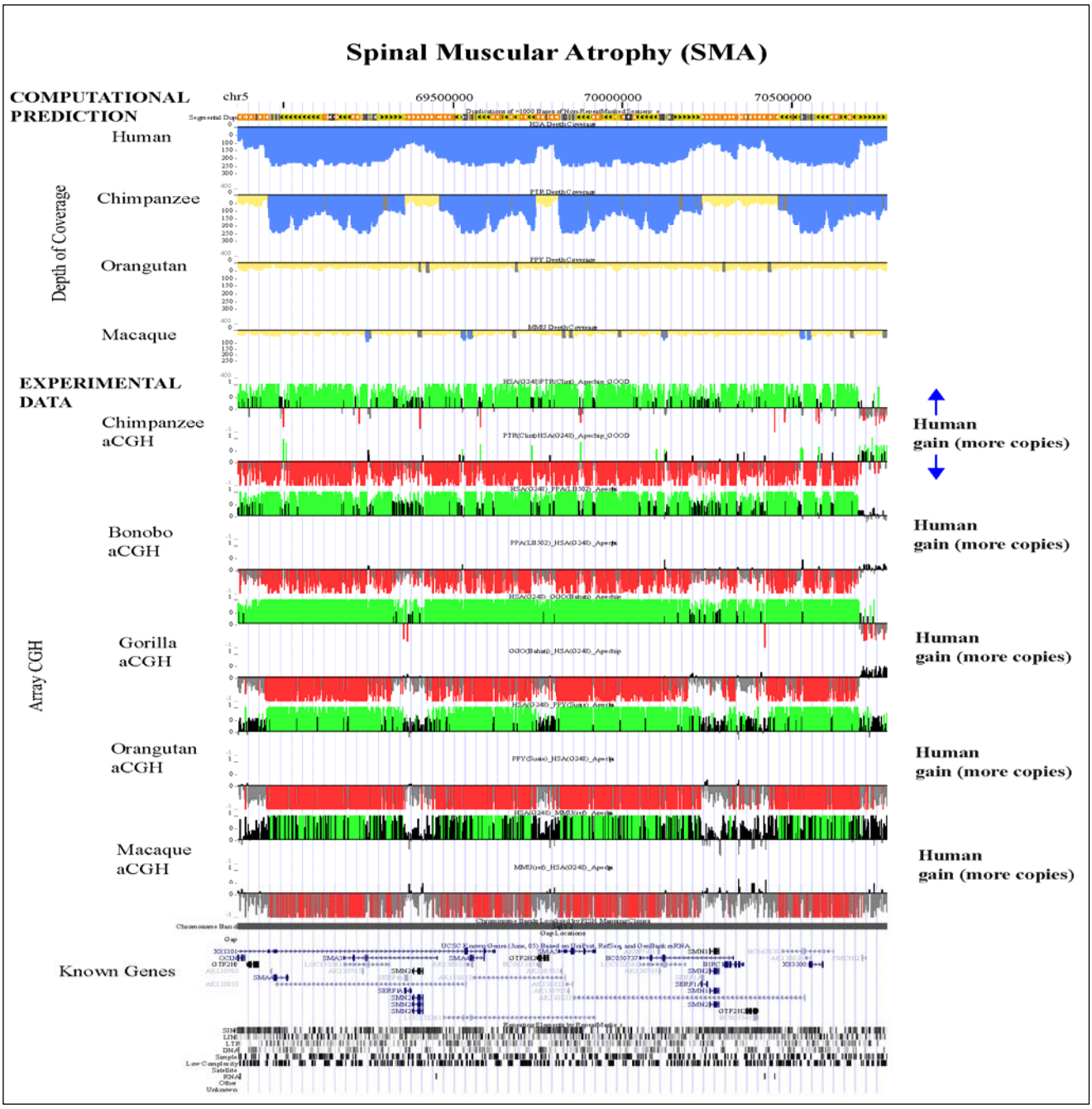
The percentage of copy-number variant basepairs was computed for segmental duplications classified by type.



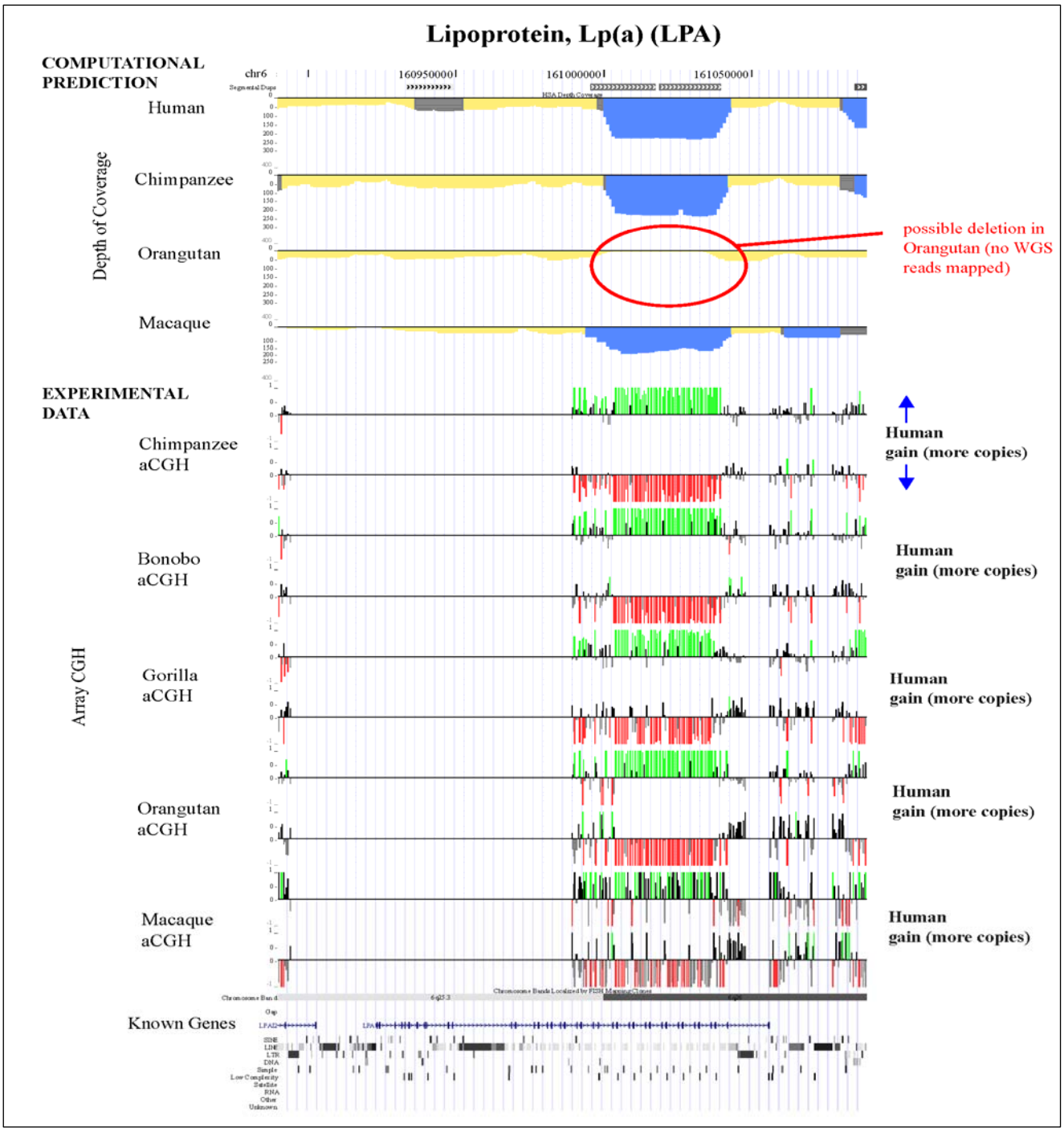
Supplementary Note Fig. 5. Copy-number polymorphism of human and great-ape SDs. *In this figure, SDs were further categorized (see Fig. 2c) using arrayCGH information from gorilla and bonobo. The same trends reported in the text are observed.*

5.2 Evolutionary history of disease genomic regions and disease susceptibility loci

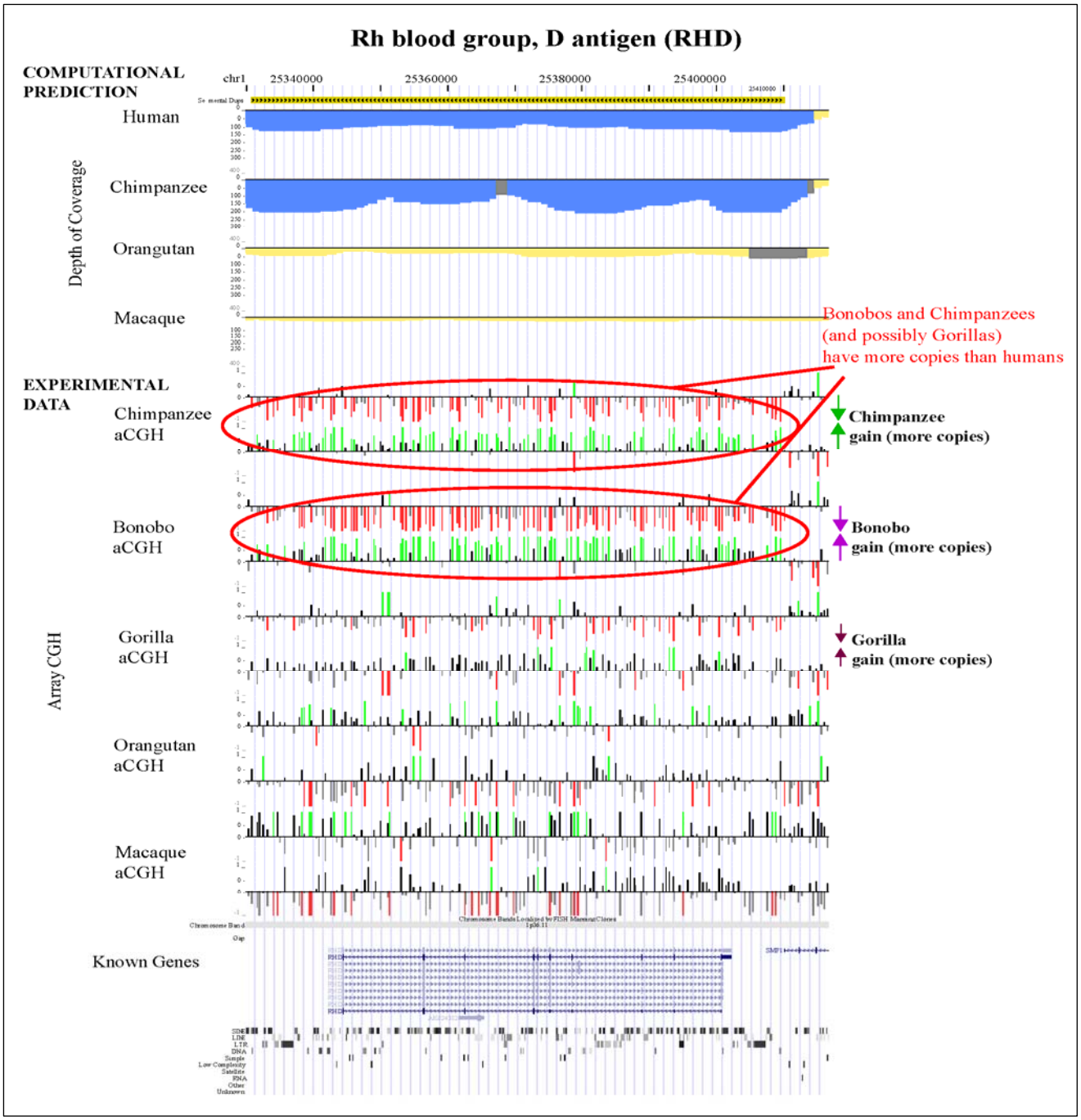
The majority of human copy-number polymorphisms (9/12) associated with common diseases and disease susceptibility loci map to human segmental duplications (Supplementary Note Table 10), despite the fact that most copy-number variants supposedly map outside of duplicated regions[14] and 25-30% of all large-scale microdeletions and microduplications associated with mental retardation and developmental delay are segmental-duplication mediated[30]. This represents a 10- to 25-fold enrichment. We have performed a comparative evolutionary history of both classes of disease-associated segmental duplications (Supplementary Note Table 10, Supplementary Note Fig. 6) and to exemplify an application of our dataset; we also show several interesting examples (Examples 1, 2, 3 and 4).



