

Table of Contents

Supplemental Section S1 – Genome Sequencing and Assembly	2
Supplemental Section S2 – Indel Assessment With the Neutral Indel Model	11
Supplemental Section S3 – Great Ape Divergence Estimate via WGS Read Mapping..	11
Supplemental Section S4 – Short Read Sequencing.....	13
Supplemental Section S5 – The Orangutan Ensembl Gene Set	13
Supplemental Section S6 – Ancestral Reconstruction	15
Supplemental Section S7 – The Genomic Distribution of Genic Evolution Rates	18
Supplemental Section S8 – Cytogenetic Characterization	22
Supplemental Section S9 – High-copy Repeat Assessment	28
Supplemental Section S10 – Processed Pseudogene Formation.....	31
Supplemental Section S11 – Segmental Duplications and Structural Variation.....	33
Supplemental Section S12 – Great Ape Gene Family Expansion	39
Supplemental Section S13 – Protease Gene Families	41
Supplemental Section S14 – Evolution of Orangutan Alpha and Theta Defensins.....	52
Supplemental Section S15 – Genic Positive Selection.....	57
Supplemental Section S16 – Bornean/Sumatran Divergence.....	62
Supplemental Section S17 – HMM Estimate of Bornean/Sumatran Divergence Time, Speciation Time and Effective Population Size	63
Supplemental Section S18 – Bornean/Sumatran Duplication Comparison	64
Supplemental Section S19 – Retroelement Polymorphisms	70
Supplemental Section S20 – SNP Calling and Ancestral Base Reconstruction	72
Supplemental Section S21 – Demographic Inference Using DaDi	72
Supplemental Section S22 – References.....	86

Supplemental Section S1 – Genome Sequencing and Assembly

DNA Resources for the Orangutan Assembly

The *Pongo abelii* whole-genome shotgun (WGS) data began initially with DNA from Susie (Studbook #1044; ISIS #71), a female Sumatran orangutan housed at the Gladys Porter Zoo, Brownsville, TX, obtained courtesy of Dr. Greely Stones.

Assembly Input Data

The orangutan genome was sequenced to 5.62x (\geq Q20 Phred bases) depth in Sanger (ABI 3730) reads in a combination of plasmid, fosmid and BAC end reads (Table S1-1). The BAC library (CHORI-276) from which BAC end sequences were produced was constructed in Pieter de Jong's lab by Yuko Yoshinaga using Susie's DNA.

Table S1-1. Input read statistics.

Read Type	Insert Size (kb)	Reads (million)	>Phred20 Bases (Gb)	Seq. Coverage (x)	Phys. Coverage (x)
Plasmid (WU/BCM)	4	22.90	17.81	5.40	12.11
Fosmid (WU)	40	0.88	0.56	0.17	3.99
BAC (WU)	180	0.25	0.16	0.05	5.35
Total		24.03	18.53	5.62	21.45

The Q20 length distribution for plasmid, fosmid and BAC end reads reflects a bimodal distribution for plasmid reads due to a technical issue at the time reads were produced at WashU (Figure S1-1).

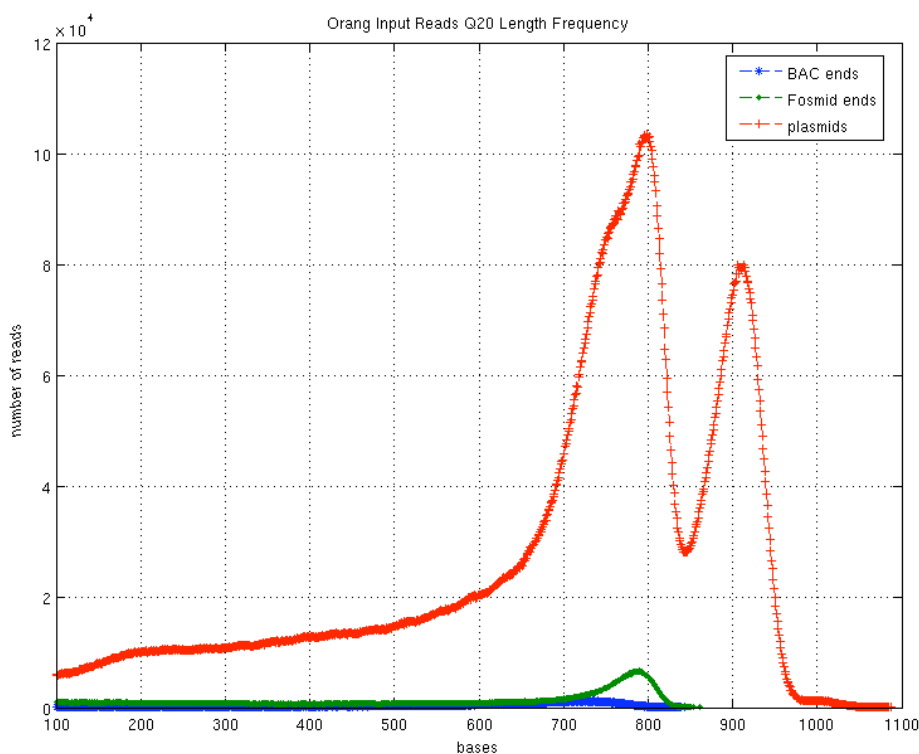


Figure S1-1. Q20 read length distribution for all input reads.

The Assembly Process

Reads were assembled using PCAP¹ with ~300 parallel PCAP jobs run on a cluster of dual processor AMD Opteron blades (AMD, Sunnyvale, CA, USA) with 2-8 Gb RAM. In addition to the CPUs mentioned above, some PCAP steps required high-memory architecture (HP Itanium with 96+ Gb RAM). The chaff rate of the assembly was ~10% (~2.7 M unplaced reads out of ~26.8 M raw input reads). The raw orangutan assembly (v2.0) then underwent several additional steps of maturation including contamination screening, small (<1 kb) contig removal, quality assessment and “A Golden Path” (AGP) creation prior to final release (v2.0.2, a.k.a. ponAbe2).

Assembly Quality Assessment - Confirming Read Origin

To confirm the orangutan reads from BCM and WashU were generated from the same individual (Susie), we checked the origin of reads underlying high quality discrepancy sites within the assembly. High quality discrepancies are potential heterozygous SNPs that occur in otherwise high quality sequence. As shown in Table S1-2, of the 1.4 million high quality discrepancies, 96% were supported by reads from both centers, with the remainder likely due to areas of low coverage.

Table S1-2. High quality discrepancies and Center origin.

Type	Number	Percentage
BCM-only	28264	1.87
WashU-only	38214	2.53
Mixed	1441380	95.59

We then checked if reads from both Centers showed heterozygosity at the 1.4 M high quality discrepancy sites, or whether the reads from WashU or BCM were biased for a single allele. As depth increased (and thus greater opportunity to observe heterozygosity) the percentage of sites with heterozygous read support from both centers also increased (Figure S1-2). These data suggest the assembly represents the genome of a single individual.

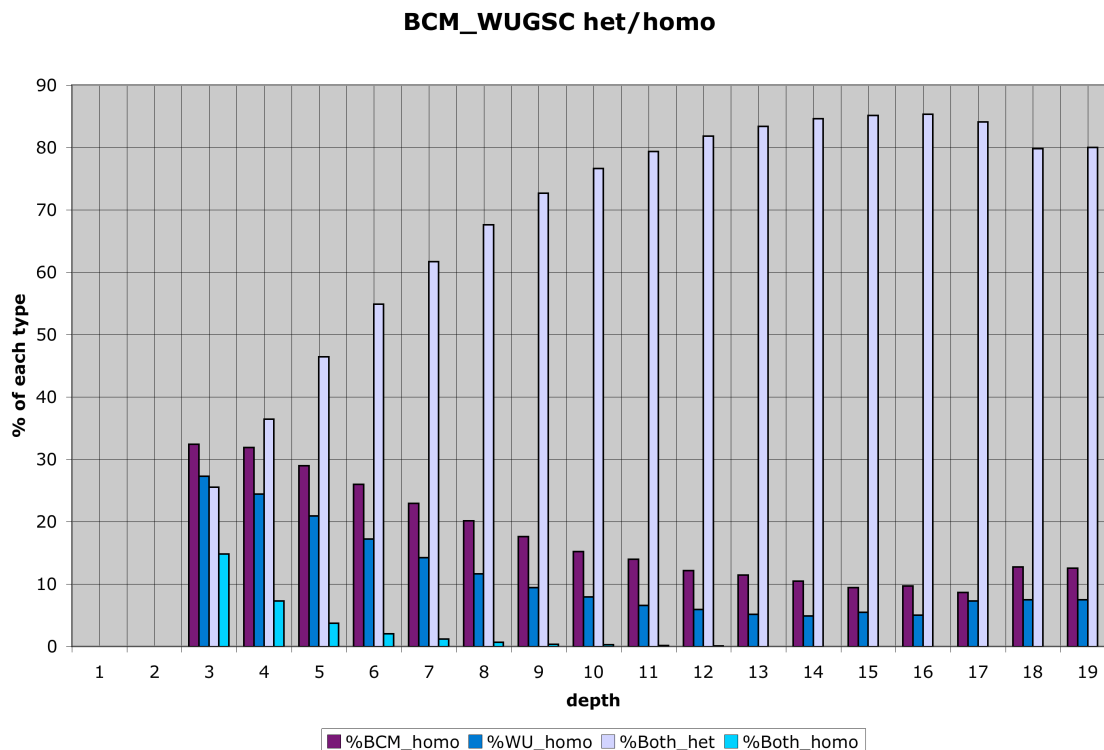


Figure S1-2. High quality discrepancies appear in reads from both Centers.

Assembly Quality Assessment - Coverage

To assess coverage, we first looked for the presence of known orangutan mRNAs within the assembly. A set of 4,667 *P. abelii* mRNA sequences² were obtained from the MIPs cDNA Consortium Group for this purpose (http://mips.gsf.de/proj/cdna/Sites/PP_cDNA_Database.htm). Using BLAT³ in client/server mode with default parameters, an alignment to the genome was found for 4,661 of the 4,667 mRNA sequences, with 95% of the 4,667 sequences aligning over $\geq 90\%$ of their length and 98% aligning over $\geq 80\%$ of their length. We also searched our set of 1,520,309 EST sequences obtained on the 454 platform (Supplemental Section S5). Of those, 197,839 (8.7%) did not align to the reference (less the mitochondrial sequence). As an additional estimate of the coverage, we downloaded the UCSC orangutan/human alignments^{4,5,6} (ponAbe2.hg18.all.gz) and found that 93% of the human genome was spanned by an alignment to the orangutan chromosomal sequences. Using the reciprocal best alignments (furnished by B. Raney, UCSC) used to generate our AGP files (see below), we obtained a more stringent estimate of 91% coverage on the autosomes, 90% when including the sex chromosomes. Lastly, we compared the assembly directly to finished BAC sequences to assess coverage and additional quality metrics (see below).

Assembly Quality Assessment – Comparison to Finished BACs

To assess the quality of the assembly we aligned the assembly to 83 finished CHORI-276 BAC sequences, totaling 16.8 Mb (Table S1-3), using `cross_match` (P. Green, unpublished). Note for a complete list of all finished CHORI-276 BACs generated by WashU to date (564 BACs as of November 2009), use the search term: 'Wilson [LASTAU] and Pongo and "complete sequence"' at the NCBI website.

Table S1-3. Finished *Pongo abelii* BAC clones used for quality assessment.

Accession	Accession	Accession
gi 102140852 gb AC186758.1	gi 109716020 gb AC188104.1	gi 110624889 gb AC189036.1
gi 102140881 gb AC186759.1	gi 109716021 gb AC188105.1	gi 110665861 gb AC189101.1
gi 102140917 gb AC186760.1	gi 109716022 gb AC188106.1	gi 111120377 gb AC189854.1
gi 102140946 gb AC186761.1	gi 109716023 gb AC188107.1	gi 111120378 gb AC189855.1
gi 102140961 gb AC186762.1	gi 109716024 gb AC188108.1	gi 111120379 gb AC189856.1
gi 102140965 gb AC186763.1	gi 109716025 gb AC188109.1	gi 111120381 gb AC189857.1
gi 102140967 gb AC186764.1	gi 109716026 gb AC188110.1	gi 112984641 gb AC190145.1
gi 102140970 gb AC186765.1	gi 109716027 gb AC188111.1	gi 112984643 gb AC190146.1
gi 102140972 gb AC186766.1	gi 109716028 gb AC187347.2	gi 112984658 gb AC190154.1
gi 102140973 gb AC186767.1	gi 109716029 gb AC188112.1	gi 113462168 gb AC187292.2
gi 102140975 gb AC186768.1	gi 109716030 gb AC188113.1	gi 113462169 gb AC187771.2
gi 102140977 gb AC186769.1	gi 109716031 gb AC188114.1	gi 113462170 gb AC188394.3
gi 102140978 gb AC186770.1	gi 109716032 gb AC188115.1	gi 113700380 gb AC187575.2
gi 102140980 gb AC186771.1	gi 109716033 gb AC188116.1	gi 113700381 gb AC190402.1
gi 102140982 gb AC186772.1	gi 109716034 gb AC187293.2	gi 113930804 gb AC190147.2
gi 102140984 gb AC186773.1	gi 109716035 gb AC188117.1	gi 113951820 gb AC187359.3
gi 105578903 gb AC187101.1	gi 109716036 gb AC188118.1	gi 113951821 gb AC190417.1
gi 109716009 gb AC188093.1	gi 109716037 gb AC187773.2	gi 114213541 gb AC139624.3
gi 109716010 gb AC188094.1	gi 109809871 gb AC188182.1	gi 114306982 gb AC191227.1
gi 109716011 gb AC188095.1	gi 110005891 gb AC188304.1	gi 114329111 gb AC158413.3
gi 109716012 gb AC188096.1	gi 110005892 gb AC188305.1	gi 114431279 gb AC158698.3
gi 109716013 gb AC188097.1	gi 110005894 gb AC188307.1	gi 114703465 gb AC138727.3
gi 109716014 gb AC188098.1	gi 110005895 gb AC188308.1	gi 115270990 gb AC158414.3
gi 109716015 gb AC188099.1	gi 110431390 gb AC144597.3	gi 115292454 gb AC138733.3
gi 109716016 gb AC188100.1	gi 110556871 gb AC187486.2	gi 115312374 gb AC191648.1
gi 109716017 gb AC188101.1	gi 110556872 gb AC188306.2	gi 94957697 gb AC138730.4
gi 109716018 gb AC188102.1	gi 110556873 gb AC188957.1	gi 94957698 gb AC158416.3
gi 109716019 gb AC188103.1	gi 110612042 gb AC189029.1	

Of the 16.8 Mb of finished BAC sequence, 16.3 Mb (97%) were covered by `cross_match` alignments with the orangutan assembly. These alignments revealed a substitution rate of 9×10^{-4} and an insertion/deletion (indel) rate of 2×10^{-4} . The substitution rate is consistent with the heterozygosity rate, as determined by PCR-based sequence sampling from Susie's genomic template DNA (35,075 bases assayed; 32 heterozygous sites identified; 1 het site/1,096 bp or 9×10^{-4}). Compared to the heterozygosity rate obtained for the chimpanzee named Clint at the time of the chimpanzee genome project (89,444 bases assayed; 66 heterozygous sites identified; 1 het site/1355 bp or 7×10^{-4}), we obtained a notably higher heterozygosity rate for Susie.

We also investigated in further detail instances where two contigs or supercontigs aligned to the same region of a finished BAC sequence, which is indicative of allelic sites remaining separated in the assembly (see below). In terms of the structural integrity of the assembly (the order and orientation of contigs) with respect to finished BAC sequences, we noted some small supercontigs (most <5 kb) were not positioned within larger supercontigs (<1 event per 100 kb). While these are not strictly errors, they

do affect overall assembly statistics. There are also small, undetected overlaps (most <1 kb) between consecutive contigs (~1.7 events per 100 kb), occasional local mis-ordering of small contigs (~0.1 events per 100 kb), and small contigs incorrectly inserted within larger supercontigs (~0.5 events per 100 kb). Overall, the rate of rearrangements with respect to finished BACs was comparable to previous WGS assemblies.

Assembly Heterozygosity Validation – PCR of Selected Regions

From our comparisons to finished clones, we identified regions where two independent supercontigs aligned to the identical region of a BAC sequence (putative allelic supercontigs). We designed 13 PCR amplicons targeting four sample regions containing 51 putative heterozygous SNPs, and we amplified and sequenced these products from both genomic and cell-line template DNA. In summary, all variants (where sequence quality allowed a call to be made) appeared to be from normal allelic variation and we saw no differences based on DNA source (Table S1-4). One site (a T at position 606 within Contig 22610.6) had insufficient sequence quality to make a conclusive call.

Table S1-4. Assembly Heterozygosity Validation.