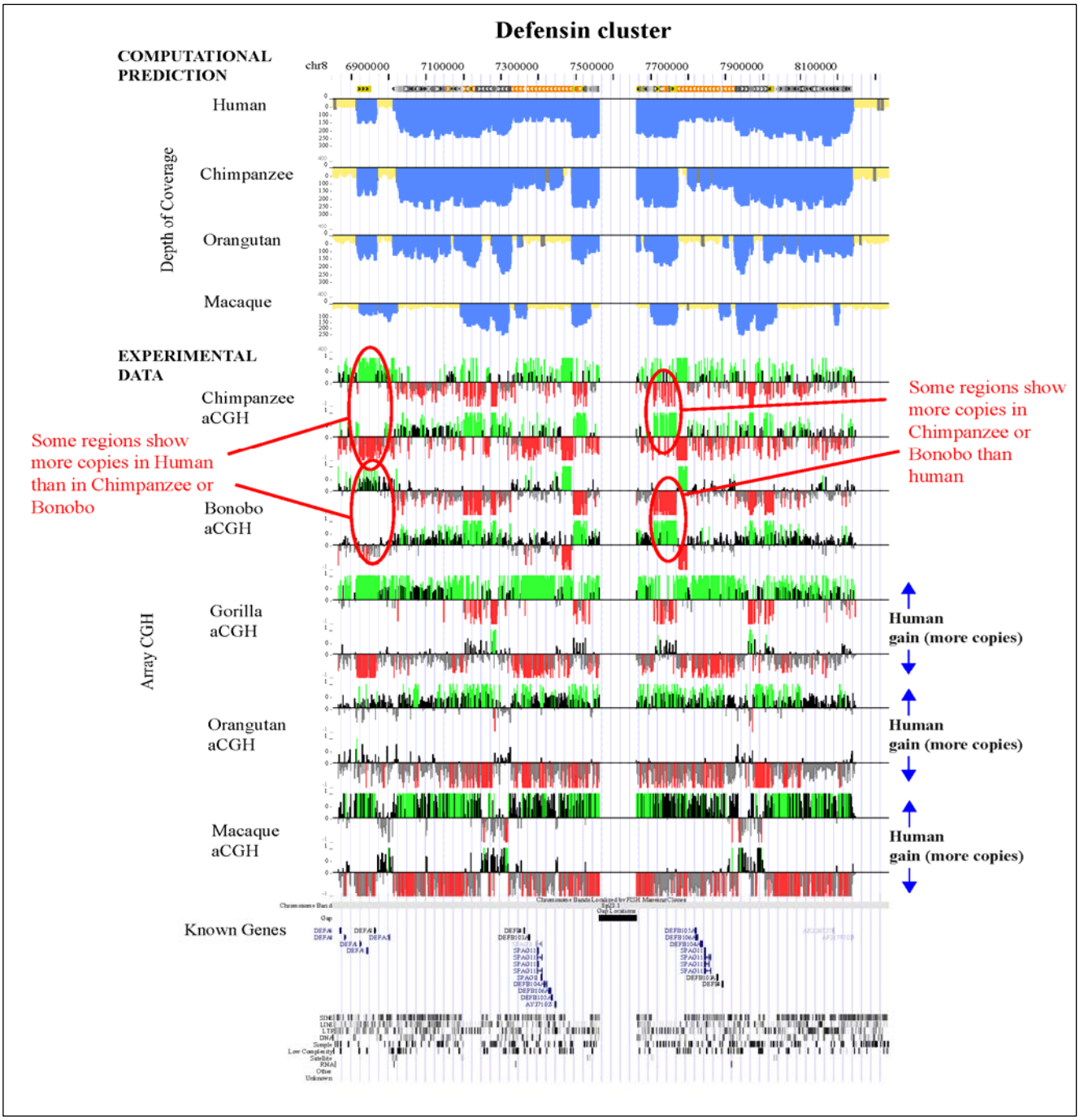
Example 1. Spinal muscular atrophy. *This region is duplicated in both human and chimpanzees but the experimental results of arrayCGH show that humans have had an expansion of this loci, since we saw a gain in copy number in all the hybridizations (even in the human/chimpanzee comparison).*



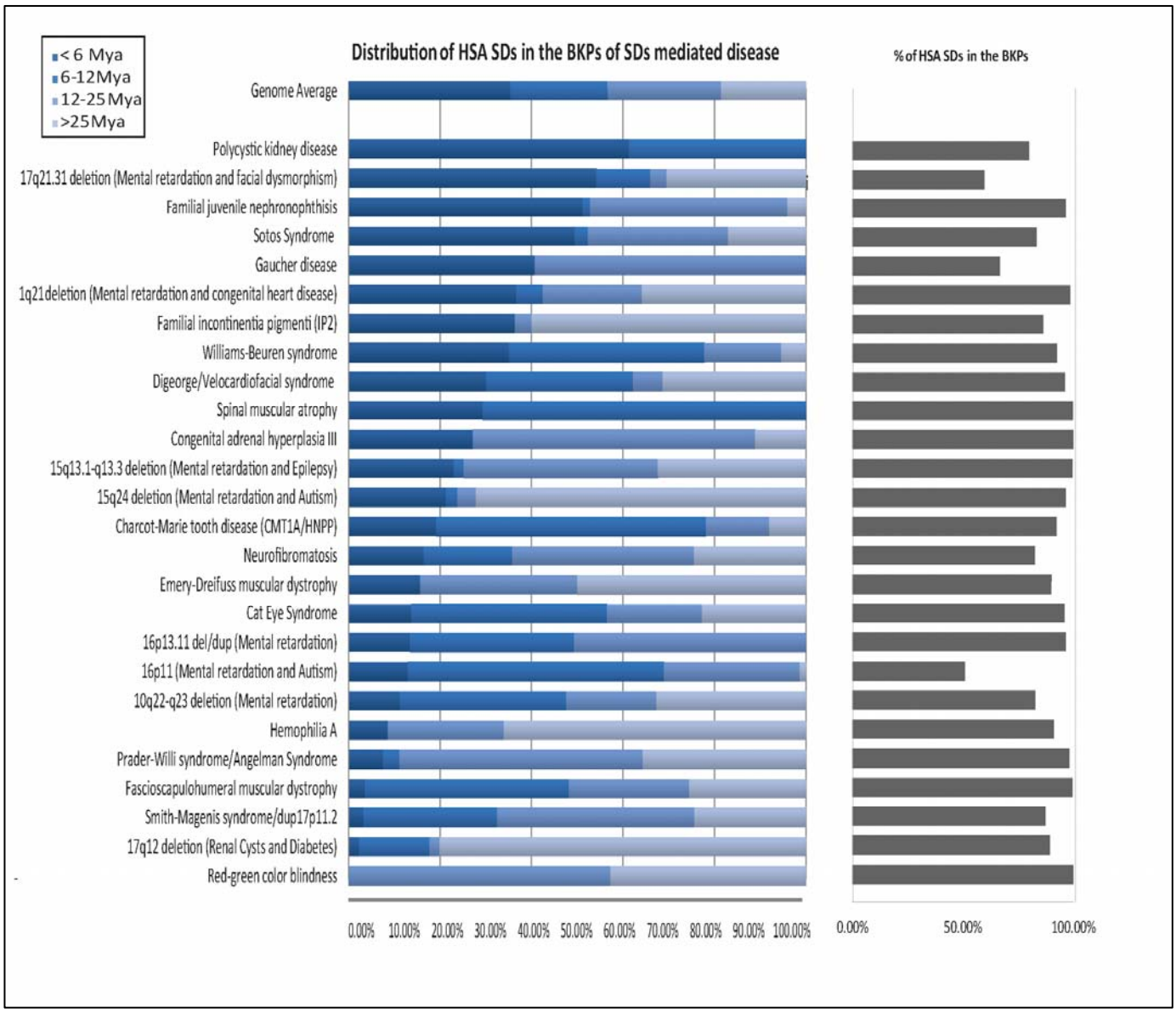
Example 2. Lipoprotein Lp(a). *This gene overlaps partially with a segmental duplication that is shared in all the species except orangutan (in which there seems to be a deletion). The duplication is a tandem expansion (size tandem ~5 Kb) that according to arrayCGH has been more expanded in humans more than in any other primate.*



Example 3. Rhesus blood group, D antigen. *This gene is found to be duplicated in both human and chimpanzee, but arrayCGH results show that it also duplicated in bonobo and gorilla, and the non-human African apes have more copies than humans as previously suggested*[31].



Example 4. Defensin cluster. *This is one of the more complex examples, in which even if human and chimpanzees have more copies (in general) than orangutan or macaque, the copy number relationship between them is difficult to disentangle since some regions show more copy number in humans (overlapping the coding regions) and other regions show more copy number in chimpanzee.*



Supplementary Note Fig. 6.Comparative analysis of disease-associated SDs. *Segmental duplications (a.k.a. low-copy repeats) mediating recurrent rearrangements associated with human disease were comparatively analyzed among the primates. Based on lineage-specific or shared status, we estimated the evolutionary age of each human duplication within each region (<6 Mya for human-specific SDs, 6-12 to duplications shared with chimpanzee 12-25 for those shared with orangutan and > 25 for those shared with macaque). Young and evolutionarily old breakpoint regions are distinguished as compared to the genome average.*

Supplementary Note Table 10: Great ape segmental duplications and disease susceptibility loci.

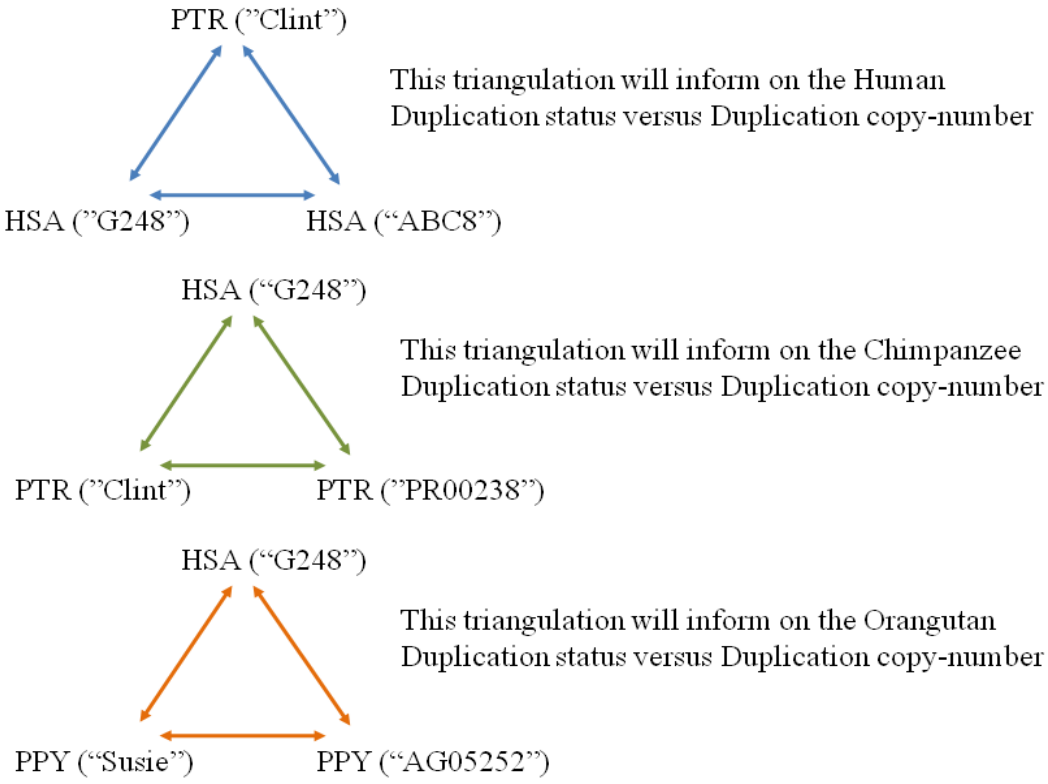
Gene name	Disease or risk factor	Gene ID	Description	Chr (gene)	Start	End	SD length (bp)	Status	HSA Dup	PTR Dup	PPA Dup	GGO Dup	PPY Dup	MMU Dup	Classification	Comments
GSTM1	squamous cell carcinoma, aplastic anemia,	NM_000561	glutathione S-transferase M1 isoform 2	chr1	109942484	109948409	18,406		0*	1*	?	?	1*	1	Great-ape/OWM	Deletion Human (some people single copy)
CYP2D6	Reduced drug metabolism	NM_000106	cytochrome P450, subfamily IID, polypeptide 6	chr22	40847001	40851379	9,004		1*	0	0	0	0	0	Human specific	
CYP21A2	congenital adrenal hyperplasia,	NM_000500	cytochrome P450, family 21, subfamily A,	chr6	32114061	32117396	32,853		1*	0*	1	?	1	0	Great-ape specific	Deletion Chimp? (Clint single copy)
LPA	Coronary heart disease	NM_005577	lipoprotein, Lp(a)	chr6	160922926	161055702	21,743	Partially duplicated	1*	1*	?	?	D	1	Great-ape/OWM	Tandem Repeat/Expansion Human
RHD	Rhesus blood group	NM_016124	Rhesus blood group, D antigen	chr1	25344355	25401018	61,003		1*	1*	1	?	0	0	African Ape ancestor	More copies Pan
CFH	Age related macular degeneration	NM_000186	complement factor H	chr1	193352798	193448288	28,649	Partially duplicated	0*	1	0	0	0	1	Great-ape/OWM	
C4A	Lupus	NM_007293	complement component 4A preproprotein	chr6	32090550	32111173	32,853		1*	1	?	0	1*	0	Great-ape specific	
C4B	Lupus	NM_000592	complement component 4B preproprotein	chr6	32057813	32078435	32,736		1*	2 STD DEV	1/2	?	1	2 STD DEV	possibly Great-ape/OWM	Half gene more copies bonobo
DEFB4	Psoriasis/Crohn's disease	NM_004942	defensin, beta 4	chr8	7789609	7791647	299,279		1*	1*	?	?	1	1	Great-ape/OWM	Deletion Orang?
DEFB103	Psoriasis/Crohn's disease	NM_018661	defensin, beta 103A precursor	chr8	7273828	7275280	309,878		1*	1	?	?	0	0	African Ape ancestor	

DEFB104	Psoriasis/Crohn's disease	NM_080389	defensin, beta 104A	chr8	7315236	7320014	309,878	1*	1*	1	?	0	0	African Ape ancestor
CCL3L1	HIV/AIDS	NM_021006	chemokine (C-C motif) ligand 3-like 1 precursor	chr17	31647958	31649843	64,469	1*	1	?	?	1*	1	Great-ape/OWM

Summary of the evolutionary origin of duplicated disease susceptibility genes. Presence or absence of the duplication in every species was assigned irrespectively of the occupancy of the duplication in the loci.

6. Duplication status vs. copy number

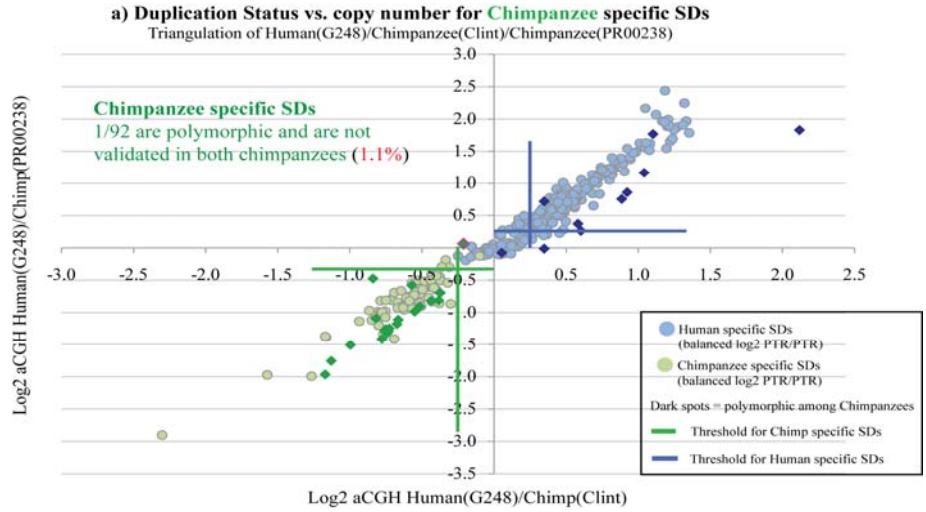
During our analysis it became important to differentiate the concepts of "duplication state" from "duplication copy number". The duplication state is simply a binary variable indicating whether or not a sequence is duplicated within a species. In contrast, duplication copy number refers to the integer number of copies of the sequence present in a given genome. Differences in duplication copy within a population may be referred to as copy-number polymorphisms. However, a given locus that differs in duplication copy across populations or species may still represent the same duplicated state (i.e. >=2 copies [duplicated]). Based on our analysis of 8 individuals within each species (see above), we found that most ape SDs are copy-number polymorphic (Fig. 2c). These CNPs might theoretically reflect changes in duplications status (i.e. absence/presence of duplications) or, alternatively, changes in the number of copies of these duplications. This is an important consideration since our estimation of the rate of duplications is based only on the analysis of a genome from a single individual. To answer this question, we performed a set of arrayCGH "triangulations" in which we investigated how copy number differences within an individual species affected our classification of the duplication status between species (Supplementary Note Fig. 7). As an example, the chimpanzee duplication status versus duplication copy-number polymorphism was retrieved from a three-way comparison of two chimpanzees ("Clint" and "PR00238") against one human ("G248"). This triangulation would allow us to infer how many of the duplications that are polymorphic among the two chimpanzees are still classified as duplicated when compared to the human.

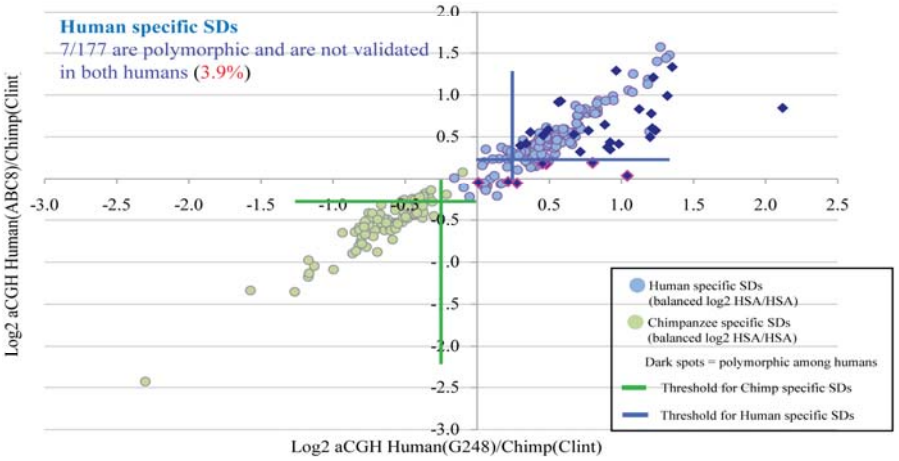


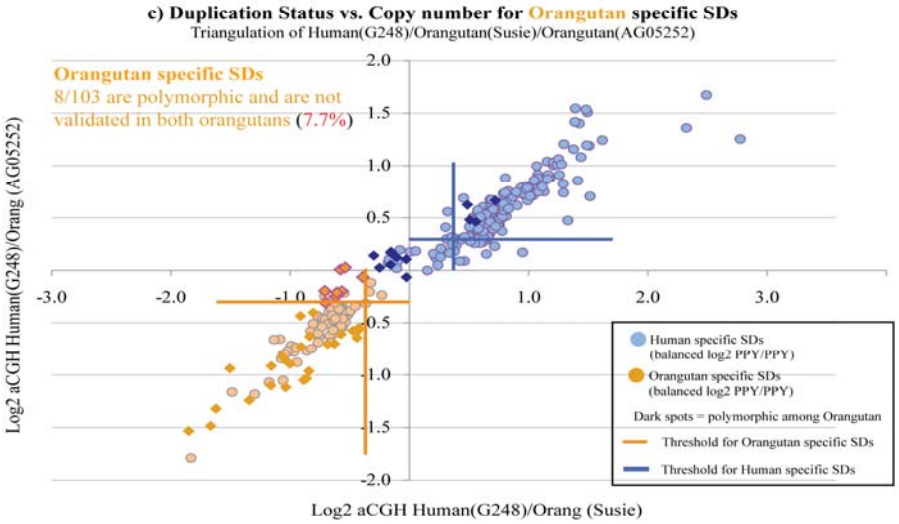
Supplementary Note Fig. 7. Schematic representation of the three triangulations performed in arrayCGH to interrogate the duplication status versus the duplication copy number in human, chimpanzee and orangutan.

The results of our analysis (Supplementary Note Fig. 8) suggest that only 1–8% of the SDs changed their duplication status while 18–32% of the duplications were copy-number polymorphic between two individuals. The chimpanzee had the lowest percentage of individual-specific SDs that lose duplication status when polymorphic (1%) whereas orangutan presented the greatest percentage (7.7%) even when the two orangutans tested are from the same sub-species (*Pongo pygmaeus abelii*, Sumatran orangutan). Human SDs showed an intermediate level (3.9% of human specific SDs). These relatively low levels of individual-specific SD suggest that while most of our detected SDs are copy-number polymorphic within the species, the duplication status (presence/absence of the duplication) remains a largely invariant feature of that species. These data are consistent with the conclusions from the duplication analysis of the Venter and Watson genomes and suggest that the effect of copy-number polymorphisms will

have a negligible effect on our duplication rate estimates (see Fig. S7).