Project ID	PCR Product ID	Heterozygous Sites	Homozygous Sites	Contigs tested	
PPAA-C1	PCR2g3 genomic	2	0	Contig2244.35-38/ Contig22610.2-7	
PPAA-C1	PCR2c3 cell line	2	0		
PPAA-C1	PCR5g3 genomic	2	0		
PPAA-C1	PCR5c3 cell line	2	0		
PPAA-C1	PCR6g7 genomic	0	0		
PPAA-C1	PCR6c7 cell line	0	0		
PPAA-C1	PCR8g7 genomic	4	0		
PPAA-C1	PCR8c7 cell line	3	0		
PPAA-C1	PCR9g10 genomic	1	0		
PPAA-C1	PCR9c10 cell line	3	1*		
PPAA-C2	PCR1g2 genomic	9	0		Contig168.55-59/ Contig34348.1-2
PPAA-C2	PCR1c2 cell line	9	0		
PPAA-C2	PCR5g6 genomic	8	0		
PPAA-C2	PCR5c6 cell line	8	0		
PPAA-C2	PCR9g10 genomic	5	0		
PPAA-C2	PCR9c10 cell line	5	0		
PPAA-C3	PCR1g2 genomic	5	0	Contig168.79-84/ Contig32382.1-2	
PPAA-C3	PCR1c2 cell line	5	0		
PPAA-C3	PCR4g5 genomic	3	0		
PPAA-C3	PCR4c5 cell line	2	0		
PPAA-C3	PCR8g9 genomic	3	0		
PPAA-C3	PCR8c9 cell line	3	0		
PPAA-C4	PCR1g2 genomic	3	0	Contig353.61-64/ Contig71362.1	
PPAA-C4	PCR1c2 cell line	3	0		
PPAA-C4	PCR4g6 genomic	4	0		
PPAA-C4	PCR4c6 cell line	4	0		

* T at 606 in Contig22610.6

Sequence Quality Assessment – Indels

Two sets of indels were investigated in the quality assessment phase of the project. The first set was identified by a conservative ortholog calling pipeline prior to screening for positive selection (Adam Siepel, personal communication and see below). The second set was based on the application of the Neutral Indel Model⁷ (Supplemental Section S2).

The Ortholog Indel Set

We identified 3311 coding sequence indels that fell in regions of orthology between the orangutan and human genomes (Adam Siepel personal communication). When comparing this number to similar ortholog sets from the macaque or chimpanzee assemblies (data not shown) only about half as many indels were identified. We then investigated the reads underlying these genic indels to understand their origin by several approaches.

First, we aligned our set of more than 1.5 million orangutan cDNA reads generated on the 454 platform against the orangutan genome and retained the best in genome alignment. Of the 3311 indels, 392 were covered by a cDNA alignment and 205 (52%) of them were confirmed as correct in the assembly based on the data from the cDNA.

Further, we aligned 20x coverage next generation sequence data from a Bornean orangutan (KB5404; Supplemental Section S4) against the assembly to identify reads that may confirm the assembly sequence. Reads were aligned using Maq⁸, a tool that does not allow for indels, and `cross_match`, one that does allow for indels. Of the 3311 indels, in 1242 cases there was a Bornean read which confirmed the Sumatran sequence (spanned by at least 5 bases) indicating the existing orangutan consensus may well be correct in these regions. In 529 cases, after low quality bases were trimmed from the “end” of the short read, the Bornean read data confirmed the Sumatran sequence (spanned by at least 5 bases). Finally, there were 497 of the indels that were within 50 bases of the contig end, suggesting they are in regions of low quality sequence. So, of the 3311, counting the above events uniquely, there are 2172 of the indels which are either within 50 bp of a contig gap or there is a Bornean read spanning the indel. Counting conservatively and using the 1242 where the entire Bornean read aligned and suggested that the current Sumatran consensus is correct, that would leave only 2,069 (3311-1242) possible indels which is more similar to the number found in comparisons of the chimpanzee assembly with the higher quality human genome (Adam Siepel personal communication).

Next generation sequence reads from KB5404 were also aligned using an alternative alignment pipeline (Supplemental Section S2). Of the 3311 indels, 852 were identified in the Bornean/Sumatran indel set defined by that analysis. The gap error rate based on their overall analysis showed an error rate in the Sumatran assembly that is 0.99 errors per kb. Of these 852 indels, the Maq analysis described in the previous paragraph identified 214 where there was a Bornean read aligned along its entire length that spanned the position of the indel suggesting those regions of the Sumatran sequence may be correct.

Finally, we manually reviewed 2872 of the 3311 indel regions. We extracted 600 bases surrounding the indel and incorporated the consensus sequence into CONSED⁹. We then assembled the corresponding region from the human genome (Hs.36) as well as the underlying ABI 3730 original orangutan data so the reviewer could view all relevant data within CONSED. In 742 of the cases, the indel was confirmed by the underlying raw orangutan data (meaning the current orangutan consensus sequence was validated as true and thus there is no indel), but in 30 of those cases there were some reads that also agreed with the human consensus in the region. In 1423 cases, all of the underlying raw orangutan data were low quality and the orangutan consensus was deemed incorrect. Finally, in 707 cases, no indel was identified; the orangutan and human sequence agreed through the region.

Consensus Quality Analysis of the Ortholog Indel Set

We wanted to see whether the consensus quality values for the bases involved in the indel regions was uniformly lower than the consensus quality of the sequence as a whole. For the 3311 possible indel locations, we excised 150 bp flanking the indel base. Of those, only 14% have an average consensus quality of 97 (the highest score possible), and only 32% had an average consensus quality >90. Figure S1-3 provides the consensus quality distribution for the WGS assemblies of the orangutan, chimpanzee and macaque genomes PCAP assemblies.

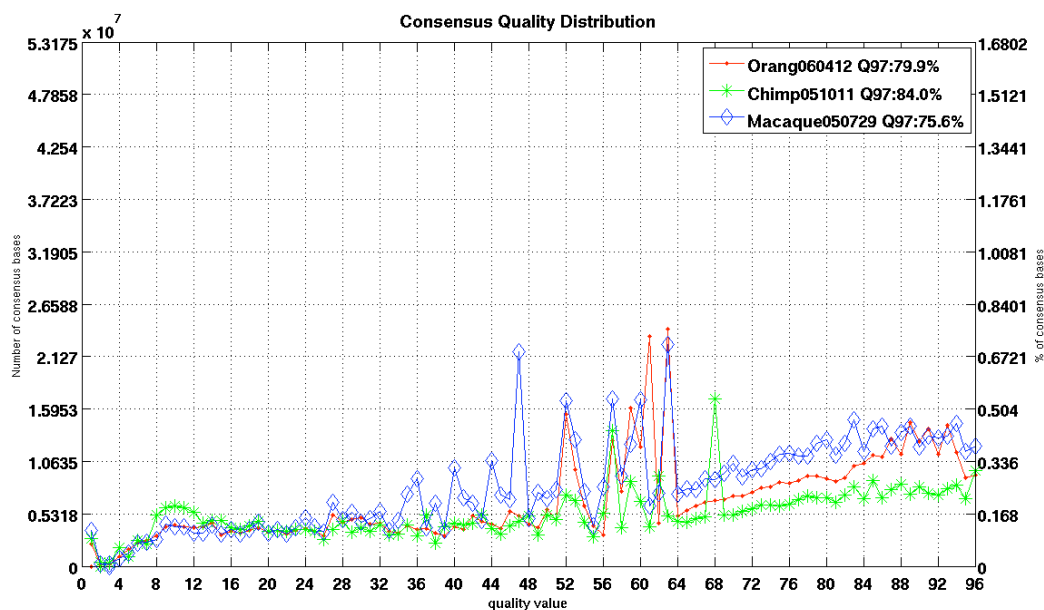


Figure S1-3. Macaque, chimpanzee and orangutan assembly consensus quality distributions.

The magenta curve in Figure S1-4 (OrangFrameShiftPoints) represents the quality of the two bases immediately flanking the indel. Here it is evident that the bases involved at indel sites are of lower quality in general.

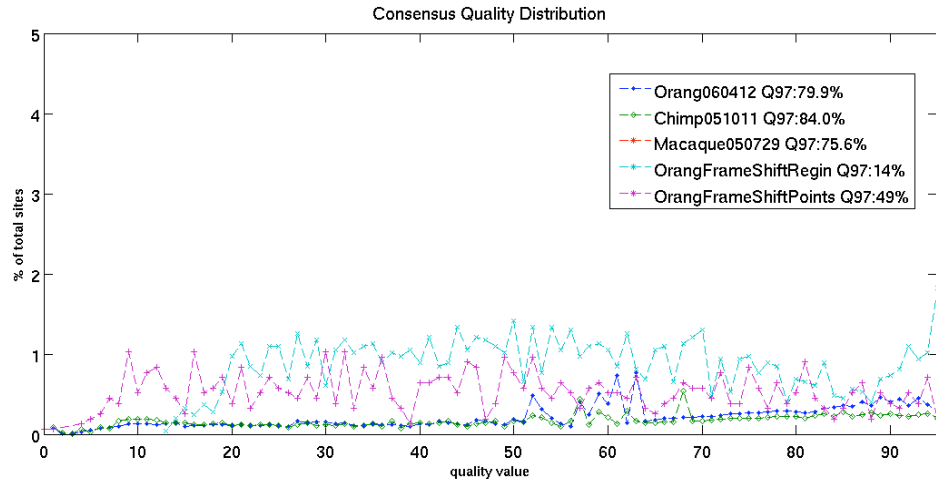


Figure S1-4. Indel sites are enriched in low consensus quality regions.

Read Depth Analysis of the Ortholog Indel Set

Low depth of coverage is directly related to low consensus quality and vice versa, i.e. the lower the depth of underlying reads, the lower the consensus quality score will be. We also examined depth of coverage in the indel regions (labeled “Read Depth FrameShift” and depicted in red in Figure S1-5) and compared that with genome average (labeled “Read Depth All” and depicted in green in Figure S1-5). From this graph it is clear that the indel regions have lower depth of coverage in general and as discussed above, manual review confirmed that many of these indels are close to the end of contigs or in regions covered by a single reads.

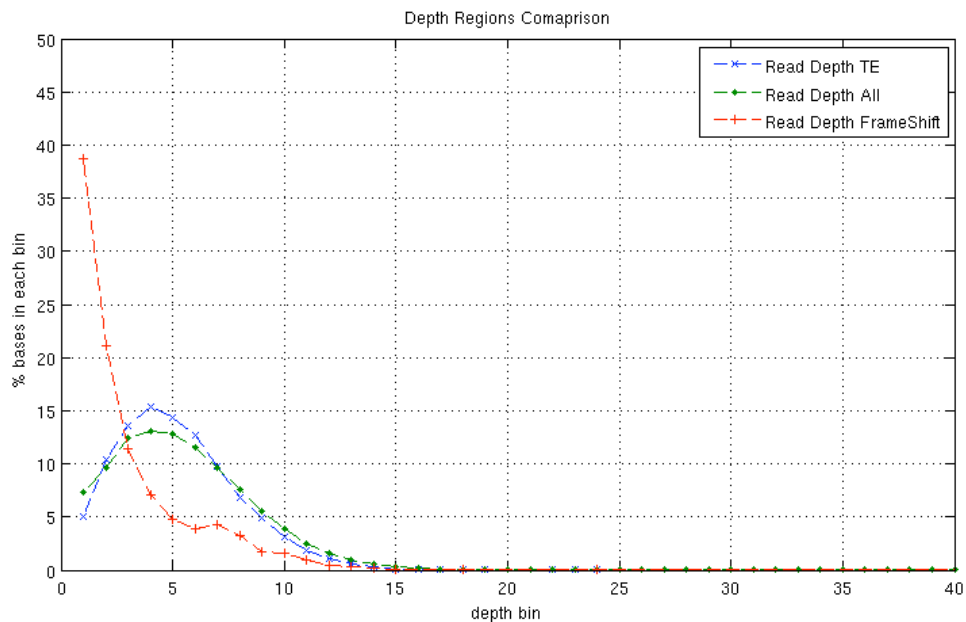


Figure S1-5. Read depth at indel sites compared to the assembly average.

As a further step, we checked the raw, intermediate, and final released assemblies to understand if the indels were introduced during merging of contigs or other manipulations of the assembly. We found 584 of the indels were due to merging neighboring contigs between v2.0 to v2.02 of the assembly accounting for 18% of the total indels. This algorithmic issue has subsequently been addressed and will not happen in subsequent assemblies.

Overall, the consensus quality score provided for each base of the orangutan assembly should be used as a guide with respect to confidence in the quality of any individual base.

Creation of Chromosomal AGP files

The assembly data were aligned against the human genome at UCSC (B. Raney) utilizing BLASTZ⁵ to align and score non-repetitive orangutan regions against repeat-masked human sequence. Alignment chains differentiated between orthologous and paralogous alignments⁶ and only "reciprocal best" alignments were retained in the alignment set. The orangutan AGP files were generated from these alignments in a manner similar to that already described¹⁰. Documented inversions based on primarily on FISH data (Rocchi, personal communication) as well as inversions suggested by the assembly and supported by additional mapping data (e.g. fosmid end sequences against the human assembly, (Chen and Eichler, personal communication)) were introduced, as was the separation of alignments to human chromosome 2 into orangutan chromosomes 2A and 2B. Centromeres were placed based on their localization relative to human based on FISH data (Rocchi, personal communication). Lastly, 78 finished CHORI-276 BAC clones were integrated into the final chromosomal sequences (after quality assessment, see below). Of the 3.09 Gb of total orangutan genome sequence, 3.08 Gb are ordered and oriented along chromosomes with gap sizes between supercontigs estimated based on their size in human.

The Relaxed Stringency Assembly (v2.2)

Due to concerns of over-aggressive separation of allelic copies of duplicate loci in the v2.0.2 assembly (Evan Eichler, personal communication), we used a module of PCAP (PCAP.poly) to generate a version of the assembly with relaxed merging stringency (v2.2). In addition to having a lower rate of false-positive segmental duplications (Evan Eichler, personal communication), the relaxed version showed greater contiguity and scaffolding (Table S1-5). The ~4-fold increase in N50 scaffold length with relaxed merging, without altering the underlying mate pair data, suggests the high heterozygosity rate of the orangutan genome may have resulted in higher fragmentation compared to the N50 scaffold statistics for other primate genomes. Also, a reassessment of indels in the v2.2 assembly eliminated 40% of the previously observed indels in orangutan-human orthologs. Future upgrades to the orangutan assembly will take these factors into account.

Table S1-5. N50 statistics for orangutan assemblies.

Assembly Version	Contig Level Stats		Scaffold Level Stats	
	N50 Length	N50 Number	N50 Length	N50 Number
Raw Stringent Assembly (v2.0)	12.9kb	69888	658kb	1180
Relaxed Assembly (v2.2)	18.3kb	48256	2.8mb	263
Released Assembly (v2.0.2)	15.5kb	55989	739kb	1031

Supplemental Section S2 – Indel Assessment With the Neutral Indel Model

Fine scale accuracy of the Sumatran orangutan assembly was assessed by a new method that analyses insertion and deletion (indel) positions in alignments with the human genome assembly. The Neutral Indel Model⁷ has recently been demonstrated to partition aligned sequences into three types: neutrally evolved sequence, sequence under purifying selection, and sequence containing clusters of spuriously inserted or deleted nucleotides. The latter do not represent the true evolutionary imprint of mutation, but rather are artifacts of the sequencing process. The Neutral Indel Model predicts that numbers of short un-gapped alignment blocks outside of selected sequence follow a geometric distribution. As assembly fidelity decreases, clusters of indel mutations appear, often due to reduced read depth, which appear as an excess of short un-gapped sequence blocks that depart from the geometric distribution.

For the alignment of Sumatran orangutan and human non-repetitive sequences, the indel error rate was 0.99 per kb (95% c.i. = 0.98-1.00). As expected, errors cluster predominantly in regions of low sequence coverage, and towards contig ends. Illumina short reads derived from KB5404 (a Bornean individual with ~20x coverage; Supplemental Section S4) were then used to simulate a Bornean orangutan genome assembly, using the Sumatran assembly as a template¹¹. Where coverage by these short reads was sufficient to call variants (80.3% of the genome), the inferred indel error rate was reduced 4-fold to 0.25 per kb (95% c.i. = 0.24-0.26), whilst the inferred rate of true neutral indels remained essentially unchanged. This demonstrates that mixing capillary reads with next-generation short sequence reads provides an effective approach to producing higher-fidelity genome sequences.