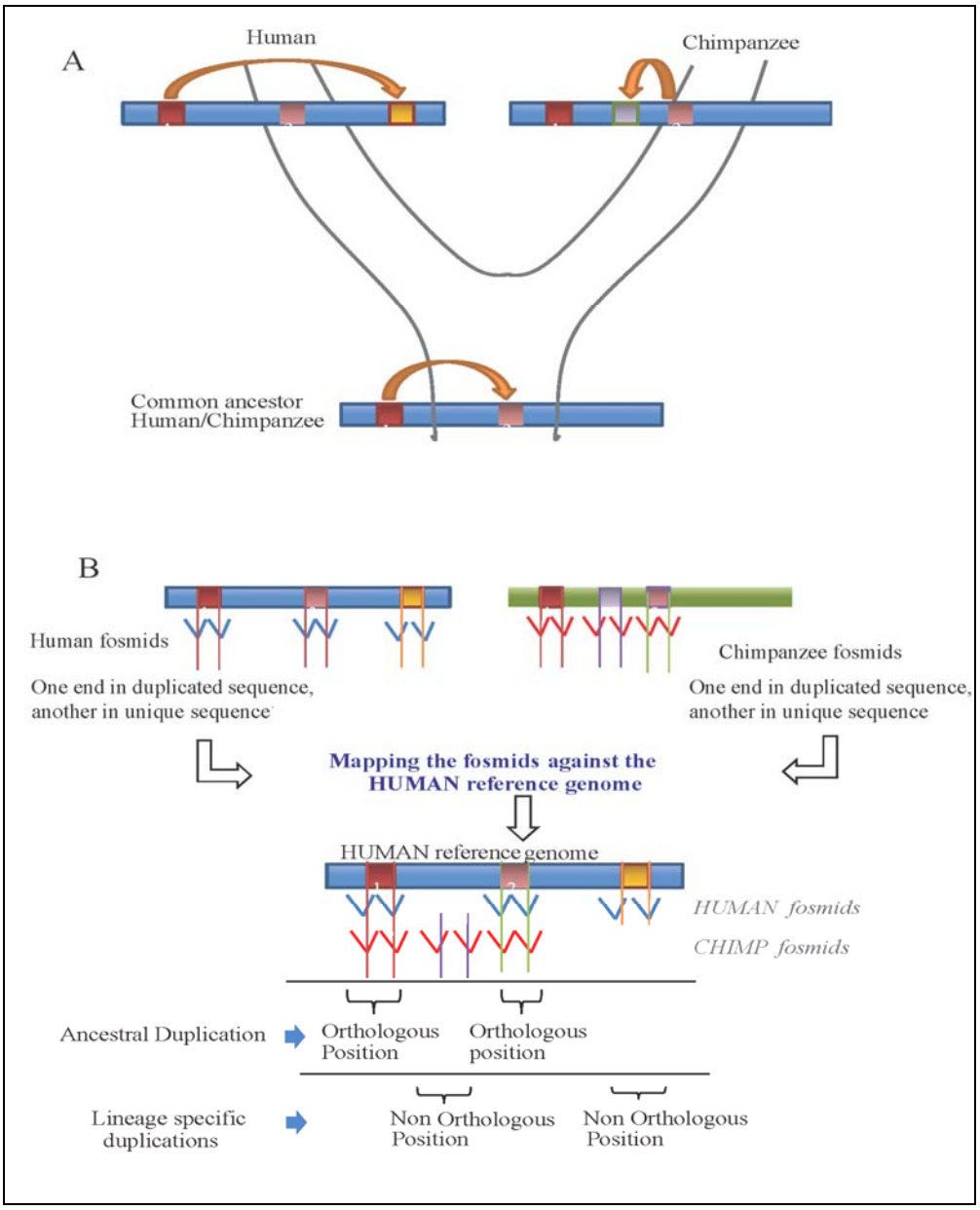
Supplementary Note Fig. 8. Duplication state vs. copy-number polymorphism. *Since we detected segmental duplications in a single individual for each species, assignment of duplications to a specific branch or lineage may be confounded by copy-number polymorphism. To assess the extent of this effect, we compared two unrelated chimpanzees ("Clint" and "PR00238") against the same reference human genome ("G248") by arrayCGH and examined the status of predicted human-specific and chimpanzee-specific duplications. Our analysis revealed that 99% of assigned chimpanzee-specific duplications and 96% of human-specific duplications were correctly classified based on our thresholds even when they were found to be polymorphic. Significant copy-number polymorphisms in the two chimps (in dark green) were observed for 23 of the intervals that were assigned as chimpanzee-specific events (N=92). Thus, while a significant fraction of these events may be polymorphic among the species (25%), the duplication status remains a relatively constant and predictable property of a species based on a sampling of a small number of individuals. The same approach was used for b) two humans ("G248" and "ABC8") against chimpanzee ("Clint") (Blue) and for c) two orangutans ("Susie" and "AG05252") versus human ("G248") (Orange).*

7. Estimates of recurrent duplication (Homoplasy)

We inferred the ancestry of duplication events based on the most parsimonious interpretation of the data from the four species. These assignments were based only on the duplication status without any consideration of the map location of the duplicated sequences. Therefore, this interpretation could be confounded by positions where a duplication is mistakenly inferred to be ancestral when, in reality, the same sequence has independently duplicated in multiple lineages. While such recurrent mutations are expected to be rare for most mutational processes, detailed studies of segmental duplications have shown that such recurrent duplications have, in fact, occurred[32]. In order to provide an estimate of recurrent duplications, we focused on 479 duplications predicted to be shared between the human and chimpanzee lineage (but not duplicated in other primate species). This number is a redundant count since many of these regions are represented multiple times within the human genome assembly. Considering the pair-wise segmental duplication relationships and requiring that 80% of each interval be covered by a given alignment reduces these to 136 sets of nonredundant shared duplications.

7.1 Orthologous human-chimpanzee shared duplications

The goal of this analysis was to assess what fraction of shared human/chimpanzee duplications were duplicated in the human/chimp common ancestor. If the sequence was duplicated in the ancestral population, we would expect to find the duplicated sequence present in two or more orthologous positions in humans and chimpanzees (although not all positions are expected to be shared since additional lineage-specific duplication events could occur without changing the duplication status of the sequence; see above). Alternatively, if the sequence independently duplicated in each lineage (without being duplicated in the common ancestor) then the only orthologous position would correspond to the original ancestral locus (see Supplementary Note Figure 9).



Supplementary Note Fig. 9. Schematic of end-sequence placement strategy to distinguish recurrent from deletion/lineage sorting events. *A) An ancestral locus (1) is duplicated to a new location (2) in the common ancestor of human and chimpanzee. Subsequent independent duplications occur in the human (gold) and chimpanzee lineages (grey). B) Fosmid end-sequences from chimpanzee (red angles) and human (blue angles) genome libraries are mapped against the human genome and identify four distinct mapping locations (three in chimpanzee and three in human). Two of these positions are shared, corresponding to the ancestral locus (1) and the duplication that occurred in the common ancestor (2), while one each is specific to the lineage. Since two or more sites are detected in orthologous position, we conclude that the initial duplication occurred in the ancestral lineage of both species.*

We assessed the genomic positions of duplicated sequences using end-sequence pairs from fosmid clones from a single human (NA15510; fosmid library WIBR2) and a single chimpanzee (Clint; CHORI-1251). For each of the shared duplications, we used the fosmid paired-end sequences to define the map location of the duplicated sequence in both NA15510 and Clint relative to the human genome reference assembly (build35). We then compared the anchored duplication positions between these two individuals and concluded that a sequence was duplicated in the ancestral population whenever at least two locations were identified in common between human and chimpanzee.

More specifically, we mapped paired-end sequence reads from 746,627 chimpanzee and 1,141,942 human fosmid clones against build35 using a previously described approach[17,33,34]. Treating chimpanzee and human separately, for each duplication we identified end-sequence pairs (ESPs) where one end mapped within duplicated sequence while the other mapped to an anchored position outside of the duplication interval. Such ESPs serve to map the location of the duplication by anchoring the duplicated sequence onto positions in the human reference genome assembly. We considered only those human ESP alignments with high quality-rescored sequence similarity (>97% for the end mapping within the duplication and >99% for the anchor placement). Due to sequence divergence between chimpanzee and human, we slightly relaxed the thresholds for chimpanzee alignments (>96% for the duplicated ESP and 98% for the anchor placement). In all cases, we required that both ESPs include at least 30 bases of Phred Q30 and 50 basepairs of non-repeatmasked sequence (2% divergence threshold from the repeat consensus). We permitted "tied" placements for ESPs mapping within the targeted duplications.

For the anchored map position, we carried out two separate analyses. First, we required that the anchored end (which places outside of the duplication interval) have a unique, best placement. Second, given the observed duplication shadowing effect (see above) and the high-quality of the human genome assembly, we also considered all end-sequence placements, including those having multiple "tied" anchored positions due to flanking duplications. In each case we then compared the anchored positions for the duplicated sequence between Clint and NA15510 and identified overlapping positions. Before intersecting positions, the size of each mapped locus was expanded by 50 kbp in each direction (50 kbp is ~3 standard deviations beyond the Clint fosmid

insert size distribution; this expansion corrects for uncertainty in anchored position caused by the clone insert size). Each of the 479 intervals was analyzed separately with results then collapsed into the corresponding 136 nonredundant interval sets (Supplementary Note Table 11). Between 10–25% of the assigned duplications mapped to a single position in the human genome assembly (Supplementary Note Table 12)—a result that may reflect that the sequence is a tandem duplication or that it is, in fact, not duplicated in the examined individual.

Supplementary Table 11. Chimp/human shared duplication map positions.

	Unique Anchors		All Anchors	
	NA15510	**Clint**	**NA15510**	**Clint**
0 defined positions	0	2	0	0
1 defined positions	35	15	26	14
2 or more defined positions	101	119	110	122
TOTAL	136	136	136	136

The map locations for 136 shared duplications in human and chimpanzee were determined based on fosmid paired end-sequence placements. An anchored position was defined by clones having one end matching the sequence of a given duplication interval and the other placing outside of the interval. Map positions were classified as uninformative (0 defined positions), mapping to a single locus (1 defined position) or mapping to multiple locations (2 or more defined positions).

Considering only those ESPs that are anchored within unique sequence, there are 92 duplications that have multiple mapped positions in both human and chimpanzee. Of these, 86% (79/92) have at least two intersecting locations (i.e. there are two or more locations in both chimpanzee and human that are orthologous to one another). Similarly, if we include ESP anchors that map within flanking duplicated sequence, we find that 85% (88/103) of ESPs map to orthologous locations in chimpanzee and human (Supplementary Note Table 12). We note that this analysis is blind to sequences that were tandemly duplicated in the ancestral population and then subsequently duplicated to dispersed locations in each lineage. However, based on these results we estimate that 85% of the shared duplications represent sequences that were already duplicated in the human-chimpanzee common ancestor.

Supplementary Table 12. Shared chimpanzee-human duplications.

	Unique Anchors	**All Anchors**
0 shared positions	0	0
1 shared position	13	15
2 or more shared positions	79	88
TOTAL	92	103

Comparison of the anchored positions for sites having two or more mapped positions in both chimpanzee and human. Sites having two or more positions in common reflect shared duplications that were present in the chimpanzee-human ancestor.

7.2 Recurrent duplications vs. lineage-specific deletions

In order to further characterize potentially recurrent duplication events, we identified a set of segments with a pattern of duplication status inconsistent with the human-chimpanzee-gorilla-orangutan phylogeny (Table S5 and S6). We focused on intervals greater than 20 kbp in size that were confirmed by our arrayCGH experimental analysis. Using duplication positions defined by paired end-sequence mapping (see 6.1 for description of methodology), we sought to distinguish between independent duplications and lineage-specific deletions. In each of these analyses we considered only those ESPs anchored within a unique position of the human genome.

H+C-G+ duplication intervals

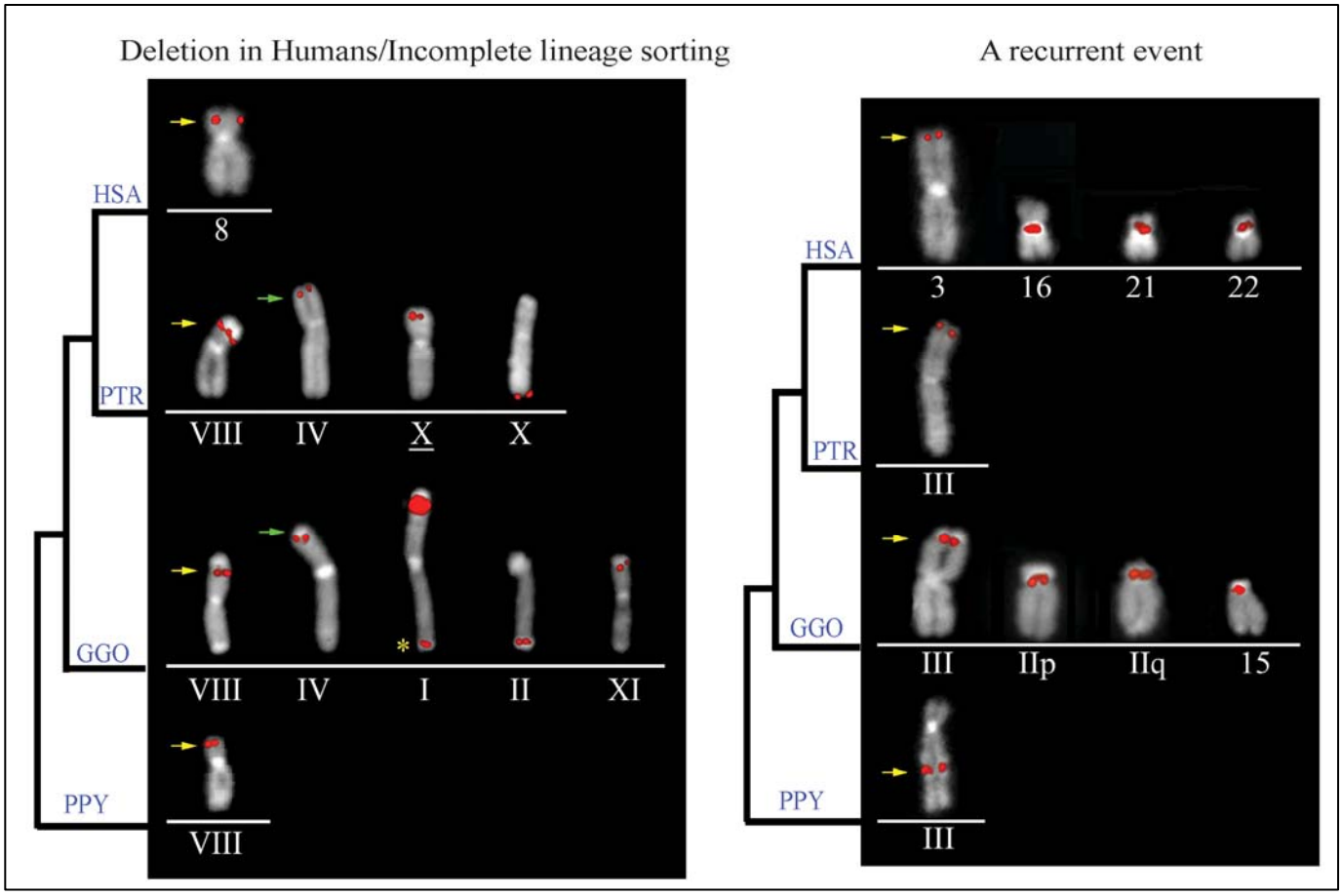
We identified 73 segments that are duplicated in human and gorilla but not in chimpanzee (termed *H+C-G+ segmental duplications)*. Considering the pair-wise segmental duplication relationships within the human genome (WGAC)[10], this set reduced to 43 nonredundant intervals duplicated in gorilla and human but not in chimpanzee. We used the ESP-duplication anchoring approach to map the location of these 43 duplications in both the gorilla and in the human genomes. For this purpose, we took advantage of the gorilla plasmid paired-end sequences generated as part of the Gorilla Genome Project by the Sanger Center (8.1 million end-sequences from a female *Gorilla gorilla,* Kamillah [NCBI trace repository]) and human fosmid paired-end sequences (2.1 million end-sequences from human PDR sample NA15510 [WIBR2 library]). We considered only those gorilla ESP alignments with high quality-rescored sequence similarity (>95.50% for the end mapping within the duplication and >97.5% for the anchor placement; NA15510 criteria same as in section 7.1). Due to the differences in clone insert sizes between the plasmid and fosmid libraries, we considered all sequences within 50 kbp of anchored human and within 10 kbp of gorilla anchored positions. Additionally, we required that each mapped position be supported by two or more clones. We found that 15/43 intervals mapped to multiple, distinct locations in the human and gorilla genome (Supplementary Note Fig. 10). Of these, 80% (12/15) had two or more anchored locations in common. This suggests that these sequences were duplicated in the human-gorilla ancestral population and subsequently lost in the chimpanzee lineage or incomplete lineage sorting.

H-C+G+ duplication intervals

We identified 37 genomic intervals that are duplicated in chimpanzee and gorilla but are not duplicated in human or orangutan (termed *H-C+G+ segmental duplications*, Table S6). Using 3,962,791 gorilla plasmid and 15,220,669 chimpanzee plasmid end-sequence pairs, we similarly mapped the position of these duplicated sequences within the gorilla and chimpanzee genome. We considered only those chimpanzee ESP alignments with high quality-rescored sequence similarity (>96% for the end mapping within the duplication and >98% for the unique anchor placement; gorilla criteria same as above) and searched map positions within 10 kbp on either side of the ESP. As above, we required that each mapped position be supported by two or more clones. We found that 24/36 intervals mapped to multiple, distinct positions within the chimpanzee and gorilla genomes. Of these, only 42% (10/24) mapped to orthologous locations within the genome indicating that their absence in the human genome was the likely result of deletion or incomplete lineage sorting. The remaining 58% (14/24) may represent independent and recurrent duplication events in the chimpanzee and gorilla lineage (see below).

H+C-G-O+ duplication intervals

We identified 44 intervals that are duplicated in humans and orangutans but not in chimpanzee or gorilla (Table S6), corresponding to 10 nonredundant duplications. Using human (NA15510) and orangutan fosmids (n = 963,199 and 567,676, respectively, trace archive query: SPECIES_CODE = 'PONGO PYGMAEUS ABELII' and TRACE_TYPE_CODE = 'CLONEEND' and INSERT_SIZE <50000), we mapped the positions of these duplicated sequences. We considered only those orangutan ESP alignments with high quality-rescored sequence similarity (>94.5% for the end mapping within the duplication and >95% for the unique anchor placement) and searched map positions within 50 kbp on either side of the ESP. Due to the lower coverage of the orangutan fosmid library, we considered locations supported by a single ESP placement. Thus, the mapping information should be interpreted more cautiously. We found that only 6/10 intervals had multiple, distinct map locations in the orangutan genome and human genome. Of these, 67% (4/6) mapped to orthologous locations between human and orangutan suggesting a potential polymorphic deletion in the ancestral human/African-ape ancestor with lineage-specific sorting among the great apes.



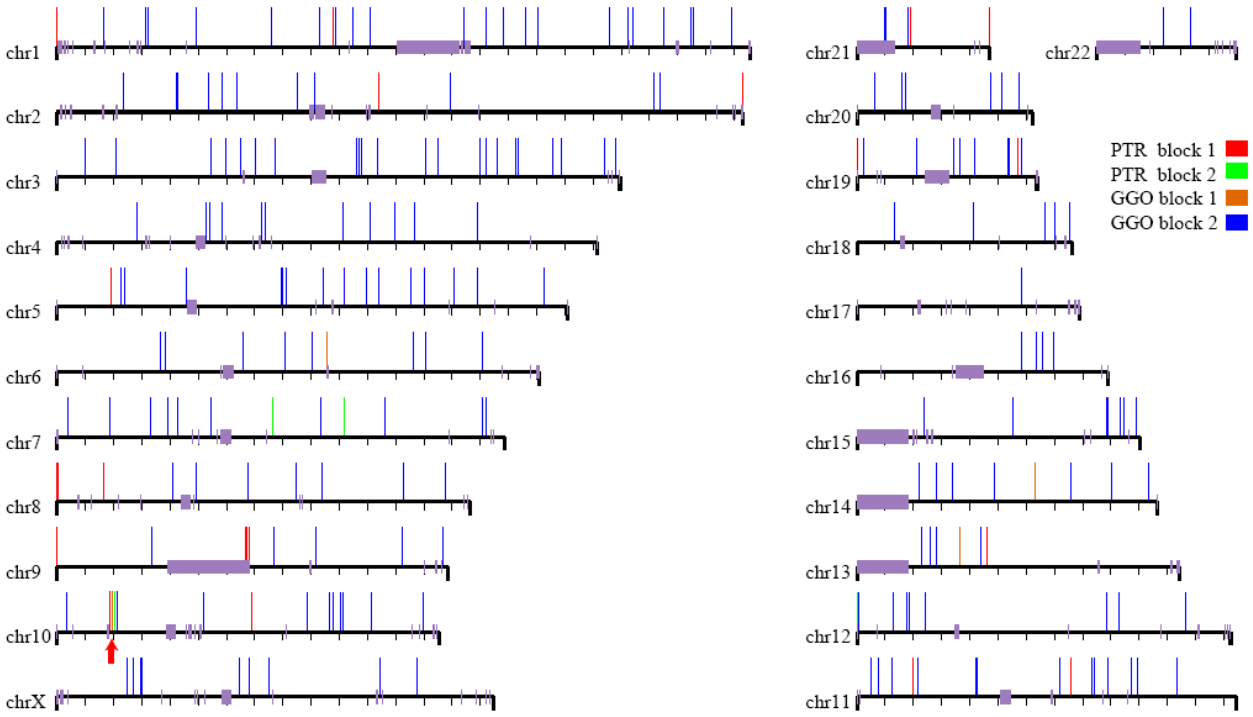
Supplementary Note, Fig. 10. Comparative FISH analysis on segments with a pattern of duplication status inconsistent with the human-chimpanzee-gorilla-orangutan phylogeny. *We selected human fosmid probes corresponding to two loci that showed a pattern "inconsistent" with the hominid phylogeny. A) Example of deletion/incomplete lineage sorting: FISH analysis show that chimpanzee and gorilla have two orthologous copies in common (the ancestral on chromosome 8 (yellow arrow) and derivative locus on chromosome 4 (green arrow). We then concluded that the chromosome 4 copy has been deleted in human, although there have been independent duplication events in chimpanzee and gorilla. Of course, this could also be the result of incomplete lineage sorting. B) Example of recurrent event: this site was classified as a recurrent duplication event because NONE of the derived loci are shared between human and gorilla, suggesting that duplications have occurred exclusively independently in both lineages. *Denotes a heterozygous duplication.*

8. A recurrent African great-ape duplication expansion

We characterized in detail one duplication mapping to human chromosome 10p12.31 that showed the most extreme evidence of a segmental duplication expansion in gorilla and chimpanzee but was a single copy in human. The duplication interval was approximately 153 kbp in length and consisted of two duplication blocks (block 1 was 86 kbp (chr10:19,438,000-19,526,000) and block 2 was 66 kbp in length (chr10:19,551,000-19,617,791); Fig. 3; Supplementary Note Fig. 12). Notably, we observed a ~182 kbp deletion (chr10:19,245,500-19,427,500) in chimpanzee mapping ~11 kbp upstream without any known gene in the region (FISH results, data not shown). Our goal was to estimate the copy number and map locations of the duplications in the gorilla and chimpanzee genomes. A detailed physical map would allow us to assess the orthology of chimpanzee and gorilla copies and address whether the duplicated copy was present in the ancestor or if it was created independently in every lineage. We used two different approaches to map the locations of the duplications: end-sequence pair mapping and FISH.