Supplemental Section S3 – Great Ape Divergence Estimate via WGS Read Mapping

We aligned whole genome shotgun reads of human, chimpanzee, orangutan, gibbon and macaque against human reference genome (Hs.35/hg17) to calculate the average divergence/identity of each species against the human genome (Table S3-1). We used windows of 5 kb of non-RepeatMasked, gap and duplication free sequence (PhredQ > 20, 27, 30, 30 and 27 for human, chimpanzee, orangutan, gibbon and macaque,

respectively). We calculated the raw average percentage identity from WGS reads with at least 200 bp of high quality bases. The calculation of percentage identity excludes RepeatMasked bases and has been corrected by Kimura2. This divergence estimate, based on the raw orangutan WGS reads, were consistent with our expectations, and served to confirm species.

Table S3-1. Basic statistics of ape-human divergence (hg17).

	AVERAGE % identity	MEDIAN % identity	STD DEV
Human	0.99927	0.99951	0.00104
Chimpanzee	0.99061	0.99101	0.00313
Orangutan	0.97463	0.97500	0.00487
Gibbon	0.97130	0.97164	0.00512
Macaque	0.94996	0.95076	0.00919

If we assume 6 MYA as the speciation time for human and chimpanzee (divergence of 0.9 %), we can extrapolate that orangutan diverged 16.7 MYA, gibbon at 19.3 MYA and macaque at 33.3 MYA. If we assume a divergence of macaque at 25 MYA (5% divergence), we extrapolate that gibbon diverged 14.5, 12.5 for orangutan and 4.5 for human/chimp. In light of the hominoid slowdown in substitution, these can serve as upper and lower bound estimates.

We found that there is a 0.4% difference in divergence between the gibbon and orangutan when compared to the human genome. So, assuming a constant molecular clock, gibbons diverged ~2 (2.5) million years earlier than orangutans. However, the distribution of sequence identity between the orangutan and gibbon genomes differs significantly; i.e. both species did not share a common ancestor after divergence from the human lineage (Figure S3-1).

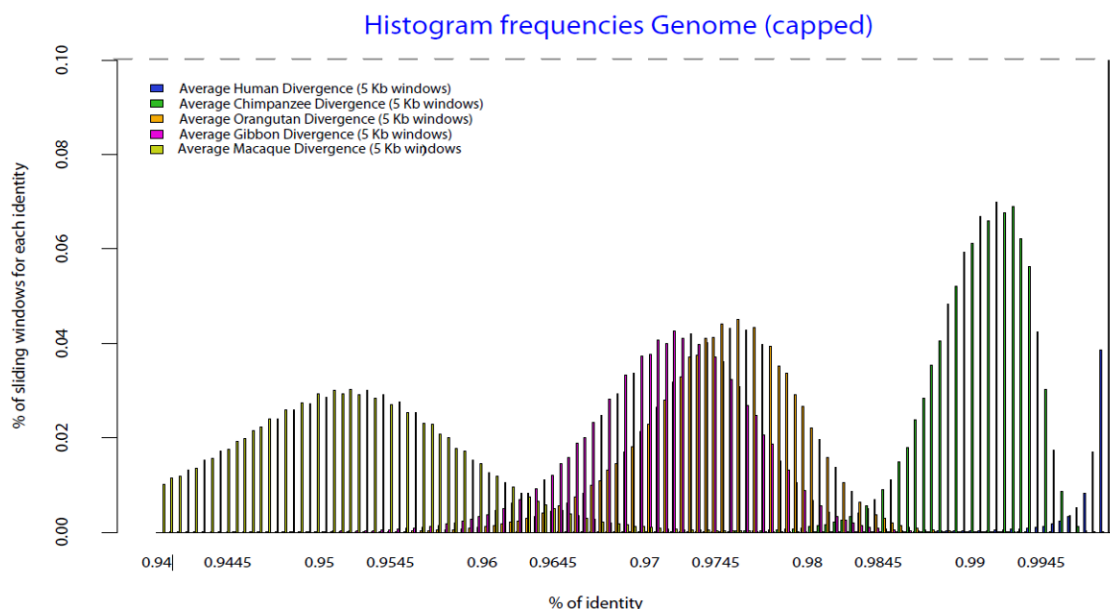


Figure S3-1. Select primate divergence histograms. Histogram of frequency of identities of human, chimpanzee, orangutan, gibbon and macaque WGS reads mapped to the human reference genome in 5 kb windows. Data is shown capped at 10% for a better view of chimp, orangutan, gibbon and macaque.

Supplemental Section S4 – Short Read Sequencing

The orangutan population diversity survey utilized DNA from 5 Sumatran and 5 Bornean wild-caught orangutans, provided by Dr. Oliver Ryder and the San Diego Zoo's Institute for Conservation Research, San Diego, California. DNA from each orangutan was individually fragmented and ligated with adapters suitable for PCR amplification and sequencing on the Illumina GA/GAI platform in accordance with the manufacturer's protocols. For a thorough Bornean vs Sumatran comparison, one female Bornean individual (KB5404) was selected for deep (20x) coverage, and the remaining individuals were targeted for ~8x coverage (Table S4-1). A mix of paired end and fragment reads (36 bp, 50 bp, and 75 bp read length) was used to reach coverage targets. Pair spans for the 9 ~8x coverage individuals are approximately 180-280 bp. Pair spans of 180-280 bp, 280-380 bp and 380-480 bp were used for KB5404. All reads have been submitted to the NCBI Short Read Archive (<http://www.ncbi.nlm.nih.gov/sites/sra>) and accession numbers are reported in the associated file "orangutan_short_read_genome_sequence_accessions.xlsx".

Table S4-1. Next generation sequence data summary.

Sample ID	Studbook#	Name	Institution When Sampled; Local ID	Origin	Sex	Raw Data (Gb)	Coverage (x)	Major Read Type
KB5404	590	Billy	Lincoln Park Zoo, Chicago; 000362	Borneo	F	61	20.3	50 bp Paired Ends
KB5406	356	Dinah	Dallas Zoo; 001036	Borneo	F	23	7.7	50 bp Paired Ends
KB5405	360	Dennis	Dallas Zoo; na	Borneo	M	26	8.7	75 bp Paired Ends
KB4204	364	Dolly	Dallas Zoo; 001041	Borneo	M	25	8.3	50 bp Paired Ends
KB5543	990	Louis	Los Angeles Zoo; 001929	Borneo	M	29	9.7	50 bp Paired Ends
SB550	53	Doris	San Diego Zoo; 148001	Sumatra	F	21	7	50 bp Paired Ends
KB9528	732	Baldy	Sacramento Zoo; 100083	Sumatra	M	28	9.3	50 bp Paired Ends
KB4361	1600	Likoe	Miami Metro Zoo; M00176	Sumatra	F	21	7	50 bp Paired Ends
KB4661	695	Bubbles	San Diego Zoo; 177257	Sumatra	M	20	6.7	50 bp Paired Ends
KB5883	550	Sibu	Atlanta Zoo; 681456	Sumatra	M	25	8.3	50 bp Paired Ends

Supplemental Section S5 – The Orangutan Ensembl Gene Set

The orangutan gene set was produced by combining 3 transcript sets, each made using a different technique. The primary transcript set consisted of 'Targeted' gene models¹² made from Genewise alignments of orangutan proteins. This was augmented with 2 transcript sets derived from human Ensembl transcripts aligned to the orangutan assembly. The first of these used a whole genome alignment to project human Ensembl transcripts on to orangutan. The second set consisted of human Ensembl translations aligned to orangutan using Exonerate¹³. The combination of approaches aimed to maximise coverage of the genome and the use of human evidence.

The transcripts produced by the human alignments were combined to remove redundancy. Where the different techniques had built conflicting models from the same evidence, we used a transcript scoring method to identify consensus transcripts. EST

and cDNA support was used to help inform the scoring. For each exon and intron we calculated the score as the number features which matched exactly, over the total number of overlapping features. Additional weight was given to cDNA and EST evidence, which included human cDNAs and orangutan 454 reads provided by WashU (see below). Finally, we gave an additional score to exons supported by WashU Illumina data. The scores were summed over all exons and introns for each transcript, and the highest scoring transcript at each locus was selected. The result was a set of well-supported human transcripts aligned to orangutan, with one transcript per locus.

UTR addition was carried out on the three transcript sets; orangutan cDNAs were used to add UTR to orangutan specific transcripts using the standard Ensembl methods. Where transcripts were derived from human data, we used our transcript scoring system to guide UTR addition, allowing UTR only where we had strong evidence. As a result we were able to add UTR based on Orangutan ESTs, 454 alignments and human cDNAs.

Finally the consensus human models were combined with Orangutan specific transcripts and alternate isoforms to produce a finished protein coding gene set. This was then scanned to identify potential pseudogenes. A second non-coding gene set was also added using standard Ensembl methods. Information about the orangutan gene set is available at: http://www.ensembl.org/Pongo_pygmaeus/Info/Index

Orangutan cDNA Data

In order to enhance *in silico* gene predictions, and create a resource for expression-based studies, we generated cDNA data from multiple RNA sources. Using an oligo-dT-based approach¹⁴ cDNA libraries were created from two fibroblast cell lines (both obtained from Coriell): 1) GM04272 – derived from a male Sumatran orangutan and 2) PR01109 – a cell line derived from Susie (the sequenced female Sumatran orangutan). Both of these cDNA libraries were sequenced on the 454 FLX platform. In addition, in collaboration with Svante Paabo's lab (Janet Kelso, Kay Pruefer and Birgit Nickel), we generated cDNA libraries from orangutan tissue-based RNA (brain, heart, kidney and liver). These libraries were also sequenced on the 454 FLX platform. Lastly, the PR01109 cDNA library was also sequenced on the Illumina GA platform (35 bp reads).

Table S5-1. cDNA data summary.

454 Data	Raw Coverage (bp)
GM04272 (male, Sumatran cell line)	94943834
PR01109 (female, Sumatran cell line, aka Susie)	162879477
Brain tissue (male, Sumatran)	69791469
Heart tissue (female, Sumatran)	42771963
Kidney tissue (male, Sumatran)	50203738
Liver tissue (female, Sumatran)	63996094
Illumina Data	Raw Coverage (bp)
PR01109 (female, Sumatran cell line, aka Susie)	821205824

cDNA Data Access

All cDNA data was deposited in the NCBI Short Read Archive and accession numbers are reported in the associated file “orangutan_cDNA_accessions.xlsx”.

Supplemental Section S6 – Ancestral Reconstruction

We partitioned the genomes of human, chimp and orangutan into 137 atoms in 100 kb resolution. These atoms cover 94% of the human genome. By recovering the ancestral order and orientation of these atoms using the genome reconstruction algorithm CARs¹⁵, we reconstructed the karyotype of the common ancestor of human, chimp and orangutan, using rhesus, mouse and dog as outgroups (Figure S6-1 & Table S6-1). We then performed a further analysis to reconstruct more detailed evolutionary operations. We used an atom set from the Ensembl Enredo pipeline¹⁶ (provided by Ewan Birney’s group) to further partition the genome at a higher resolution (5 kb minimum atom size). In this analysis, we used rhesus as the sole outgroup with 6552 atoms at this level of resolution. Note that this dataset contains duplications as well as large insertions and deletions. These smaller atoms cover 91% of the human genome sequence. We then ran our Reverse Evolution reconstruction algorithm¹⁷ to reconstruct the evolutionary operations, including rearrangements and duplications.

We inspected intervals around breakpoints, looking for genomic properties in the human genome that might help explain why breaks occur at some positions but not others. We used 20 kb intervals centered at each side of an atom where the breakpoint happened. We compared the segmental duplications and repeats in these breakpoint regions with genome-wide average. Our observations are summarized in Table S6-2.

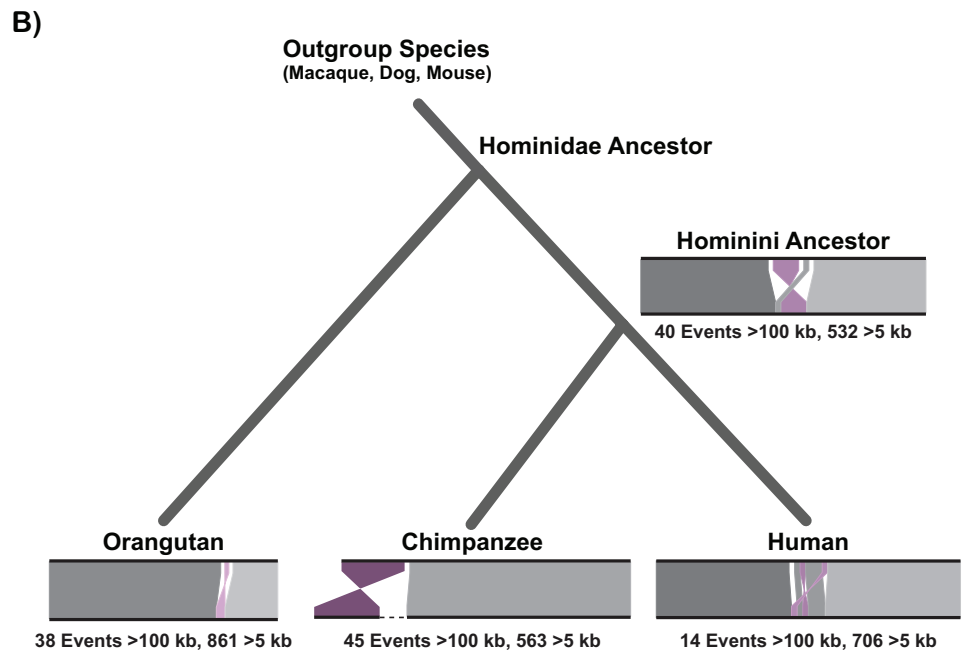
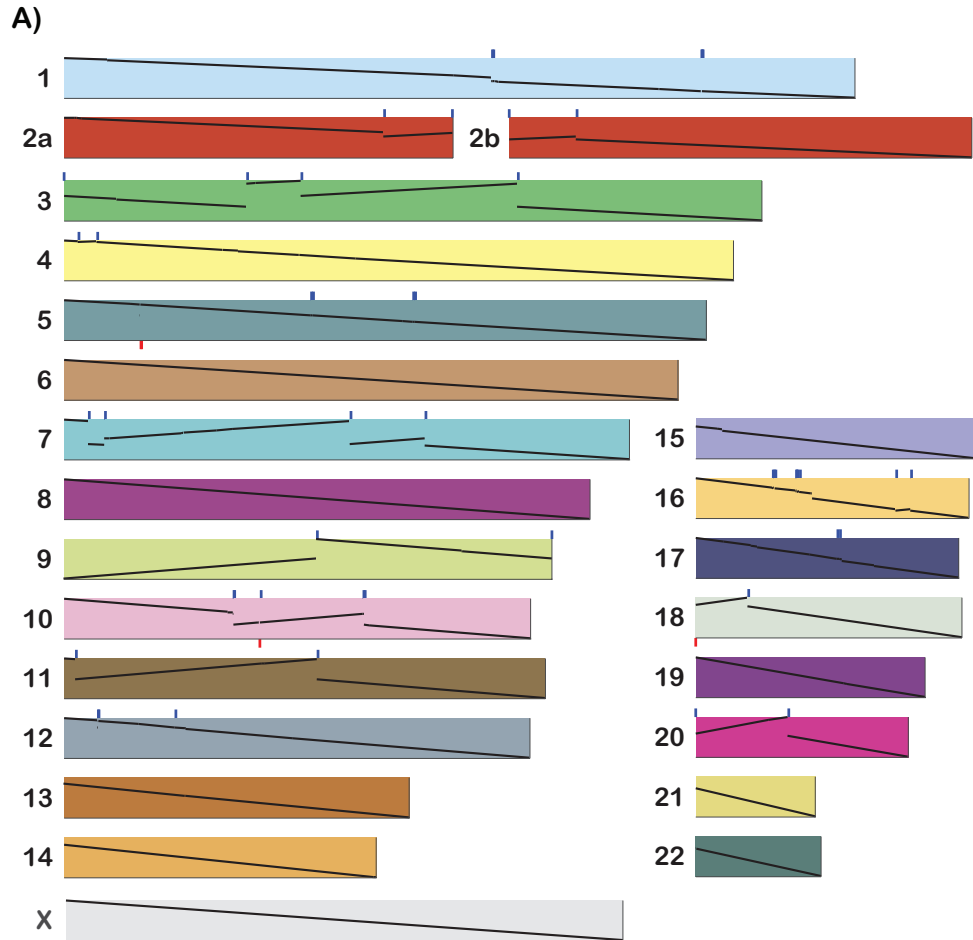


Figure S6-1. Reconstruction of the ancestral Hominidae genome. **a**, The karyotype of the ancestral Hominidae genome is shown with the corresponding human chromosome number indicated on the left of each block. Diagonal lines within each block show the orientation and position in the human chromosome, highlighting the intra-chromosomal rearrangements that occurred on the human branch. Blue tick marks above the bars indicate human-specific breakpoints and red tick marks below the bars show the human breakpoints that were confirmed only with the orangutan genome and not more distantly related outgroup species (rhesus macaque, dog and mouse). **b**, The number of rearrangements within a lineage is indicated at each node, with estimates from both 100 kb and 5 kb resolution analyses. Overall, the orangutan genome displayed fewer rearrangements at both levels of resolution. The figure also highlights a 50 Mb region of chromosome 5 subject to lineage-specific rearrangements (hg18.chr5:86M-138M). In the detailed analysis using finer resolution atoms, the reconstruction shows an inversion (hg18.chr5:98.9M-99.6M) occurred along the shared human-chimp lineage after orangutan divergence, but before human-chimpanzee divergence. This reconstruction also reflects other operations around the region. On the chimpanzee lineage, part of this region (panTro2.chr5:19M-29M) was involved in the pericentric inversion on chimpanzee chromosome 5 (panTro2.chr5:19M-97.5M).