8.1 Fine-scale mapping of African ape duplication loci using end-sequence pairs

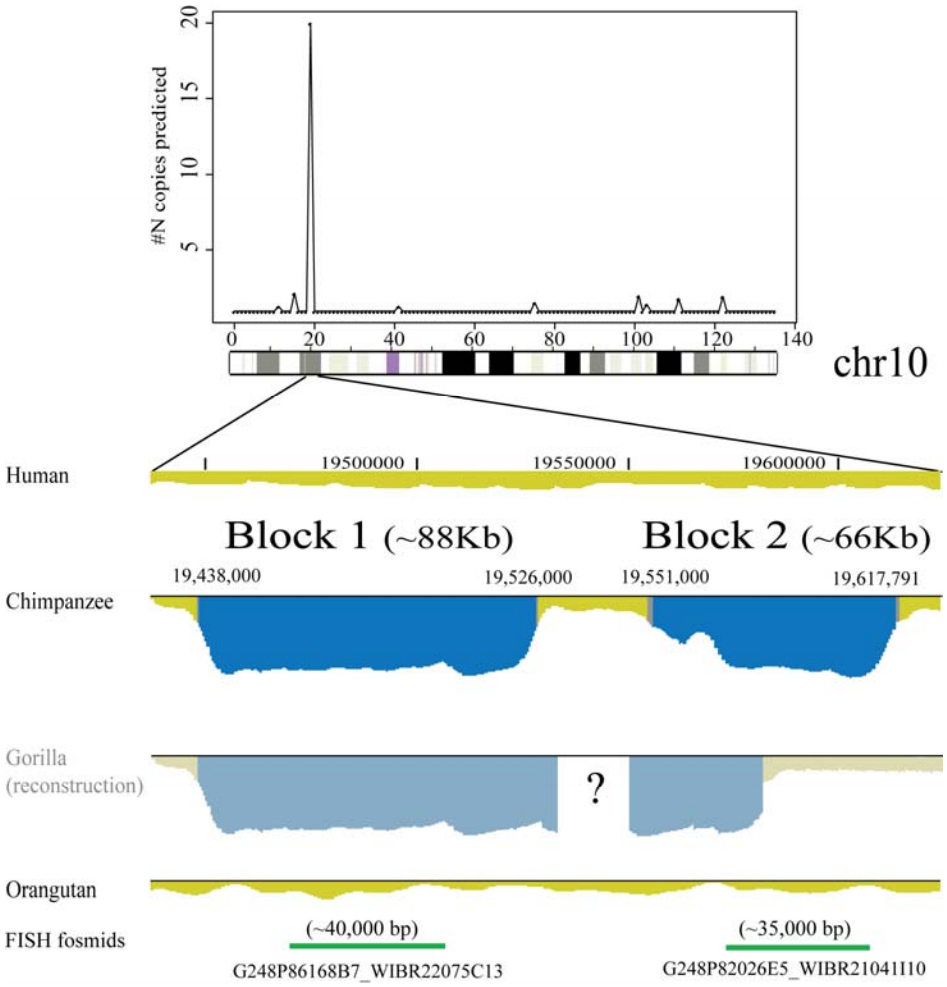
We used chimpanzee and gorilla paired end-sequences to determine if chimp and gorilla loci were orthologous (as described above in section 7.1; we considered only those map locations with unique anchors and required two or more independent clone ESPs to define a map location). Using end-sequence pairs from 15.22 million chimpanzee plasmids and 3.96 million gorilla plasmids, we identified 21 chimpanzee and 4 gorilla map locations for the block 1 duplication interval. With the exception of the ancestral locus on chromosome 10, none of the map locations were orthologous between the species. We repeated the analysis for block 2 and identified 11 chimpanzee and 199 gorilla map locations. Once again none of these were orthologous, with the exception of the ancestral locus (Supplementary Note Fig. 11). We compared block 1 and block 2 locations and found that only 6 in chimpanzee and 1 in gorilla were shared, suggesting largely independent duplications of both blocks in both species.



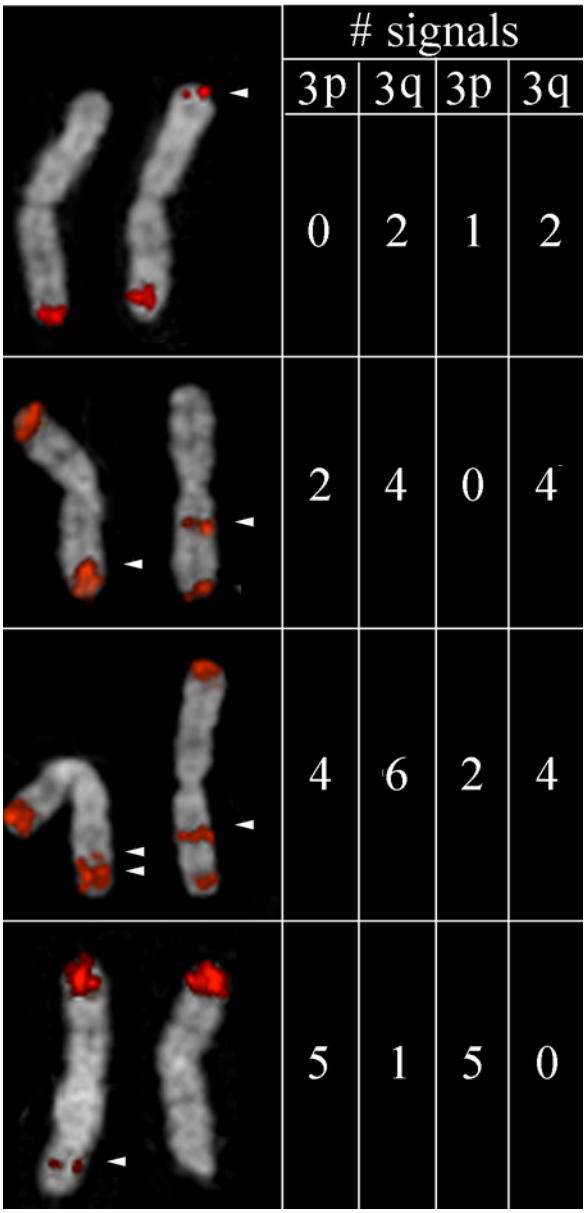
Supplementary Note Fig. 11. Gorilla and chimpanzee Chr10 SD integration sites based on ESP mapping. *The ancestral locus is indicated by the red arrow (note: overlapping positions at the ancestral locus have been separated for display purposes).*

8.2 FISH analysis

As a final confirmation, we selected two human fosmid clones (Table S11) and performed FISH on chromosomal metaphase preparations from gorilla ("GGO13") and chimpanzee ("Clint") lymphoblastoid cell lines. Within the limits of metaphase FISH resolution, the results were in agreement with the ESP mapping data, suggesting independent duplications in both lineages. Although hybridization signals were enriched near the ends of primate chromosomes, we observed numerous signals within interstitial, euchromatic regions. Due to the extraordinary number of block 2 duplications observed within the gorilla lineage, we examined additional gorilla lymphoblastoid lines (Coriell AG20600, "GGO5" and "GGO8") using a fosmid probe (WIBR2-1041I10). Surprisingly, we found evidence of copy-number polymorphism, as well as, variation in segmental duplication locations between different homologous chromosomes (see Figure 5c for chromosome 1 and Supplementary Note Fig. 13 for chromosome 3). Such extensive variation in duplication number and map location is without precedent in studies of hominoid evolution and suggests continued segmental duplication within this particular great-ape lineage.



Supplementary Note Fig. 12. Schematic of the location of the FISH probes with respect to the duplications.



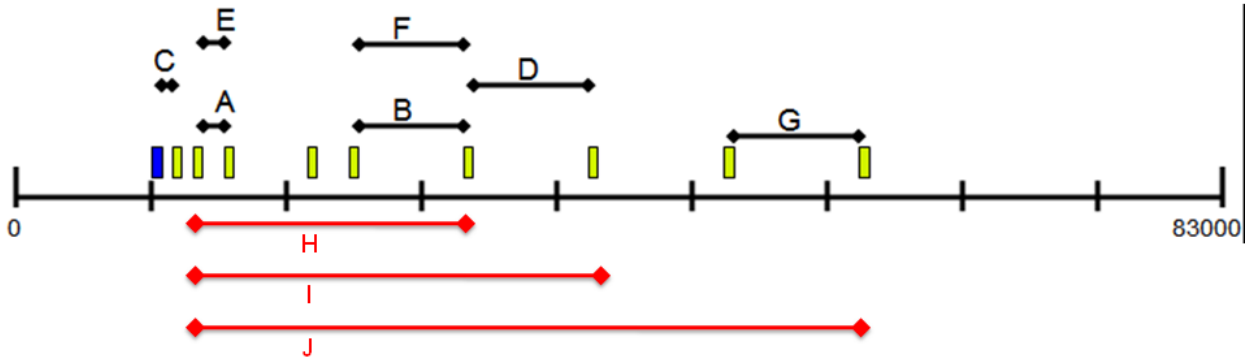
Supplementary Note Fig. 13. Gorilla CNP and variation in location of SD. *Four different gorillas were tested by FISH for the presence of copies of chr10 expansion (block 2) on chromosome 3. As seen in Fig. 3c, homologous chromosomes from the same individual show different pattern of copies. From top to bottom: "AG20600", "GGO5", "GGO8" and "GGO13".*

8.3 Gene characterization and RT-PCR

We performed a BLASTX nucleotide sequence similarity search to identify potentially uncharacterized proteins within the block 1 and block 2 duplications. Although the UCSC genome browser showed no RefSeq gene models in the region, our search identified a gene

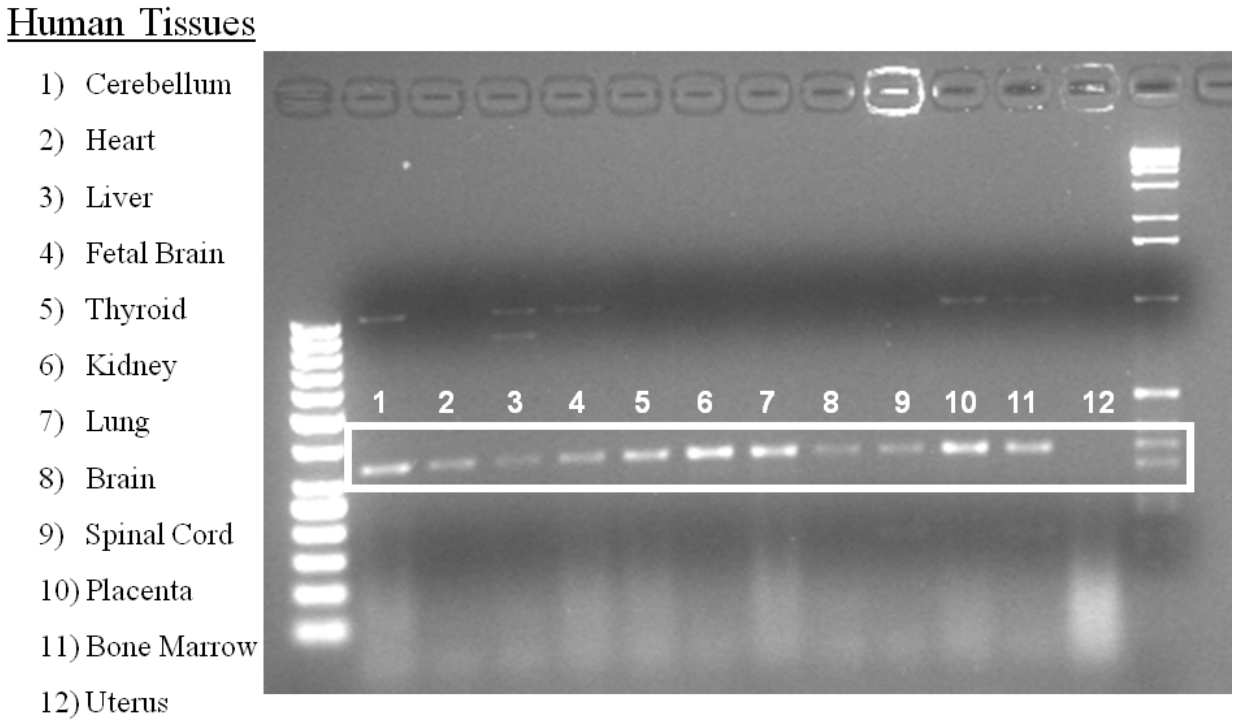
model within block 1 with sequence similarity to other predicted genes (apical early endosomal glycoprotein precursor in macaque (ref|XP_001095292.1, 386/414 amino acids (a.a.) (93% Identity)) and chimp (ref|XP_001171694.1, 56/287 a.a. (89% Identity)) and to a novel MAM domain containing protein in humans (279/283 a.a. (98% Identity)). A subsequent scan of the region using *ab initio* gene prediction software (FGENES[35] and GENSCAN[36]) predicted a 10 exon gene model within the duplicated region from a larger model of 14 exons. Using SPIDEY[37], we constructed a consensus gene model and designed seven RT-PCR primer sets to span across various exons (Supplementary Note Fig. 14). We performed RT-PCR against cDNA derived from 12 chimpanzee tissues: cortex, brain, medulla, cerebellum, brain stem, heart, kidney, liver, lung, muscle, ovary and testis (courtesy of J. Rogers South Western Primate Center). We observed products of the expected size (based on the gene model predictions) in all tissues for all primer pairs with the exception of primer set C.

We repeated the analysis using human cDNA synthesized from the following 12 human tissues: cerebellum, heart, liver, fetal brain, thyroid, kidney, lung, brain, spinal cord, placenta, bone marrow and uterus. RT- PCR products were observed in all of the tissues except uterus. Sequencing of subcloned RT-PCR products from chimpanzee confirm the gene model and suggest that transcription is limited to relatively few sites with most transcripts consistent with chromosome 10 ancestral gene model. However, since the chimpanzee duplications have not yet been characterized at the sequence level and have occurred relatively recently with a high degree of sequence identity, it is unclear whether both ancestral and derivative copies are expressed. Nevertheless, these results confirm mRNA transcription and suggest a previously uncharacterized gene model within the block 1 duplication. The significance of this gene with respect to the independent duplications is unknown although the data are consistent with a gene family expansion in chimpanzee and gorilla that is single copy in humans and other apes.



Supplementary Note Fig. 14. Gene model prediction (boxes indicate exons) within the block 1.

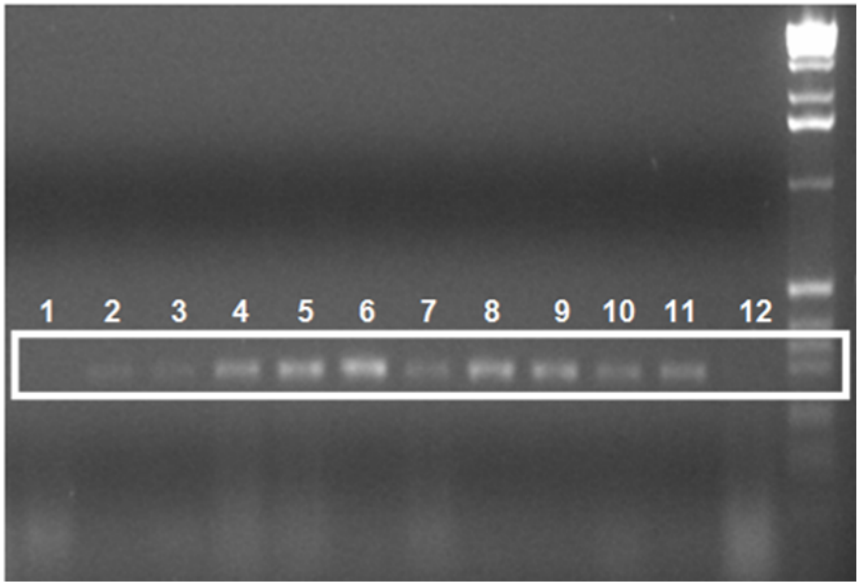
RT-PCR assays (A-J) are indicated by horizontal lines.



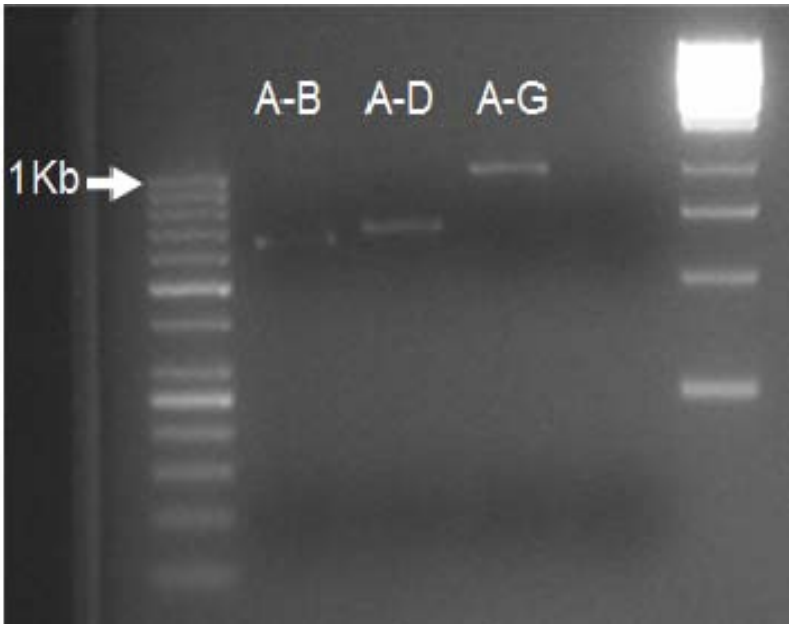
Supplementary Note Fig. 15. Human RT-PCR results using assay G.

Chimp Tissues

1) Cortex
2) Brain
3) Medulla
4) Cerebellum
5) Brain Stem
6) Heart
7) Kidney
8) Liver
9) Lung
10) Muscle
11) Ovary
12) Testis



Supplementary Note Fig. 16. Chimp RT-PCR results using assay G.



Supplementary Note Fig. 17. RT-PCR assay H, I and J on chimp brain tissue.

9. Rates of Duplication

9.1 Rates of duplication (Mbp)

Using the generally accepted phylogeny of primates, we assigned duplicated regions to either terminal branches or ancestral nodes based on our cross-species categorization. For each species we estimated the number of duplicated Mbp based on the depth of coverage of WGS reads (Supplementary Note Section 1.2) and used this number to estimate effective accumulation of segmental duplications within each lineage. The amount of shared duplication (assigned to ancestral nodes) was estimated similarly but was quantified independently for each species (Fig. 4). With few exceptions (Fig. 3), estimates of shared duplication were remarkably similar between the different species. Interspecific arrayCGH results from bonobo and gorilla DNA hybridizations were used to further recategorize duplications within the primate phylogeny. For example, if a human-specific SD validated by human/chimpanzee arrayCGH was found to show no significant difference by human/gorilla arrayCGH, the segmental duplication was reclassified as a human/gorilla shared SDs. Likewise, we used the gorilla/orangutan arrayCGH results to reclassify a proportion of orangutan-specific SDs as shared with gorilla (i.e. both showed gains in signal intensity with respect to human). Using 6, 8, 12 and 25 million years for the divergence of humans from chimpanzee, gorilla, orangutan and macaque species, as well as 1-2 million years for the separation of bonobo and chimpanzee, we calculated the effective rates of duplication for each branch. We repeated the analysis using estimates of genetic distance (number of substitutions per 1000 bp) in lieu of millions of years divergence[38]. The analysis revealed two interesting features: 1) there has been a 2- to 5-fold increase in segmental duplications in great-ape/human when compared to macaque; and 2) the highest rate of duplication occurs within the common ancestor of humans and African great-apes when compared to lineage-specific duplications.

Supplementary Note Table 13. Rates of segmental duplication per Myr.

	WSSD (>20 kbp)	WGAC <94% (no WSSD)	Copy number corrected	MMU WSSD corrected +WGAC (no WSSD)	Million years of the branch	Rate Duplications (Mb/Myr)
Macaque specific SDs	4,149,107	3,566,303	14,700,353	18,266,656	25	**0.731**
Great-ape SDs (human perspective)	52,916,372		55,729,091		25	**2.229**
HSA/PTR/PPY	15,973,342				13	**1.229**
HSA/PTR/GGO	16,822,662				4	**4.206**
HSA/PTR	9,319,525				2	**4.660**
HSA	10,800,843		13,613,562		6	**2.269**
Human/chimp/orang/macaque SDs	10,111,509		12,363,179			

Duplications longer than 20 Kbp from the two lineages (great apes and Old World monkeys) were compared using the human and the macaque reference genome. Duplications were detected by WSSD, and for macaque, we conservatively added more divergent predictions (Gibbs et al. 2007) from the macaque assembly. Duplications were corrected for copy number according to the depth of coverage of WGS reads. Human SDs were refined with arrayCGH information from bonobo and gorilla.

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.

9.2 Rates of duplication (events)

As a surrogate for actual duplication events, we separately considered the ancestral nonredundant set of chained duplicons (n=950; >20 kbp in length)[11]. This ancestral nonredundant set of duplications (>90% sequence identity) may be considered to represent evolutionary distinct duplication events that have occurred specifically within the human lineage. Based on our computational and experimental predictions, we assigned each of these 950 duplicons to different timepoints in the primate phylogeny and estimated the rate of duplication per million years or as a function of genetic distance (substitutions per 1000 bp). These data similarly support a burst of duplication events during the time of African great ape and human speciation (6-8 million years ago; Fig. 4c).

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].

9.3 Maximum likelihood model

We developed a maximum likelihood framework to model the accumulation of SDs in the primate lineage and to assess rates of SD accumulation in primates taking into account our estimates of recurrent mutation (homoplasy). We aimed to resolve two issues. First, is the rate of SD accumulation in the MMU branch slower than the rate in the great apes? And second, is the rate of SD accumulation in the common ancestor of human, chimpanzee and gorilla larger than lineage-specific estimates?

We measured the amount of SDs in every branch of the phylogeny in either of two units: SD events/genome or SD bps/genome. Branch length was also measured in two ways, either using estimates of the length of each branch in millions of years, or calibrating branch lengths with substitution rates in the nucleotide sequence. In the latter case, our time unit would be the number of substitutions per 1000 bp in each branch as determined in a previous analysis[38]. The latter time unit has two advantages: it eliminates the need to use imprecise time inferences from the fossil record and, in addition, allows for direct comparison with sequence divergence rates.