Branch	Large-scale breakpoints (>100 kb)	Detailed operations (>5k)	
		2-breakpoint operations	3-breakpoint operations
Hominidae Anc → Orangutan	38	823	38
Hominidae Anc → Hominini Anc	40	482	50
Hominini Anc → Chimpanzee	45	495	68
Hominini Anc → Human	14	566	140

Table S6-1: Summary of the number of large-scale breakpoints and the number of different types of operations on each branch in the phylogenetic tree for human, chimp, and orangutan.

	100 kb resolution	5 kb resolution	Genome wide average
Repeats density (%)	49.67	47.15	48.58
Segmental duplication (%)	52.32	26.82	5.24

Table S6-2: Summary of the segmental duplication and repeats density in the breakpoint regions in different resolutions. Repeats were identified with RepeatMasker, and segmental duplications were obtained from UCSC Genome Browser segmental duplication track.

Supplemental Section S7 – The Genomic Distribution of Genic Evolution Rates

Summary

Our results support previous findings suggesting that rates of genic divergence strongly depend on chromosomal location, which may have the implication of making adaptation contingent on chromosomal location.

- There are very significant differences in rates of evolution in different parts of the genome:
 - Genes near telomeres tend to diverge faster than genes elsewhere in the genome.
 - Genes near centromeres tend to diverge slower than genes elsewhere in the genome
 - Genes in SDs and CNVs tend to present more divergence than genes in single-copy regions.

Some, but not all, of these patterns can be explained by GC content.

- As previously demonstrated, in the human and chimpanzee lineages genes within rearrangements tend to present lower divergence rates. The pattern is exactly the opposite in the orangutan branch: genes within rearrangements diverge faster.

No evidence for chromosomal speciation in the great apes.

- The pattern above may seem suggestive of chromosomal speciation (since chromosomal speciation theory predicts more divergence around rearrangements that took part in speciation processes), but after appropriate testing, we find no evidence supporting it.

Introduction

The fact that rates of divergence are not uniformly distributed in the genomes of primates has raised considerable interest, since it may imply that rates and patterns of adaptation and speciation depend contingent factors such as nucleotide composition and the location of genes. The Sumatran orangutan (*Pongo abelii*) genome, together with the available genome sequences of human, chimpanzee and macaque, provides an opportunity to perform a complete analysis of rates of genic evolution all over the genomic landscape of the great apes. In particular, we investigated how rates of evolution are affected by the location of genes in centromeres, telomeres, structural variants and chromosomal rearrangements.

Results

Rates of genic evolution in major genomic regions

Examining any divergence measure along a chromosome unveils considerable heterogeneity, even when binning the chromosome in large 5 Mb windows (an example

can be found in Figure 1, Appendix 1; see files S7_Appendix1a_chr1_dSorang_humanCoordinates.xls and S7_Appendix1b_chr1_dSorang_orangCoordinates.xls), so it makes sense to examine major regions separately.

We first examined rates of divergence in telomeres and centromeres from the human and orangutan viewpoints. Full results can be found in **Supp. Tables 1-4** (Appendix 2; see file S7_Appendix2_Overall_Tables.xls). Genes close to telomeres tend to present faster rates of divergence than genes elsewhere in the genome, while divergence tends to be slower for genes close to centromeres. Telomeric trends are the strongest and can affect rates of protein evolution (as measured by w). Thus, rates of protein evolution are not independent of the location of genes across the genome. The trend observed in telomeres is explained by increased GC content (Appendix 4, **Supp. Tables 16-18**; see file S7_Appendix4_GCcontent.xls). Lower divergence in centromeres, even if weaker, cannot be explained by nucleotide composition, since they genes close to centromeres also have a higher than average GC content. Their lower divergence is probably linked to lower rates of recombination in these regions.

Structural Variation has been shown to affect rates of evolution; with duplicated genes presenting faster rates^{18,19}. After removing genes in telomeres and centromeres (Appendix 2, **Supp. Table 5**; see file S7_Appendix2_Overall_Tables.xls) we can see that rates of divergence tend to be faster for genes overlapping SDs and CNVs, the only exception seems to be a marginally significant reduction of rates of protein evolution in the orangutan branch, probably caused by a very large increase in dS for that branch. Results are similar when analyzing separately SDs and CNVs (not shown).

Finally, we analyzed rates of divergence in major chromosomal rearrangements. **Supp. Tables 6 and 7** (see file S7_Appendix2_Overall_Tables.xls) contain the results of the overall analysis. After removing all factors studied above, genes within rearrangements present significantly lower synonymous (dS) and intronic (ki) divergence than genes outside them, which is consistent with the latest findings for the human and chimpanzee lineages²⁰. In **Supp. Tables 8 and 9** (see file S7_Appendix2_Overall_Tables.xls), divergence rates are examined for each branch. There is a major difference between the orangutan branch and the rest. While genes within CRs in the human and chimpanzee branch tend to diverge less than genes elsewhere in the genome, as has been shown before²⁰, genes within orangutan rearrangements seem to evolve faster, just as predicted by extant models of chromosomal speciation²¹.

To ascertain the causes of this pattern, we performed an analysis of rearrangement breakpoints (**Supp. Table 10**; see file S7_Appendix2_Overall_Tables.xls) on the basis that any effects of rearrangements mediated by recombination have been predicted to be stronger there^{20,21}. Results were inconclusive due to lack of statistical power: the number of genes located close to each individual breakpoint is too small to allow for significant results. To settle the issue, we performed a chromosome-per-chromosome analysis for the rearrangements in the hominid and the orangutan branches (**Supp. Tables 11-14**; see file S7_Appendix3_Per_Chr_Tables.xls). The overall effect of

accelerated evolution within orangutan rearrangements is due to genes in chromosome 3. Table 12 shows that rapid divergence of the genes within that rearrangement is not exclusive of the orangutan or hominid lineage, but is also present in chimpanzees and, thus, it seems unlikely to be related to the orangutan speciation. In addition, we checked divergence in the outgroup branch leading to the macaque and found that dN and w were higher for these genes even before the origin of the rearrangement in chromosome 3 (dN, 0.014 within the rearrangement vs. 0.008 outside, p-value < 10^{-4} ; w, 129 vs. 133, p-value < 10^{-4}). Therefore, evidence does not sustain the hypothesis that chr3 took part in a chromosomal speciation process.

Methods

Genomic features

Centromere and telomere coordinates for human and orangutan were retrieved from UCSC. Genes were classified in three categories; pericentromeric (within 5 Mb of centromeres), (sub)telomeric (within 10 Mb of the tips of chromosomes), and rest. The coordinates of human structural variants (SV), including segmental duplications (SDs) and copy number variants (CNVs) were downloaded from: <http://eichlerlab.gs.washington.edu/database.html> and <http://projects.tcag.ca/variation/>, respectively, while chimpanzee and orangutan SV coordinates were obtained through computational and experimental work carried-out in A. Navarro's and E. Eichler's laboratories. Genes were classified according to their overlap with any known SVs in these species as genes overlapping structural variants (SVs) or genes single copy regions (SCRs).

The coordinates of large chromosomal rearrangements (CRs) and their breakpoints were cytogenetically determined based on FISH-BAC experiments performed in M. Rocchi's laboratory (see below). CRs were classified according to the five branches of the phylogenetic tree we were studying: Human (H), Chimpanzee (C), Gorilla (G), Orangutan (O), Human-chimpanzee ancestor (HC) and Human-Chimpanzee-Gorilla ancestor (HCG). Coordinates of CRs in the human and chimpanzee branches were refined with data from the literature^{22,23,24,25,26,27,28}. Additionally, we used an ad-hoc procedure based on gene position to identify new rearrangements and to confirm and refine breakpoint coordinates. Annotated genes were classified, in every particular branch, as rearranged (REA) or collinear (COL) according to their location inside or outside chromosomal regions that rearranged in that branch. In addition, genes within 1 Mb from regions containing rearrangement breakpoints were labelled accordingly (Break).

Rates of divergence

For annotated genes, coding region divergence was estimated based on non-synonymous substitutions (dN), synonymous substitutions (dS) and their ratio (w) for each branch (H, C, HC, HCO, O), using the *codeml* program in PAML²⁹. Estimates were obtained assuming a free-rate branch model and the GTR model of nucleotide

substitution. We used the same alignments and the same species requirements as the positive selection analysis. Mean intronic divergence per intron, K_i , was estimated for every branch using *baseml*, also in PAML²⁹. For each gene, K_i was estimated as the weighted average of its introns. Here, we used the same set of alignments that was used for detection of the action of positive selection in introns. Different filters were subsequently applied to exclude false orthologies. After estimating Mean, Median (Md) and Standard Deviations (SD) for dN , dS and K_i in every branch, we filtered out, per branch and per variable, any values that are above the median plus two standard deviations and the corresponding omega values. In addition, w values were filtered out for genes with $dS < 0.001$ or set to zero for genes with $dS \geq 0.001$ and $dN = 0$.

Supplemental Section S8 – Cytogenetic Characterization

Note: To simplify the orangutan-human comparison, we used phylogenetic nomenclature, but with Arabic figures instead of Latin figures.

Synteny Organization

A detailed characterization of the synteny organization of the orangutan genome with respect to the human genome was obtained by Fluorescence In Situ Hybridization (FISH) of ~470 human BAC-based probes to orangutan (*Pongo abelii*) metaphases. A full understanding of synteny organization illuminated the cascade of rearrangements that each chromosome underwent since the divergence of the great apes from a common Hominioid ancestor. This comparison is crucial to determine the origin of lineage-specific rearrangements within the Hominoidea. The synteny organization of the orangutan genome, compared to the human genome, is presented at <http://www.biologia.uniba.it/orang> (username: pongo, password: pygmaeus2).

The borders of each synteny segment were defined by a split signal from a single BAC probe or by two overlapping BACs spanning each breakpoint. Many breakpoints were further validated by using orangutan BAC clones (CHORI-253 library) hybridized to human metaphases. Candidate BACs were also identified during the assembly process (L. Hillier, personal communication). For a detailed description of this molecular cytogenetic approach see Stanyon et al. 2008³⁰.

Synteny Lineage Tracking

The flow of chromosomal rearrangements from the Hominioid ancestor to orangutans and humans is graphically illustrated at the bottom of each chromosome-specific web page (<http://www.biologia.uniba.it/orang>; username: pongo, password: pygmaeus2). The flow of rearrangements, inversions in particular, are also important to validate the reciprocal orientation of the synteny segments. Note that chromosome 3 of the Sumatran orangutan (*Pongo abelii*) is derivative with respect to the Bornean orangutan (*Pongo pygmaeus*) (see www.biologia.uniba.it/orang/PPY/PPY_03.html and Figure S8-1).

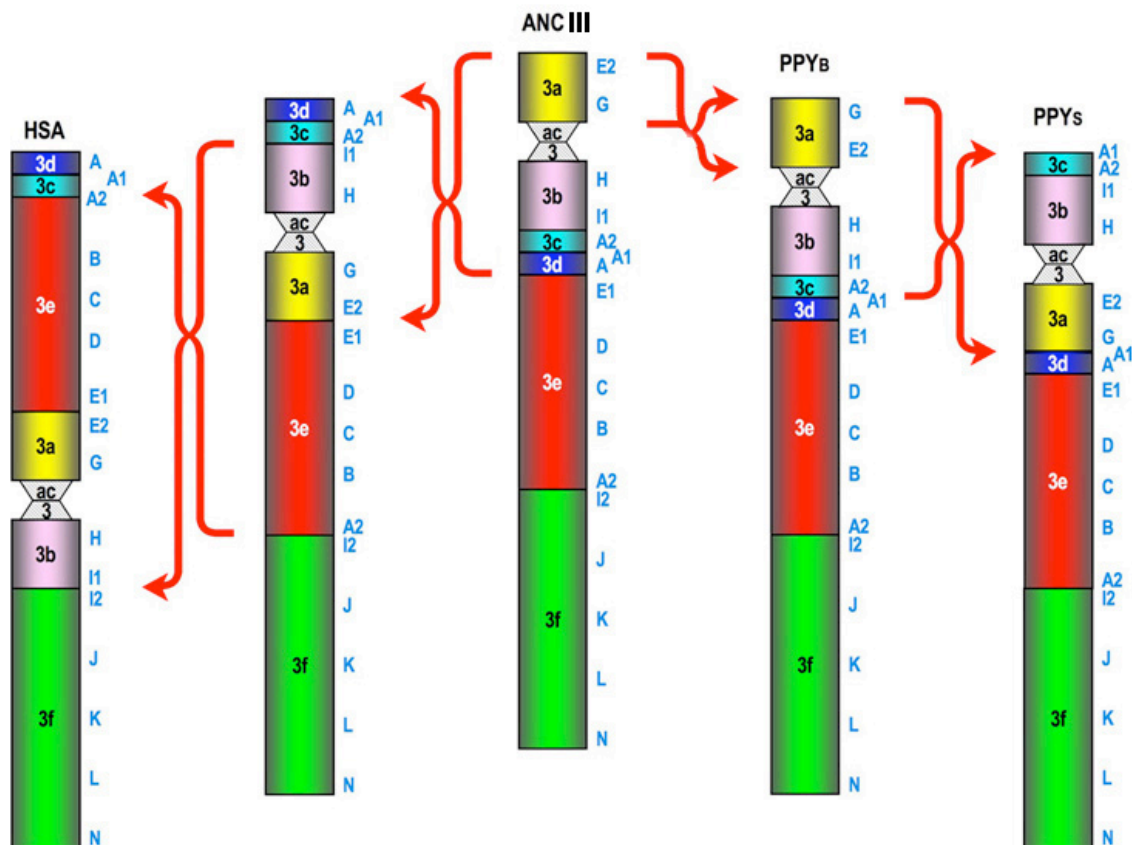


Figure S8-1. Large-scale evolution of chromosome 3 in select Hominoidea. Depicted is the cascade of large-scale rearrangements of chromosome 3 descending from a common hominoid ancestor (ANC). Note that evolutionary synteny blocks are numbered according to their position in the ancestor. Human (*Homo sapiens*) is designated HSA, The Bornean orangutan (*Pongo pygmaeus*) is designated PPYb, the Sumatran orangutan (*Pongo abelii*) is designated PPYs.

Isolating The Chromosome 3 Inversion Differentiating Bornean and Sumatran Orangutans

The Sumatran orangutan differs from the Bornean orangutan by a pericentric inversion of chromosome 3 (shown below, and also see Figure S8-1). FISH mapping of human BAC clones from the area syntenic to the breakpoint identified a breakpoint-spanning clone (RP11-732C9; Figure S4-2)

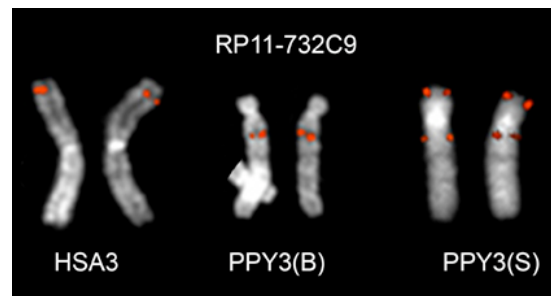


Figure S8-2. Isolating the inversion breakpoint. Metaphase chromosome 3 from a human (HSA3), a Bornean orangutan (PPY3(B)) and a Sumatran orangutan (PPY3(S)), hybridized with human BAC RP11-732C9 (hg18.chr3:12,441,757-12,649,037). Note the split of the BAC probe signal in the Sumatran orangutan indicating this clone crosses the inversion breakpoint.

The Orangutan Neocentromere

In 1976, a putative pericentric inversion of orangutan chromosome 12 was identified, heterozygous in both Sumatran and Bornean orangutan populations³¹. Susie proved homozygous for the more common ancestral form of chromosome 12 (data not shown); therefore, we studied a heterozygous individual (a Sumatran male) with the intent of refining the inversion breakpoints. Surprisingly, the “inverted” chromosome showed no difference in marker order with respect to the ancestral configuration, yet the position of the centromere had changed (Figure S8-3A). From Figure S8-3A, the clone names of the BAC probes used and corresponding human genome coordinates are as follows: yellow - RP11-12I7 (hg18.chr12:28,168,830-28,312,751); blue - RP11-80I23 (hg18.chr12:41,090,617-41,278,100); green - RP11-10O10 (hg18.chr12:71,906,687-72,062,213); red - RP11-20L19 (hg18.chr12:89,601,022-89,774,035). The functionality of this neocentromere was confirmed with ChIP-on-chip analysis (see below).

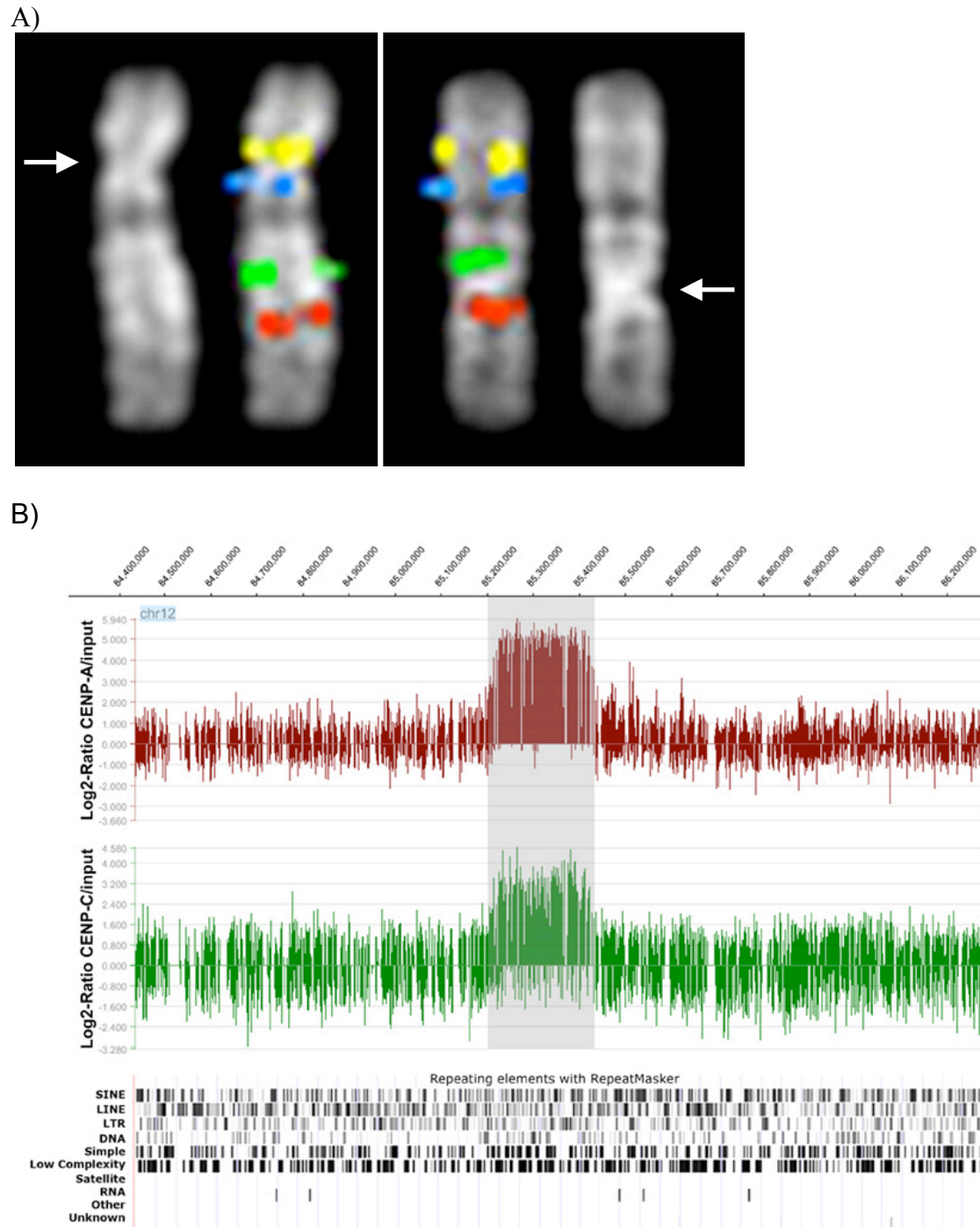


Figure S8-3. A polymorphic neocentromere of chromosome 12 in the orangutan population. **a**, Note the identical order of BAC FISH probes between the “normal” (left) and “inverted” (right) configuration of orangutan chromosome 12, despite the contrasting centromere positions, indicated by arrows. **b**, ChIP-on-chip analysis delineated the CENP-A (red track) and CENP-C (green track) binding domains, indicating the neocentromere is functional. The y-axis of each SignalMap (NimbleGen Systems Inc.) view represents the log₂ fluorescence intensity ratio of immunoprecipitated DNA with respect to controls. The x-axis indicates position along orangutan chromosome 12. The shaded area highlights the CENP-A and CENP-C co-localizing domains. Note the lack of satellite sequence in this region identified by RepeatMasker (adapted from the UCSC Genome Browser).

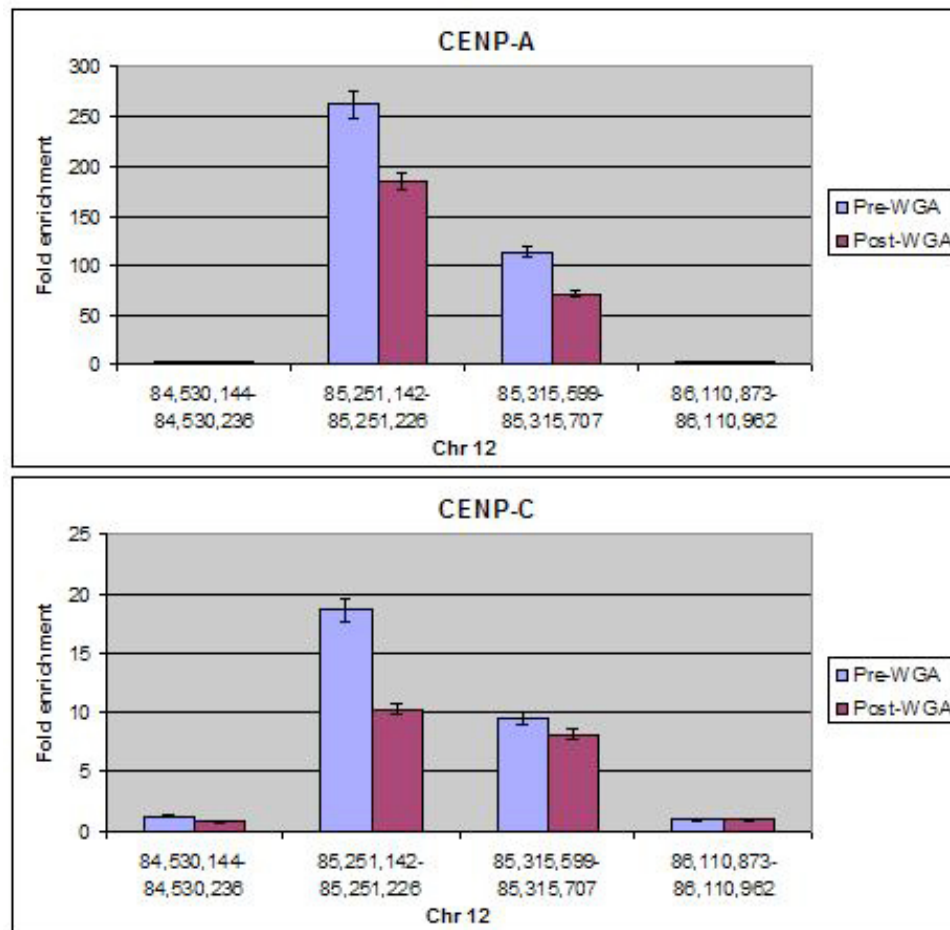
Functional Characterization of the Neocentromere

To further refine the location of the neocentromere, and confirm its functionality, we investigated by ChIP-on-chip analysis (see Figure S8-3B). Experiments were performed using two rabbit polyclonal antibodies directed against CENP-A or CENP-C human centromeric proteins. These DNA binding-proteins are required for kinetochore function and are exclusively targeted to functional centromeres (reviewed in Carroll and Straight 2006³²). Thus, the immunoprecipitation of the DNA bound to these proteins allows the isolation of centromeric sequences, including those of the orangutan neocentromere. Immunoprecipitated DNA from an individual bearing the neocentromere was amplified and hybridized to a NimbleGen custom oligo array with an average resolution of 1 oligo per ~100 bp across a ~74 Mb interval of the assembly (v2.0.2, a.k.a. ponAbe2) encompassing the neocentromere. A solitary peak was identified for both CENP-A and CENP-C, using very stringent conditions (98th percentile threshold and $P < 0.0001$).

Methods for ChIP-on-chip analysis

To identify the sequences bound by CENP-A, native chromatin immune-precipitation (N-ChIP) analysis was performed, as previously described³³. Briefly: lymphoblastoid cells derived from the father were processed and the native chromatin was prepared by nuclease digestion of cell nuclei, then the immunoprecipitation was performed using polyclonal antibodies against the centromeric protein CENP-A and CENP-C. Crosslinked chromatin immune-precipitation (X-ChIP) analysis, as previously described³⁴, was performed to identify the sequences bound by CENP-A/C. Briefly: cells were crosslinked in situ by adding formaldehyde to a 1% final concentration directly to the culture medium and chromatin was immunoprecipitated with an anti-CENP-A and anti-CENP-C polyclonal antibodies³⁵. In both methods, purified DNA fragments were amplified using the Whole Genome Amplification kit (Sigma-Aldrich, St.Louis, USA). The enrichment of ChIP DNA before and after amplification was validated by Real-Time PCR (Figure S8-4).

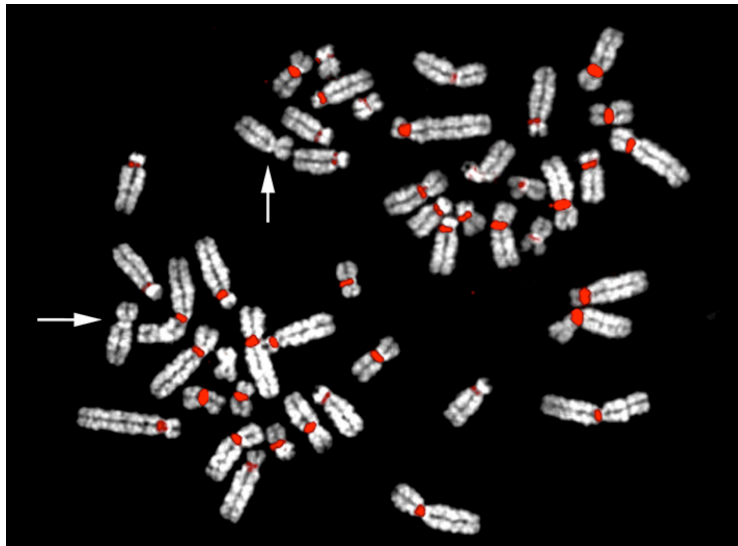
The labeled ChIP and total DNAs were co-hybridized to a NimbleGene custom oligo tiling array, which has an average resolution of 100 bp. The oligos were designed on the masked orangutan assembly (v2.0.2), and covered the region chr12:58,836,844-132,289,483. DNA binding peaks were identified by using the statistical model and methodology described at: <http://chipanalysis.genomecenter.ucdavis.edu/cgi/tamalpais.cgi>³⁶.



Supplemental Figure S8-4. Validation of ChIP enrichments by Real-Time PCR. Real-Time PCR products show enrichment of chromosome 12 orangutan DNA from the neocentromeric region immunoprecipitated with anti-CENP-A or anti-CENP-C antibodies before and after WGA amplification. Fold enrichment of the indicated sample was calculated as the ratio between the immunoprecipitated and the input DNA. The enrichment was evaluated in two regions inside the CENP-A/CENP-C domain (orangutan chr12:85,251,142-85,251,226 and chr12:85,315,599-85,315,707) and in two regions outside the domain (orangutan chr12:84,530,144-84,530,236 and chr12:86,110,873-86,110,962). Results were averaged among three independent ChIP experiments.

Lack of Orangutan Chromosome 12 Alphoid Sequences

Discovery of the polymorphic neocentromere raised the question of the distribution of alphoid sequences in the orangutan genome, specifically concerning chromosome 12. We addressed this using FISH with a pancentromeric alphoid probe and *in situ* oligo-primed synthesis (PRINS)³⁷ to characterize and quantify alphoid sequences in the orangutan genome. Interestingly, chromosome 12 was the only chromosome that repeatedly failed to show any detectable alphoid signal, suggesting an unusual reduction of the alphoid array (Figure S8-5). Unequal crossing-over can drive expansion and contraction of satellite DNA arrays³⁸. It can be hypothesized that the extreme contraction, by unequal crossing-over, could have played a role on the repositioning event.



Supplemental Figure S8-5. Lack of alphoid FISH signal on orangutan chromosome 12. FISH with alphoid clone aTLp on a *Pongo pygmaeus* metaphase spread shows a consistent lack of signal on chromosome 12 (arrows).

Supplemental Section S9 – High-copy Repeat Assessment

Methods

BLASTZ generated human-orangutan chains were downloaded from UCSC (<http://genome.ucsc.edu>) and analyzed for the presence of orangutan-lineage specific L1 and *Alu* insertions.

PCR amplifications were performed in 25 μ l reactions containing 15-50 ng of template DNA; 200 nM of each oligonucleotide primer; 1.5 (full-length L1 and SVA analyses) to 2.0 mM (all other L1 and *Alu* element analyses) $MgCl_2$, 10x PCR buffer (50 mM KCl; 10 mM TrisHCl, pH 8.4); 0.2 mM dNTPs; and 1-2 U *Taq* DNA polymerase. Primer sequences and PCR conditions can be found at <http://batzerlab.lsu.edu>. Due to the large insertion size for most L1 and SVA elements, a second (internal) PCR with one primer residing within the retrotransposon insertion was required to verify insertion presence/absence. In addition, for some PCR reactions 0.125 μ l T4 Gene 32 Protein (#M0300L, New England Biolabs) was added to the reagent mix to enhance the yield of the PCR amplification.

PCR reactions were performed under the following conditions: initial denaturation at 94°C for 90 sec, followed by 32 cycles of denaturation at 94°C for 20 sec, 20 sec at primer annealing temperature (specific to each primer combination, see primer list for specifics at <http://batzerlab.lsu.edu>), extension at 72°C for 30 to 60 sec depending on the predicted PCR amplicon size. PCRs were terminated with a final extension at 72°C for 3 min. 20 μ l of each PCR product were fractionated in a horizontal gel chamber on a

2% agarose gel containing 0.1 $\mu\text{g/ml}$ ethidium bromide for 50-60 minutes at 175V. UV-fluorescence was used to visualize the DNA fragments.

Allele-Specific Alu PCR (ASAP) is a technique designed to selectively amplify *Alu* insertions from a primate genomic DNA sample of interest^{39,40}. The elements retrieved are sequenced along with some unique flanking sequence and then compared to the reference genome. *Alu* elements recovered using ASAP that are absent from the reference genome are considered novel insertion polymorphisms. This approach has been successfully utilized in primate studies to identify young *Alu* insertions that were undetected in the reference genome³⁹ or from primate species without a reference genome⁴¹. We designed PCR primers to target two of the youngest orangutan-lineage specific *Alu* subfamilies with indication of recent retrotransposition activity (polymorphic insertions) in both Sumatran and Bornean orangutans to identify putative novel insertions undetected in the draft reference sequence.

Results

We investigated the mobile element composition of the orangutan draft genome sequence (ponAbe2). With approximately half of the orangutan genome occupied by repetitive sequences, the orangutan has a comparable mobile element content compared to the other three primate genome sequence assemblies completed to date: human, chimpanzee, and rhesus macaque^{10,42,43}. As with other primate genomes, there is no evidence of DNA transposon mobilization activity. We also investigated the endogenous retrovirus (ERV) composition of the *P. pygmaeus abelii* draft genome sequence. Similar to other primate genomes⁴², about 8% of the *P. pygmaeus abelii* draft genome sequence can be attributed to endogenous retroviruses. Endogenous retroviruses fall into the group of Class I elements and are Long Terminal Repeat (LTR)-retrotransposons. We found evidence for orangutan-lineage specific expansion of HERV-E (Human ERV-E). However, we did not recover any recent ERV insertions with less than 2% divergence from the consensus sequence indicating that ERVs may no longer be actively mobilizing in orangutans. In addition, we were unable to identify evidence of lineage-specific retroviral invasions into the *P. pygmaeus abelii* draft genome sequence. Consequently, no orangutan-lineage specific endogenous retrovirus subfamilies were identified.