First, we used our primate SD map together with parsimony criteria to estimate where within the primate phylogeny did any given SD arise (Supplementary Note Table 2). By using maximum parsimony alone, we would be assuming that any SD that is shared by two given sister branches occurred within the common ancestor. Such an assumption, however, ignores homoplasy (i.e. it is possible that any shared SD appeared twice, once in each branch, instead of only once in the common ancestor). To account for homoplasy we incorporated a recurrent mutation rate of ~20% (e.g. an upper bound based on our estimation of 15–16% (Supplementary Note Section 7)) to parsimoniously assigned duplications and we recursively corrected the assignments of SDs according to this estimate. We consider this treatment conservative as it will tend to "erase" the signature of any SD accumulation burst within internal branches. Results are presented in Supplementary Note Table 16.

Supplementary Note Table 16. Adjusted rates incorporating 20% of homoplasy.

Branches	Mbs (after correction)	Events	Myrs	subts/1000bps
HSA	**19,480,933**	210	6	5.40
PTR(PTR+PPA)	**12,021,849**	232	6	5.56
PPY	**23,524,428**	434	12	15.03
HSA/PTR/GGO/PPY	**12,778,674**	170	13	14.72
HSA/PTR+ HSA/PTR/GGO	**23,469,484**	307	6	8.69
MMU	**18,266,656**	499	25	29.44

The second step was to build a model of SD accumulation and a Likelihood Ratio Test (LRT). Given that the topology of the tree and the lengths of branches are fixed (in whatever time unit we choose to use), the observations render well to a likelihood model that considers the amount of SDs in every branch and the observation for which we need to compute a probability. We are considering accumulation of SDs in branches, a phenomenon that results from the interaction between many different complex factors but can be modelled as a pure birth process. The probability of observing a certain number, α_i, of SDs in branch *i* will be given by a Poisson distribution with parameter $\lambda_i t_i$, where, λ_i is the rate of duplication per unit time in that branch and t_i is its length.

$$\frac{e^{-\lambda_i t_i}\left(\lambda_i t_i\right)^{\alpha_i}}{\alpha_i}$$

The simplest model for the accumulation of SDs in the primate lineage assumes a single duplication rate, λ, over the entire tree. In that case, the likelihood of our observation is given by the product of the Poisson probabilities across all branches:

$$L = \prod_i \frac{e^{-\lambda t_i}\left(\lambda t_i\right)^{\alpha_i}}{\alpha_i}$$

A second model with an extra parameter considers two rates, λ_T for the branch or branches we want to test and λ_R for the rest of the tree. The likelihood is now given by a very similar expression:

$$L = \prod_i^R \frac{e^{-\lambda_R t_i}\left(\lambda_R t_i\right)^{\alpha_i}}{\alpha_i} \times \prod_j^T \frac{e^{-\lambda_T t_i}\left(\lambda_T t_i\right)^{\alpha_j}}{\alpha_j}$$

We can test whether the second model explains our observations better that the first using a Likelihood Ratio Test. Two times the difference between the two log likelihoods will be asymptotically distributed as a Chi-Square with 1 degree of freedom (since we are estimating an extra parameter in the second model). The results of our two tests are presented in Supplementary Note Table 17. Each test has been carried out several times using different duplication and time units.

Results are highly significant and indicate that Old World monkeys have slower rates of duplications than great apes and that, within great apes, there was a burst of duplication activity in the common ancestor of humans and chimpanzees.Our prediction is in strong agreement with the degree of sequence divergence among human intrachromosomal segmental duplications that shows a mode at 97-99% sequence identity. We note that this burst of duplication activity corresponds to a time when other mutational processes, such as point substitutions and retrotransposon activity, were slowing along the hominoid lineage.

One possibility for this dichotomy may be reduction in the effective population size of primate hominid populations as has been proposed recently for the burst of nuclear mitochondrial insertion sequences at the prosimian-anthropoid divergence[39]. If we assume that most large segmental duplications are weakly deleterious, such variants may be disproportionately fixed as a result of the whims of genetic drift as opposed to being eliminated by purifying selection in a large effective population size. Such an excess of deleterious mutations has been seen in certain cases, such as gene control regions in comparisons between human and chimpanzees[40] or, at smaller scale, in human populations that underwent bottleneck[41].

Supplementary Note Table 17. Summary of the likelihood estimates of SD accumulation rates and LRT test results.

Accumulation of SD events. Time unit =>Myrs			
	Model 1 (all identical rate)	Model 2 (two different rates)	P-values
(T1) MMU against great apes.	λ = 27.23SDEv/Myrs	λR(G. Apes) = 31.47 SDEv /Myrs λT (MMU) = 19.96 SDEv /Myrs	$<10^{-10}$
(T2) HSAPTR/GGO against the rest of great apes.	λ = 31.47 SDEv /Myrs	λRest = 28.27 SDEv /Myrs λT (HSAPTRandGGO) = 51.17 SDEv /Myrs	$<10^{-10}$
Accumulation of SD Mbs. Time unit =>Myrs			
	Model 1 (all identical rate)	Model 2 (two different rates)	P-values
(T1) MMU against great apes.	λ = 1.61 Mbs//Myrs	$\lambda_{R(G.\ Apes)}$ = 2.12 Mbs/Myrs $\lambda_{T\ (MMU)}$ = 0.73 Mbs/Myrs	$<10^{-10}$
(T2) HSAPTR/GGO against the rest of great apes.	λ = 2.12 Mbs/Myrs	λ_{Rest} = 1.18 Mbs/Myrs $\lambda_{T\ (HSAPTRandGGO)}$ = 3.92 Mbs/Myrs	$<10^{-10}$
Accumulation of SD events. Time unit =>subst/1000 bps			
	Model 1 (all identical rate)	Model 2 (two different rates)	P-values
(T1) MMU against great apes.	λ = 23.49 SDEv /subst1000	λR(G. Apes) = 27.39 SDEv /subst1000 λT (MMU) = 16.95 SDEv /subst1000	$<10^{-10}$
(T2) HSAPTR/GGO against the rest of great apes.	λ = 27.39 SDEv /subst1000	λRest = 25.69 SDEv /subst1000 λT (HSAPTRandGGO) = 35.32 SDEv /subst1000	1.9231x10-6
Accumulation of SD Mbs. Time unit =>subst/1000 bps			
	Model 1 (all identical rate)	Model 2 (two different rates)	P-values
(T1) MMU against great apes.	λ = 1.39 Mbs/subst1000	$\lambda_{R(G.\ Apes)}$ = 1.85 Mbs/subst1000 $\lambda_{T\ (MMU)}$ = 0.63 Mbs/subst1000	$<10^{-10}$
(T2) HSAPTR/GGO against the rest of great apes.	λ = 1.85 Mbs/subst1000	λ_{Rest} = 1.67 Mbs/subst1000 $\lambda_{T\ (HSAPTR\ and\ GGO)}$ = 2.70 Mbs/subst1000	$8.1507x10^{-10}$

Two tests were performed. First, the rate in the MMU branch is compared with that of great apes. Second, the macaque was removed and the rate in the branch corresponding to the common ancestor of humans, chimpanzees and humans, chimpanzees and gorillas was compared with the rest of the tree.

References

1. Bailey, J.A. et al. Recent segmental duplications in the human genome. *Science* **297**, 1003-7 (2002).
2. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7**, 203-14 (2000).
3. Cheng, Z. et al. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437**, 88-93 (2005).
4. Smit, A., Hubley, R & Green, P. RepeatMasker Open-3.0. (1996-2004).
5. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-80 (1999).
6. Liu, G. et al. Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res* **13**, 358-68 (2003).
7. Consortium, C.S.a.A. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87 (2005).
8. Gibbs, R.A. et al. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**, 222-34 (2007).
9. Jiang, Z., Hubley, R., Smit, A. & Eichler, E.E. DupMasker: A tool for annotating primate segmental duplications. *Genome Res* **18**, 1362-8 (2008).
10. She, X. et al. A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications. *Genome Res* **16**, 576-83 (2006).
11. Jiang, Z. et al. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat Genet* **39**, 1361-8 (2007).
12. Wheeler, D.A. et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-6 (2008).
13. Levy, S. et al. The diploid genome sequence of an individual human. *PLoS Biol* **5**, e254 (2007).
14. Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444-54 (2006).
15. Sharp, A.J. et al. Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* **77**, 78-88 (2005).
16. Iafrate, A.J. et al. Detection of large-scale variation in the human genome. *Nat Genet* **36**, 949-51 (2004).
17. Kidd, J.M. et al. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64 (2008).
18. Bentz, M., Dohner, H., Cabot, G. & Lichter, P. Fluorescence in situ hybridization in leukemias: 'the FISH are spawning!'. *Leukemia* **8**, 1447-52 (1994).
19. Fortna, A. et al. Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol* **2**, E207 (2004).
20. Dumas, L. et al. Gene copy number variation spanning 60 million years of human and primate evolution. *Genome Res* **17**, 1266-77 (2007).
21. Yang, Z. & Swanson, W.J. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol Biol Evol* **19**, 49-57 (2002).
22. Pond, S.L. et al. Adaptation to different human populations by HIV-1 revealed by codon-based analyses. *PLoS Comput Biol* **2**, e62 (2006).

23. Yang, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* **15**, 568-73 (1998).
24. Nickel, G.C., Tefft, D. & Adams, M.D. Human PAML browser: a database of positive selection on human genes using phylogenetic methods. *Nucleic Acids Res* **36**, D800-8 (2008).
25. Bakewell, M.A., Shi, P. & Zhang, J. More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc Natl Acad Sci U S A* **104**, 7489-94 (2007).
26. Kosiol, C. et al. Patterns of positive selection in six Mammalian genomes. *PLoS Genet* **4**, e1000144 (2008).
27. Engel, K., Zhou, M. & Wang, J. Identification and characterization of a novel monoamine transporter in the human brain. *J Biol Chem* **279**, 50042-9 (2004).
28. Lupski, J.R. An evolution revolution provides further revelation. *Bioessays* **29**, 1182-4 (2007).
29. Popesco, M.C. et al. Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains. *Science* **313**, 1304-7 (2006).
30. de Vries, B.B. et al. Diagnostic genome profiling in mental retardation. *Am J Hum Genet* **77**, 606-16 (2005).
31. Westhoff, C.M. & Wylie, D.E. Investigation of the RH locus in gorillas and chimpanzees. *Journal of Molecular Evolution* **42**, 658-668 (1996).
32. Johnson, M.E. et al. Recurrent duplication-driven transposition of DNA during hominoid evolution. *Proc Natl Acad Sci U S A* **103**, 17626-31 (2006).
33. Newman, T.L. et al. A genome-wide survey of structural variation between human and chimpanzee. *Genome Res* **15**, 1344-56 (2005).
34. Tuzun, E. et al. Fine-scale structural variation of the human genome. *Nat Genet* **37**, 727-32 (2005).
35. Salamov, A.A. & Solovyev, V.V. Ab initio gene finding in Drosophila genomic DNA. *Genome Res* **10**, 516-22 (2000).
36. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**, 78-94 (1997).
37. Wheelan, S.J., Church, D.M. & Ostell, J.M. Spidey: a tool for mRNA-to-genomic alignments. *Genome Res* **11**, 1952-7 (2001).
38. Elango, N., Thomas, J.W. & Yi, S.V. Variable molecular clocks in hominoids. *Proc Natl Acad Sci U S A* **103**, 1370-5 (2006).
39. Gherman, A. et al. Population bottlenecks as a potential major shaping force of human genome architecture. *PLoS Genet* **3**, e119 (2007).
40. Keightley, P.D., Lercher, M.J. & Eyre-Walker, A. Evidence for widespread degradation of gene control regions in hominid genomes. *Plos Biology* **3**, 282-288 (2005).
41. Lohmueller, K.E. et al. Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**, 994-U5 (2008).