The non-LTR retrotransposons are also class I elements. Non-LTR retrotransposons – currently ongoing retrotransposition in great ape genomes – are the autonomous Long INterspersed Elements (LINEs) and the non-autonomous *Alu* elements – a primate specific family of Short INterspersed Elements (SINEs) – and SVAs – a composite mobile element containing three subcomponents: SINE-R /VNTR (Variable Number of Tandem Repeats)/ and an *Alu*-like region^{10,42,44}. A full-length LINE (L1) is about 6 kb long and encodes two Open Reading Frames (ORFs). Just a small fraction of full-length L1s are capable of retrotransposition due to inactivating point mutations within the ORFs. In addition, the majority of L1s are 5' truncated upon insertion into primate genomes. L1 is also responsible for the insertion of its non-autonomous counterparts, SVA and *Alu* elements. Although the identification of active *Alu* elements is far less well understood than for L1, recent research into human *Alu* mobilization has identified

several factors that alter the retrotransposition activity of *Alu* elements. These include polyA-tail length, nucleotide substitutions within the polyA-tail, distance of polyA TTTT termination signal from the end of an *Alu* element, sequence variation from the consensus sequence of a currently active subfamily, and the interaction ability of SRP9/14 to build RNA/protein complexes^{45,46,47,48,49}.

Our initial analysis of the orangutan draft genome sequence (ponAbe2) queried a BLASTZ generated orangutan-human chain for retrotransposon insertions in the orangutan draft assembly. This method revealed that L1 elements are still very active in orangutan with over 4700 lineage specific insertions. Using the same approach, we retrieved only about 200 *Alu* elements. This was in sharp contrast to analyses of the human and chimpanzee genomes in which previously approximately 5,000 and 2,300 lineage specific *Alu* insertions were identified, respectively^{10,42}. The orangutan estimate was believed to be potentially somewhat of an underestimate as some sections were not included in the stringent BLASTZ chains. Therefore, we performed a comparison of recent lineage-specific *Alu* insertions (less than 2% diverged from their respective consensus sequences) with the BLASTZ generated chain and showed that most of the very youngest insertions (41 out of 45) were overlooked using the BLASTZ approach. Consequently, we searched the orangutan genome draft sequence for the presence of all *Alu* insertions that corresponded to an indel in the human genome [hg36.2]. Using this approach we identified 248 orangutan lineage-specific *Alu* insertions. Thus, we estimate that the orangutan draft genome contains approximately 250 lineage-specific *Alu* insertions.

In addition to the computational approaches described above, evidence of a low retrotransposition rate of *Alu* insertions in the orangutan genome was further supported by wet bench experimentation. First, we used polymerase chain reaction (PCR) to test all young looking *Alu* candidate loci on a phylogenetic DNA panel including 7 orangutan individuals (5 Sumatran and 2 Bornean) and 9 other primate species (Table S9-1). Only a small fraction (30%; 13 of 44) were shown to be polymorphic within Sumatran and Bornean orangutans. The fixation of the majority of *Alu* insertions provides strong evidence that even the youngest appearing elements identified in the orangutan draft genome, inserted before the divergence of Bornean and Sumatran orangutans and thus *Alu* elements have been relatively inactive.

Next, we implemented a display-based PCR strategy termed Allele-Specific *Alu* PCR (ASAP)³⁹ using two orangutan lineage specific *Alu* subfamilies. Using this approach we identified 167 subfamily specific *Alu* insertions from the genome of a Bornean orangutan (AG05252, Coriell Cell Repositories). Only one of the 167 was not found in the reference genome of "Susie" (PR01109) and this was due to a region of N's in the assembly proximal to the insertion site. In addition, of the 167 elements retrieved by ASAP, many were identified anywhere from 3 to 17 times each. In other words, we were detecting the same young *Alu* insertions over and over again, further supporting a low retrotransposition rate and low copy number of young *Alu* insertions in the orangutan lineage.

	Species Names	Common Names	Origin	ID number
1	<i>Homo Sapiens</i>	Human	ATCC	HeLa CCL-2
2	<i>Pan paniscus</i>	Bonobo	IPBIR	PR00661
3	<i>Pan troglodytes</i>	Common Chimpanzee	IPBIR	NS06006
4	<i>Gorilla gorilla</i>	Lowland gorilla	Coriell ^a	AG05251
5	<i>Pongo pygmaeus pygmaeus</i>	Bornean orangutan	SDFZ ^b	OR315
6	<i>Pongo pygmaeus pygmaeus</i>	Bornean orangutan	Coriell ^a	AG05252A
7	<i>Pongo pygmaeus abelii</i>	Sumatran orangutan	Coriell ^a	GM06213A
8	<i>Pongo pygmaeus abelii</i>	Sumatran orangutan	Coriell ^a	GM04272A
9	<i>Pongo pygmaeus abelii</i>	Sumatran orangutan	Coriell ^a	NG12256
10	<i>Pongo pygmaeus abelii</i>	Sumatran orangutan	Coriell ^a	NG06209
11	<i>Pongo pygmaeus abelii</i>	Sumatran orangutan	SDFZ ^b	OR823
12	<i>Hylobates syndactylus syndactylus</i>	Siamang	IPBIR	PR00598
13	<i>Hylobates syndactylus syndactylus</i>	Siamang	SDFZ ^b	KB11539
14	<i>Hylobates lar</i>	white-handed gibbon	IPBIR	PR00715
15	<i>Hylobates gabriellae</i>	red-cheeked gibbon	IPBIR	PR00652
16	<i>Chlorocebus aethiops</i>	African Green Monkey	ATCC	CCL70
17	<i>Macaca mulatta</i>	Rhesus Macaque	Coriell ^a	NG07109

ATCC: From cell lines provided by the American Type Culture Collection

IPBIR: Integrated Primate Biomaterials and Information Resource

a Coriell Institute for Medical Research, 403 Haddon Avenue, Camden, NJ

b San Diego Frozen Zoo,
Conservation and Research for Endangered Species (CRES)

Supplemental Section S10 – Processed Pseudogene Formation

Processed pseudogenes, or retrocopies, represent the vast majority of pseudogenes in mammalian genomes and derive from mRNAs that are reverse-transcribed into cDNA and then inserted into the genome^{50,51}. We assessed the rate of retrocopy formation on each branch of the phylogeny composed by human, chimp, orangutan, and rhesus macaque (Figure S10-1). Because of the retrotransposition process, in most cases gene retrocopies inserted in the genome lack introns that are present in their parent genes⁵¹. Our method to identified retrocopies relied on this structural difference with parent genes, as has been previously applied to several genome-wide studies^{52,53}.

We initially retrieved coding sequences corresponding to 9,743 one-to-one orthologous genes shared by these species; this data set was chosen in order to correct for differences in the number of annotated genes from different genomes. In addition, the lack of duplicated genes in this dataset made the identification of pseudogenes more straightforward. Orthologs with no introns in their coding region were excluded from the analysis, as it would be not possible to infer whether copies of these genes were

generated by retrotransposition or DNA duplication given the method used. The final gene set included 8,701 one-to-one orthologs.

For each genome, the species-specific set of coding nucleotide sequences was used as query in a BLASTn⁵⁴ search against the respective genome assembly. Every blast hit with at least 70% identity and 70% coverage (length) compared to the query was retrieved. We verified that these hits were overlapping two or more exons of the original queries, thus representing *bona fide* retrocopies, by using BLAT to search against the genome assemblies at the UCSC Genome Browser (<http://genome.ucsc.edu/>) and inspecting the match with the parent gene. Intronless pseudogenes that appeared to have originated via DNA duplication of existing processed pseudogenes were also removed.

Finally, each retrocopy was assigned to a specific branch in the four species phylogeny by using synteny information retrieved from the UCSC Genome Browser (<http://genome.ucsc.edu/>) (Figure S10-1). We considered genome position information from human, chimp, orangutan, and macaque. For instance, retrocopies generated in the human lineage only were required to be absent in the orthologous genomic position in chimp, orangutan, and macaque. Similarly, processed pseudogenes present in the orangutan genome but not in the other three primate genomes were assigned to the orangutan branch of the phylogenetic tree. While it is possible that the number of retrocopies specific to the rhesus macaque was slightly overestimated because of possible losses of some of them in the common ancestor of human, chimp, and orangutan—and vice versa—it is unlikely that our estimates along the hominid branches were significantly affected, as this would require losses in several lineages.

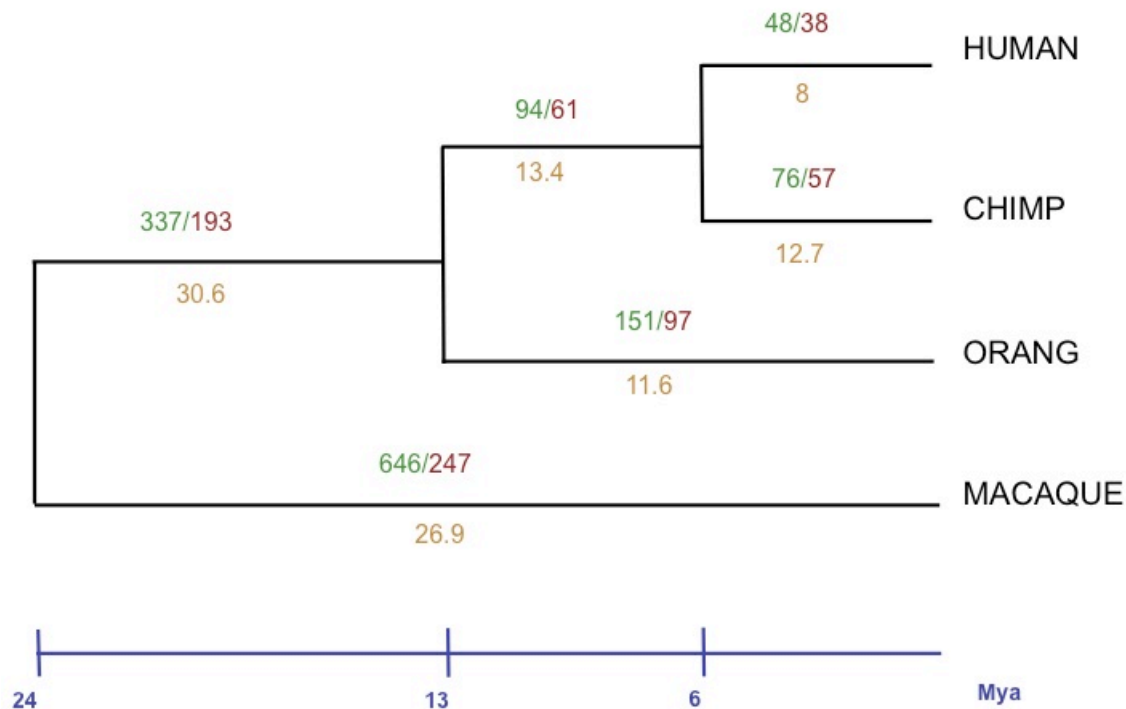


Figure S10-1. Rate of retrocopies formation in hominids and macaque. Retrocopies were obtained using genes with more than one coding exon as queries for similarity searches against the genome sequence of each species. To normalize for different gene content of different genomes, only 1-to-1 orthologs were used as queries. Intronless sequences with more than 80% identity over more than 70% of the length of queries were counted as retrocopies. Lineage-specific copies were inferred from this dataset using syntenic information (Blastz alignments) available at the UCSC Genome Browser (<http://genome.ucsc.edu/>). Green: number of retrocopies. Red: number of parent genes. Orange: retrocopies generated per Mya. Age of divergence among lineages is indicated.

Supplemental Section S11 – Segmental Duplications and Structural Variation

Assessment of Orangutan Segmental Duplications

Two different methods were used to detect genome duplications in the orangutan assembly; one dependent on the assembly itself (WGAC) and one based on an assessment of excess depth-of-coverage of whole-genome shotgun sequence data (WSSD). The BLAST-based whole genome assembly comparison (WGAC) method was used to identify pairwise alignments representing >1 kb and >90% identity⁵⁵. As larger, high-identity duplications (>94%) are frequently collapsed within working draft sequence assemblies, we compared these assembly-based results to whole genome shotgun sequence detection (WSSD) database of orangutan segmental duplications⁵⁶. The WSSD approach identifies regions >10 kb in length with a significant excess of

high-quality WGS read depth within overlapping 5 kb windows. WSSD analysis was based mapping 25,514,441 orangutan WGS reads back to the assembly. The reads were mapped based on the following criteria (>94% sequence identity; >200 bp non-RepeatMasked bp and at least 200 bp of PhredQ > 30).

In the analysis of the released assembly (v2.0.2.) a total of 349.65 Mb (12.8%) of non-redundant sequence was detected by the BLAST-based WGAC (assembly dependent) method (>1 kb and >90% sequence identity). Of that sequence, 159.45 Mb (6.7%) were found within the autosomes and sex chromosomes (the remaining 190.2 Mb were localized to chromosomal random bins and unplaced contigs (chrUn)). More non-redundant duplication basepairs were mapped to intrachromosomal duplications (284 Mb intra vs. 200 Mb interchromosomal). A total of 194.62 Mb (7.1%) of duplicated sequences (>94%; >10 kb) were predicted based on WSSD analysis (assembly independent method). Only 54.42 Mb (2.3%) are in autosomes and sex chromosomes while 135.18 Mb are in Chromosome Random and unplaced contigs (chrUn). Only 31% (110 Mb/350 Mb) of the duplications detected by assembly methods (WGAC) were supported by WSSD.

Given the low concordance (31%) of duplication estimates based on WGAC and WSSD analysis of v2.0.2 we also analyzed a “relaxed stringency” version of the assembly (see Supplemental Section S1, v2.2). We found that this version has a significant reduction of total WGAC pairs (from 152,443 to 104,533), especially the interchromosomal WGAC pairs (Table S11-1), and the identity distribution showed a significant reduction of high-identity WGAC pairs (>99%), which are frequently assembly artifacts (data not shown). In general, the two versions are similar for larger SD (>10 kb), but v2.2 demonstrated greater support of the intersection of assembly dependent and independent method (50% over the previous 31%). The union of WGAC and WSSD intersect, plus WGAC < 94% identity (most WGAC with lower identity are likely real), is 113 Mb (~3.8%). This value is less than what was found in the human genome⁵⁶ and agrees with the increased rate of duplication in the common ancestor of humans and chimpanzees⁵⁷.

Table S11-1. Summary duplication statistics for orangutan assemblies.

Category	v2.0.2	v2.2
Total genome length	3.1Gb	2.98Gb
Chrom length	2.7Gb	2.73Gb
Number of WGAC pairs	152443	104533
Number of inter chrom	112627	35129
Number of intra chrom	39816	69404
nr length	349.6 mb	188 mb
nr length of inter chrom	200.8 mb	96 mb
nr length of intra chrom	284 mb	134 mb
WSSD	194 mb	174 mb
Shared WGAC/WSSD	110 mb	94 mb
Percentage WGAC supported WSSD	31.52%	50.00%

Assessment of Structural Variation in the Orangutan Genome

We have applied several complementary strategies to detect structural variants in chimpanzee and orangutan (and hence in humans, if one assume parsimoniously that an event shared for those two species would have occurred specifically in the human lineage). We used a clone end pair mapping approach using chimpanzee and orangutan fosmids, BACs and plasmids (derived from CHORI-1251 and CHORI-276 Libraries, respectively). All end-sequences were optimally aligned and paired against the reference human genome sequence (hg17) using a previously described method^{58,59}. There are four steps: (1) initial recruitment of end-sequences, (2) optimal realignment with quality rescoring, (3) determination of paired-end read placements, and (4) rearrangement detection. We considered only those orangutan and chimpanzee 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. We applied further criteria and sites spanning more than 1 Mb and overlapping any given GAP in the human genome were also removed. In all the analyses, all the calls were performed with at least 2 high-quality mapped clones supporting the site. However, and due to the coverage of the orangutan fosmid library we considered orangutan sites supported only by a single ESP placement and limited our analysis to deletions supported by at least 2 clones.

As a second measure, and to minimize false positives, we also detected deletions in the chimpanzee and the orangutan genomes by assessing the coverage of WGS reads against the human reference genome. We first mapped the ~31 million chimpanzee and ~25 million orangutan WGS reads the human reference genome (hg17) using MEGABLAST v.2.2.1. We then excluded all common repeats defined by RepeatMasker with less than 10% sequence divergence from their consensus, as well as primate-specific L1P and satellite repeat sequences. The objective was to detect sites with low coverage in read mapping (< 10 reads) suggesting homozygous or heterozygous deletions. Our classification criteria are summarized in Table S11-2.

Table S11-2. Classification of events detected by paired end methods.

	Chimpanzee End Pair	Orangutan End Pair	WGS reads
Chimpanzee specific Deletions	>= 2 Large clones mapping	0 Discordant clones	< 10 reads WGS Chimp
Orangutan specific Deletions	0 Discordant clones	>= 2 Large clones mapping >= 1 Large clones mapping (Fosmids)	< 10 reads WGS Orang
Human specific Insertion	>= 2 Large clones mapping	>= 2 Large clones mapping >= 1 Large clones mapping (Fosmids)	< 10 reads WGS Chimp and Orang

We detected 71 putative deletions (38 chimpanzee deletions, 29 orangutan deletions and 4 human specific insertions) affecting a total of 5.6 Mb using our classification scheme. This is an upper bound since the end mapping approach is inaccurate with

respect to the exact breakpoints of an event. We then performed array comparative genomic hybridization to confirm the deletions. We used a customized oligonucleotide microarrays (NimbleGen, 385,000 isothermal probes) targeted specifically to the deletions. We performed 3 hybridizations, human (G248) against chimpanzee (“Clint”), human (G248) against orangutan (“Susie”), and chimpanzee (“Clint”) against orangutan (“Susie”). We validated 58/71 deletions (82%). Combined with the validation data, we detected 38 chimpanzee deletions and 15 orangutan deletions encompassing a total length of 4.1 Mb, comprising 28 complete and 8 partial genes (Tables S11-3 & S11-4). Selected examples are presented as Figures S11-1 and S11-2 below.

Table S11-3. Summary of chimpanzee and orangutan deletions.

	Number of events	Length (bp)
Chimp Deletions	38	2,253,998
Orangutan Deletions	15	1,811,248
Total	53	4,065,246

Table S11-4. Chimpanzee and orangutan deleted genes.

GeneID (Description)	Complete/Partial	Category
NM_001005182-NP_001005182-OR6C1-olfactory receptor, family 6, subfamily C,	Complete	Chimpanzee Specific DEL
NM_001005183-NP_001005183-OR6C76-olfactory receptor, family 6, subfamily C,	Complete	Chimpanzee Specific DEL
NM_001005497-NP_001005497-OR6C75-olfactory receptor, family 6, subfamily C,	Complete	Chimpanzee Specific DEL
NM_001005499-NP_001005499-OR6C70-olfactory receptor, family 6, subfamily C,	Complete	Chimpanzee Specific DEL
NM_001005518-NP_001005518-OR6C65-olfactory receptor, family 6, subfamily C,	Complete	Chimpanzee Specific DEL
NM_001005519-NP_001005519-OR6C68-olfactory receptor, family 6, subfamily C,	Complete	Chimpanzee Specific DEL
NM_001039884-NP_001034973-ZNF826-zinc finger protein 826	Complete	Chimpanzee Specific DEL
NM_001098506-NP_001091976-CEACAM21-carcinoembryonic antigen-related cell adhesion	Complete	Chimpanzee Specific DEL
NM_014439-NP_055254-IL1F7-interleukin 1 family, member 7 isoform 1	Complete	Chimpanzee Specific DEL
NM_021571-NP_067546-CARD18-ICEBERG caspase-1 inhibitor	Complete	Chimpanzee Specific DEL
NM_054104-NP_473445-OR6C3-olfactory receptor, family 6, subfamily C,	Complete	Chimpanzee Specific DEL
NM_054105-NP_473446-OR6C2-olfactory receptor, family 6, subfamily C,	Complete	Chimpanzee Specific DEL
NM_181600-NP_853631-KRTAP13-4-keratin associated protein 13-4	Complete	Chimpanzee Specific DEL
NM_181622-NP_853653-KRTAP13-3-keratin associated protein 13-3	Complete	Chimpanzee Specific DEL
NM_015678-NP_056493-NBEA-neurobeachin]	Partial	Chimpanzee Specific DEL
NM_018465-NP_060935-C9orf46-hypothetical protein LOC55848]	Partial	Chimpanzee Specific DEL
NM_019609-NP_062555-CPXM1-carboxypeptidase X, member 1 precursor]	Partial	Chimpanzee Specific DEL
NM_030643-NP_085146-APOL4-apolipoprotein L4 isoform 1]	Partial	Chimpanzee Specific DEL
NM_152347-NP_689560-C17orf57-hypothetical protein LOC124989]	Partial	Chimpanzee Specific DEL
NM_001001913-NP_001001913-OR52N1-olfactory receptor, family 52, subfamily N,	Complete	Orangutan Specific DEL
NM_001002905-NP_001002905-OR8G1-olfactory receptor, family 8, subfamily G,	Complete	Orangutan Specific DEL
NM_001003443-NP_001003443-OR56A3-olfactory receptor, family 56, subfamily A,	Complete	Orangutan Specific DEL
NM_001005165-NP_001005165-OR52E4-olfactory receptor, family 52, subfamily E,	Complete	Orangutan Specific DEL
NM_001005167-NP_001005167-OR52E6-olfactory receptor, family 52, subfamily E,	Complete	Orangutan Specific DEL
NM_001005168-NP_001005168-OR52E8-olfactory receptor, family 52, subfamily E,	Complete	Orangutan Specific DEL
NM_001005174-NP_001005174-OR52N2-olfactory receptor, family 52, subfamily N,	Complete	Orangutan Specific DEL
NM_001005198-NP_001005198-OR8G5-olfactory receptor, family 8, subfamily G,	Complete	Orangutan Specific DEL
NM_001005338-NP_001005338-OR5H1-olfactory receptor, family 5, subfamily H,	Complete	Orangutan Specific DEL
NM_001005479-NP_001005479-OR5H6-olfactory receptor, family 5, subfamily H,	Complete	Orangutan Specific DEL
NM_001005514-NP_001005514-OR5H14-olfactory receptor, family 5, subfamily H,	Complete	Orangutan Specific DEL
NM_001005515-NP_001005515-OR5H15-olfactory receptor, family 5, subfamily H,	Complete	Orangutan Specific DEL
NM_052997-NP_443723-ANKRD30A-ankyrin repeat domain 30A	Complete	Orangutan Specific DEL
NM_138800-NP_620155-TRIM43-tripartite motif-containing 43	Complete	Orangutan Specific DEL
NM_001005482-NP_001005482-OR5H2-olfactory receptor, family 5, subfamily H,]	Partial	Orangutan Specific DEL
NM_021943-NP_068762-ZFAND3-zinc finger, AN1-type domain 3]	Partial	Orangutan Specific DEL
NM_033056-NP_149045-PCDH15-protocadherin 15 isoform CD1-4 precursor]	Partial	Orangutan Specific DEL

Figure S11-1. Example of an orangutan specific deletion of a complete gene (olfactory receptor, family 8, subfamily G)

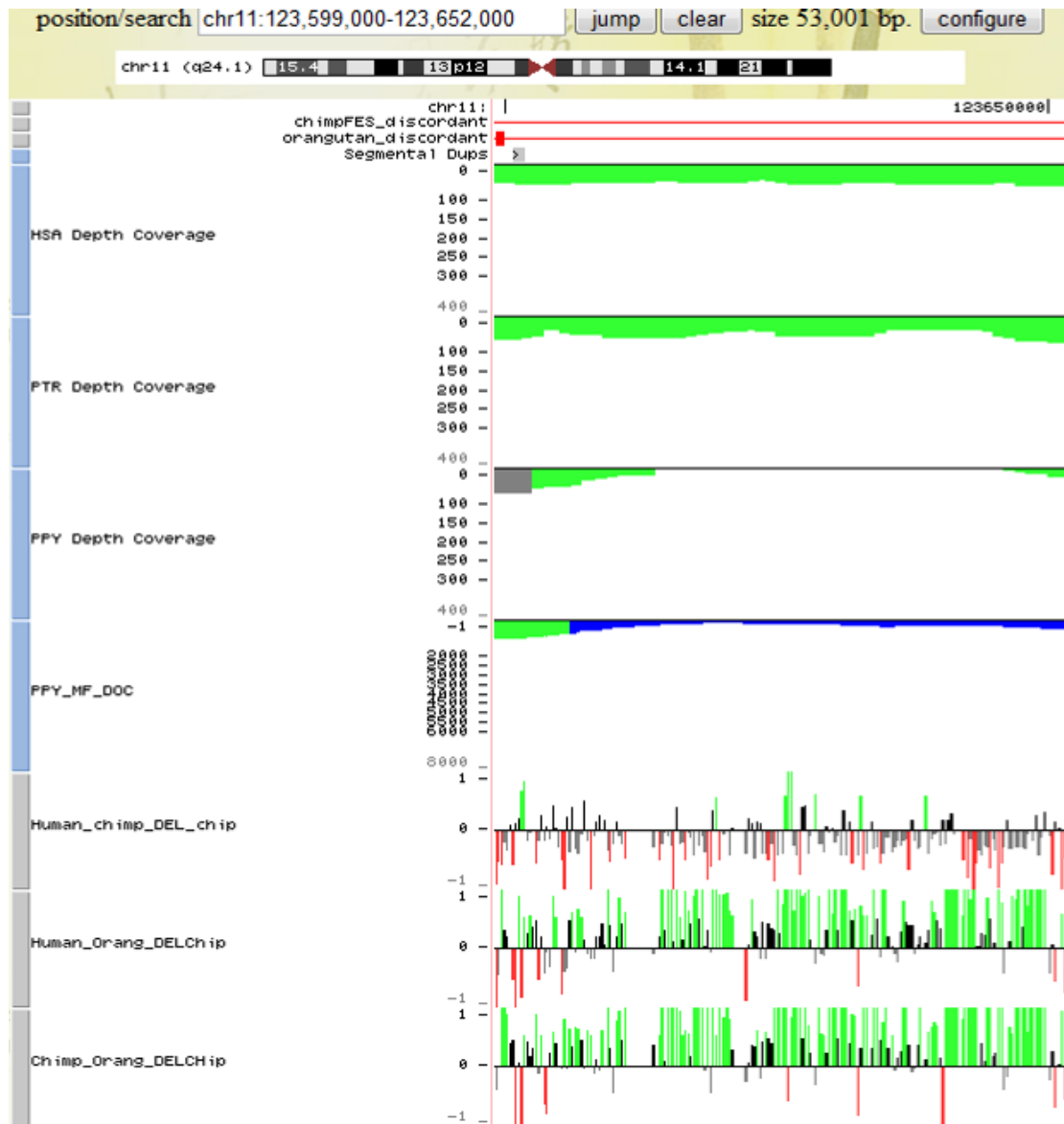
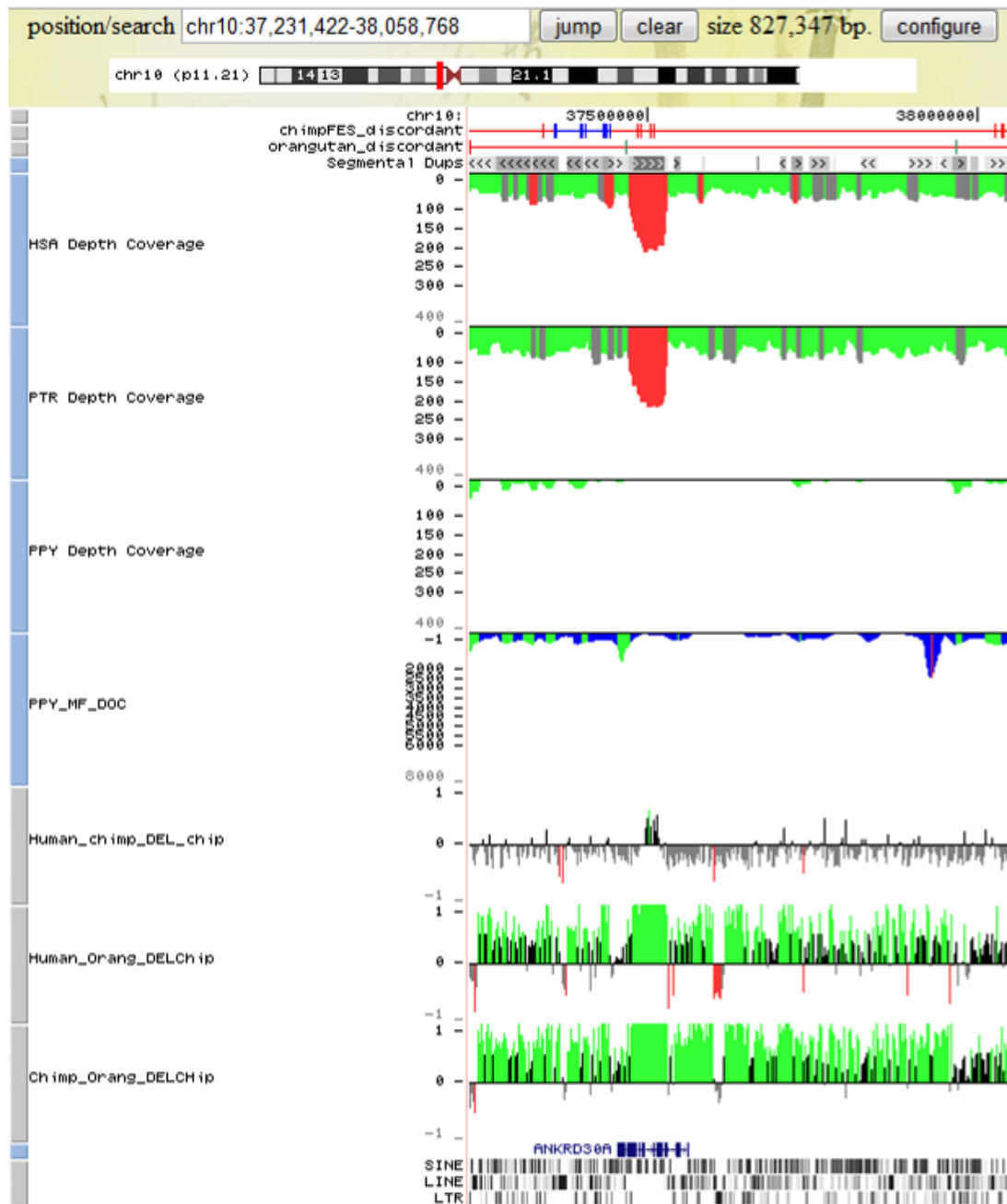


Figure S11-2. The largest orangutan specific complete gene deletion (ankyrin repeat domain 30A). This gene is partially duplicated in humans and chimpanzees.



Supplemental Section S12 – Great Ape Gene Family Expansion

Exon sequences were acquired for all protein-coding genes in human, chimpanzee, orangutan, rhesus macaque, mouse, rat, and dog from Ensembl v49. The genome assemblies corresponding to these genomes are as follows: NCBI36 for human, CHIMP2.1 for chimpanzee, PPYG2 (aka v2.0.2) for orangutan, MMUL1 for macaque, NCBI37 for mouse, RGSC3.4 for rat, and CanFam2.0 for dog. For each gene, any overlapping exon sequences from alternative transcripts were merged, and exons were then concatenated. When UTR sequences were available for the 5'- and 3'-most exons of the gene, they were removed from the concatenated sequences, resulting in the full set of protein-coding DNA for each annotated gene. In cases where multiple UTRs were present for the same exon, only the smallest UTR sequence was removed (i.e. sequence that was protein-coding in any transcript was included).

Gene families were defined using the MCL clustering algorithm⁶⁰. Each of the concatenated exonic sequences constituting a single gene from all genomes (150,127 genes total) were BLASTed against every other gene in all species (BLASTn). A weighted undirected graph was then created, where genes are represented as nodes. Gene pairs where the average BLAST E-values $\leq 10^{-2}$ were connected by an edge, with the weight of the edge equal to the negative log of their average E-value. MCL was then run on this graph using an inflation parameter of 2.3⁶¹; this resulted in 25,777 gene families. Families with members only in primates, only in rodents, or only in dog were then removed from the set of gene families to avoid the problem of inferring ancestral states for families that are not as old as the ancestor of all the mammals considered here. This left 15,960 mammalian gene families including 127,311 genes for our final analysis.

In order to estimate rates of gene gain and loss, we applied an updated version of the likelihood model originally proposed in Hahn et al. (2005) and implemented in the software package CAFE v2.1^{36,61,62,63}. This method models gene family evolution as a stochastic birth and death process, where genes are gained and lost independently along each branch of a phylogenetic tree. A parameter, λ , describes the rate of change as the probability that a gene family either expands (via gene gain) or contracts (via gene loss) per gene per million years, and can now be estimated independently for all branches. For gene families inferred to be present in the MRCA of mammals ($n=15,960$), parameters are estimated by maximizing the likelihood of the observed family sizes.

Previous analyses had found that a 3-parameter model best fit the available genome sequence data, with the human and chimp branches evolving at the fastest rate, followed by the macaque lineage, followed by the rodent and dog lineages⁶⁴. In order to test initially whether orangutan also showed an accelerated rate of gene family evolution, we compared two different 3-parameter models: one with orangutan grouped with the human and chimp branches, and one with orangutan grouped with the macaque branch (holding the other 2 parameter assignments the same). We found that

the latter model was a significantly better fit to the data using a likelihood ratio test ($P < 1 \times 10^{-9}$), indicating that the acceleration occurred after the split between human+chimp and orangutan (Figure S12-1). As a further test of this hypothesis, we estimated independent rate parameters for all 12 branches of the mammalian tree, running the maximum likelihood estimator three times to ensure convergence. The estimated l -values for each branch were:

(((((chimp_0.012712:6,human_0.008302:6)_0.010194:7,orang_0.005700:13)_0.001290:18,macaque_0.004832:24)_0.000661:63,(mouse_0.003301:17,rat_0.005108:17)_0.001104:70)_0.004785:6,dog_0.000502:93)

These data also indicate that the rates for the human and chimpanzee branches (and the human-chimpanzee ancestor) are higher than the rates in the rest of the tree.

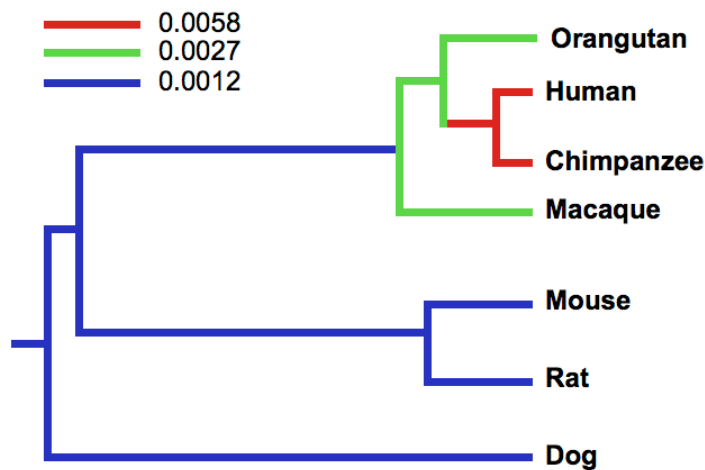


Figure S12-1. Rates of gene family evolution among select sequenced mammals. Numerical values represent gain/loss events per gene per million years. Note the acceleration in the human-chimpanzee lineage compared to orangutan and macaque, and the slower rate among non-primates.

Supplemental Section S13 – Protease Gene Families

Introduction

Proteases form a diverse group of enzymes that share the ability to hydrolyse peptide bonds. The biological and pathological significance of this enzymatic activity has prompted the definition of the degradome as the complete repertoire of proteases in an organism⁶⁵. From a genomic point of view, the degradome is highly attractive for several reasons. First, the degradome is composed of a large number of genes. Thus, the human degradome includes more than 560 protease genes, which represents about 1.7% of the total annotated human genes⁶⁶. Additionally, protease genes form a very diverse group in sequence and genomic organization. Human proteases can be grouped into five unrelated catalytic classes, which can be further subdivided into 67 different families. Thus, while proteases share a biochemical function, catalytic domains exhibit a high sequence diversity, which is further increased by the frequent attachment of auxiliary, non-proteolytic domains to the catalytic moieties⁶⁷. Some of the protease genes have been shown to occur in genomic clusters, which is convenient for the study of short-term evolution. By contrast, most of the protease genes are randomly distributed throughout the annotated genomes. Therefore, the degradome forms a representative subset of the coding genome of a species. Finally, the genomic study of the degradome may be performed using a semi-automated approach, including manual curation for a high-quality result. Importantly, previous studies have rendered the virtually complete sets of manually curated protease sequences, corresponding to the human, chimpanzee, mouse, rat, and platypus degradomes^{66 68 69 70}. We intend to use this information to characterize the orangutan degradome, which, added to the data originated from human and chimpanzee, is expected to extend our knowledge about the hominization process.

Results and discussion

We have used the set of human protease sequences to search for orthologous sequences in the orangutan genome. We have validated multiple dubious results by inspection of raw genomic traces, as well as two EST sets (GM04272 and PR01109, Supplemental Section S5). As expected, we have found that the orangutan degradome is highly similar to its human and chimpanzee counterparts. Thus, almost all of the human proteases have a closely related orangutan orthologue. However, we have found a number of differences with interesting physiological and pathological implications. In agreement with previous works⁶⁹ most of these differences putatively affect the immune and reproductive systems. Interestingly, some differences may also affect brain physiology.

Immune system

- **PRSS33**, which is a functional protease gene in human, has been pseudogenized in the orangutan genome due to the presence of a premature stop codon in exon 3. PRSS33 is a macrophage-specific serine-protease whose expression is up-regulated in activated macrophages⁷¹. Strikingly, it has been demonstrated that the chimpanzee PRSS33 orthologue has been deleted by an Alu-mediated recombination mechanism^{69,72}, whereas the Rhesus monkey counterpart also has a premature stop codon at a different position (Figure S13-1). These data show that PRSS33 has been independently inactivated in several primates, but not in human.
- Orangutan **NAPB** contains all of the features of a functional aspartyl-protease gene, including a stop codon at the equivalent position of its paralogue NAPA. In contrast, the human orthologue of NAPB is predicted to be a transcribed pseudogene, since no stop codon has been found in its ORF. The human NAPB mRNA has been found in spleen, thymus, and several types of lymphoid and myeloid cells⁷³.
- The *tre-2* oncogene (**USP6**) seems to have been specifically pseudogenized in orangutan. This cysteine-protease gene, which has been shown to be hominoid-specific, arose from the fusion of duplicates from USP32 and TBC1D3⁷⁴, as can be seen in Figure S13-2. While functional in both human and chimpanzee, the orangutan orthologue of USP6 displays a premature stop codon after the N-terminal TBC-like domain, which presumably prevents the translation of the proteolytic domain. USP6 was first identified by its ability to transform mouse NIH3T3 cells⁷⁴ and its ectopic expression through genomic transpositions has been implicated in bone malignancies^{75,76}. The link of human USP6 to the immune system has been recently discovered, after the finding that the expression of this gene is necessary for HIV infection⁷⁷.
- Orangutan **CASP12** is a functional cysteine-protease gene, whereas its human orthologue is a pseudogene or encodes an inactive protease⁷⁸. The expression of a functional CASP12 transiently inhibits the activity of caspase-1, which slows down inflammatory cytokine processing in response to septic infections⁷⁹. This change is in agreement with previous reports suggesting that the human immune system has been selected for overactivation of the immune cells⁸⁰.
- Orangutan seems to have three functional **haptoglobin-like** genes. In hominoids and some primates, haptoglobins are located in a cluster which shows evidence for multiple events of duplication, deletion, and conversion^{81,82}. Three different loci named **HP** (haptoglobin), **HPR** (haptoglobin-related), and **HPP** (haptoglobin primate) have been characterized in this cluster. However, none of the primate genomes studied up to date show all three functional loci. Thus, humans have lost HPP after a deletion event, and chimpanzees show a truncated copy of HPR⁸¹. Likewise, rhesus monkey shows one or two haptoglobin-like loci with evidence for conversion events⁸². Therefore, orangutan seems to be the first organism known to contain all three characterized loci for haptoglobin genes. Haptoglobins have been shown to participate in the infections caused by *Trypanosoma brucei*. Haptoglobins are

contained in serum high-density lipoproteins (HDLs) and bind hemoglobin in a complex that is recognized and endocytosed by parasites like *T. brucei*, thereby competing for iron with the host. However, HDLs from humans and other species contain ApoL1, which acts as a lytic factor if endocytosed by the parasite. In these cases, haptoglobins help the intake of this lytic factor, thus helping the immune system from the host⁸³. Since orangutans lack ApoL1, the presence of three haptoglobins is expected to be detrimental for the host.

- Moreover, we have confirmed and extended results from previous studies on tryptases and chymases, which play an important role in mast cell biology⁸⁴. Orangutan **δ -tryptase** seems to be a fully functional protein, which confirms a previous study⁸⁵. In humans and chimpanzees, δ -tryptase shows a premature stop codon and an R-3Q mutation that hampers protease activation. We have also found that the chymase cluster in orangutan contains the same genes as in humans. In mammals, this cluster has frequently evolved by tandem duplication events⁸⁶. Notably, **CTSG**, which is contained in the chymase cluster, encodes one of the most diverging proteases between orangutans and humans (85% identities). These data suggest that tryptases and chymases evolved in hominoids by mutations and inactivations, rather than by duplication events.

Reproductive system

- The **TESSP2** serine protease gene displays a stop codon in its second exon in orangutan. In humans, this gene is functional and specifically expressed in meiotic or postmeiotic spermatogenic cells⁸⁷.
- Orangutan **ADAM6** metalloprotease gene appears to be functional, whereas its human orthologue is a pseudogene due to a premature stop codon. Interestingly, chimpanzee ADAM6 is also a pseudogene due to a different mutation, which suggests that it has been independently pseudogenized. In rat, ADAM6 is one of the genes that are specifically expressed in meiotic germ cells and may play a role in regulation of fertility⁸⁸.
- Likewise, orangutan **ISP2** is a functional gene, while its human and chimpanzee counterparts have been pseudogenized. Interestingly, both human and chimpanzee ISP2 genes contain a mutation that abolishes the putative first methionine residue, suggesting that this pseudogenization event occurred in a common ancestor to humans and chimpanzees, but not to orangutans. On the other hand, the orangutan genome lacks any orthologue of the related ISP1 gene, present in mouse and rat. In those organisms, ISP1 and ISP2 products have been suggested to form a heterodimer that functions as a hatching enzyme, allowing the preimplantation embryo to digest the surrounding *zona pellucida* and invade the uterus⁸⁹. If this is confirmed, and given the lack of conservation of ISP1 in orangutans, it seems likely that alternative proteolytic mechanisms have taken over this conserved function in primates, as previously suggested for humans⁹⁰.
- The orangutan genome contains a one-exon complete ORF almost identical to the metalloprotease **PA2G4** (also called EBP-1). This novel PAG2-like gene has no orthologues in humans or chimpanzees. The product of PA2G4 has been implicated in cellular proliferation and immune response to influenza^{91 92}. Several

one-exon PA2G4-like pseudogenes are known in humans⁹³. We have also found similar pseudogenes in marmoset, rhesus monkey, and chimpanzee. Only the orangutan ORF is complete and contains an in-frame stop codon. Interestingly, this putative PA2G4-like gene is inserted between the first and second exons of LRRC1, a conserved gene mainly expressed in placenta. This suggests that orangutan PA2G4-like may be a functional gene sharing the promoter of LRRC1. Therefore, PA2G4-like may be an orangutan-specific gene expressed in placenta.

- In orangutans, **HTRA4** is a pseudogene because of an in-frame stop codon in its first exon. This gene is functional in other mammals from mouse to humans. The HtrA family of serine-proteases appears to be involved in important processes, such as cell growth, apoptosis, inflammatory reactions, and control of cell fate via regulated protein metabolism⁹⁴. Interestingly, HTRA3 may play a role during placentation⁹⁵, and human ESTs from HTRA4 suggest that this gene features a placental-specific expression⁹⁴. Therefore, loss of HTRA4 seems to be an orangutan-specific feature with potential consequences in reproductive processes.
- We have found that orangutan **KLK3** is a functional gene similar to its human orthologue. This protease, also called prostate specific antigen, degrades semenogelins and changes the physical properties of ejaculated semen. Interestingly, we have confirmed that orangutan semenogelin 1 is larger than its human orthologue due to an expansion of repetitive sequences⁹⁶. Therefore, the different properties of ejaculated semen in hominoids are best explained by changes on substrates of KLK3, and not on the protease itself.

Brain biology

- The current assembly of the orangutan genome predicts that **PRSS12** (neurotrypsin) is a pseudogene because of a frameshift at its first exon. A raw genomic trace (PPAC-blu78f11.b-1) unambiguously confirms this finding. However, a second high-quality raw genomic trace (PPAE-ahj61f04.g1.ab1) shows the same sequence without any frameshift or inactivating mutation. The possibility that this reflects a polymorphism in the orangutan population is very exciting, since lack of functional neurotrypsin has been linked to serious brain defects in multiple species. Namely, a 4 bp deletion in human neurotrypsin mRNA is believed to cause mental retardation⁹⁷. Likewise, a *Drosophila melanogaster* strain lacking the orthologue of neurotrypsin has been shown to suffer a long-term memory formation defect⁹⁸. If this polymorphism is confirmed, it would be extremely interesting to search for novel compensatory mechanisms for neurotrypsin defects in orangutans.
- Orangutan **PRSS3** (mesotrypsinogen and trypsinogen IV) shows a missense mutation at exon 3 which causes a premature stop codon. To confirm this result, a PCR amplification of this region followed by direct sequencing was performed. This experiment showed that the sequenced individual is heterozygous in this position. Thus, this individual has one functional and one non-functional allele of PRSS3, which suggests that there is a non-functional allele of PRSS3 in the orangutan population. PRSS3 is a trypsin-like serine protease expressed mainly

in the brain. This gene displays two alternatively spliced forms, named mesotrypsinogen and trypsinogen IV, which differ only in their first exon. The position of this mutation in orangutan suggests that both isoforms are inactivated. Interestingly, trypsinogen IV shows a brain-specific expression pattern, and has been shown to selectively activate proteinase activated receptor-1, but not proteinase activated receptor-2⁹⁹.

- Orangutan **CAPN7** (calpain-7) shows a frameshift at the beginning of exon 7 which causes a premature stop codon upstream from its protease domain. The human orthologue of this cysteine-protease gene is expressed in brain, and its product can cleave Htt (huntington disease protein) *in vitro*. Inhibition of this cleavage decreases Htt expanded polyglutamine toxicity¹⁰⁰. Interestingly, **CAPN14**, another less studied member of the calpain family, also seems to be a pseudogene in orangutans because of a similar frameshift at the beginning of exon 2. CAPN14 is a functional gene in human and chimpanzee, but it is not present in mouse and rat. These results need confirmation, since alternative splicing events might render both CAPN7 and CAPN14 active. Furthermore, although the genomic trace PPAC-anj48c07.g1 clearly shows the exon 7 frameshift for CAPN7, several ESTs contain the same sequence without any frameshift. A possible explanation for this discrepancy is that the sequenced individual has two different alleles, one functional and one non-functional.

Additional findings derived from orangutan degradome analysis

- **MMP23A** is also absent in orangutan reinforcing the idea that it is a human-specific gene, with no known orthologue in any other species.
- The orangutan orthologue of **CTRL** (chymotrypsin-like) is inactivated by a 1-bp insertion that disrupts its catalytic site. Interestingly, there is a similar allele described in humans (dbSNP rs35178715), although the most common allele is functional. The expression of CTRL in humans is restricted to pancreas¹⁰¹. Therefore, the pseudogenization of CTRL in orangutans may be related to differences in their digestive systems.
- Orangutan **KLKBL1** contains a premature stop codon in its second exon. Interestingly, KLKBL1 is one of the less conserved proteases between human and chimpanzee, with a 95.01% identity⁶⁹.
- We have found an orangutan-specific one-exon complete ORF almost identical to the β 1 proteasome subunit gene, **PSMB1**. The cellular concentration of PSMB1 determines to a large extent the composition and function of the proteasome. This PSMB1-like ORF may be a recently acquired pseudogene which still has not accumulated deleterious mutations. On the other hand, it is possible that this PSMB1-like ORF has evolved into a transcribed gene. Thus, this putative gene may affect the levels of PSMB1 product under certain circumstances, which in turn has potential implications for antigen presentation in response to interferon gamma stimulation¹⁰².
- Likewise, we have found an orangutan-specific one-exon complete ORF very similar to the C6.1A (**BRCC36**) metalloprotease gene. Interestingly, there are C6.1A-like orthologous pseudogenes both in human and chimpanzee. These pseudogenes show the same frameshifts that result in premature stop codons.

Rhesus monkey also has an equivalent pseudogene with two different frameshifts. By contrast, similar one-exon C6.1A-like complete ORFs can be found both in mouse and rat. Therefore, this ORF seems to have been specifically conserved in orangutan and pseudogenized in other primates. Its paralogue C6.1A is a subunit of the BRCC E3 ubiquitin ligase, which enhances cellular survival following DNA damage¹⁰³. Thus, C6.1A-like may be involved in cancer-protecting mechanisms following exposure to ionizing radiation.

- Finally, it is also noteworthy the presence of a Dobzhansky incompatibility in the gene encoding cationic trypsinogen (**PRSS1**). The orangutan orthologue of human PRSS1 displays an N29T change which, when present in humans, causes hereditary pancreatitis¹⁰⁴. Since the chimpanzee orthologue of this gene also shows the same N29T allele⁶⁹, this finding confirms that the ancestral residue for this position is disease-associated in humans.

In addition to these specific events discussed above, the orangutan degradome analysis has also extended some previous ideas regarding primate evolution. Thus, it is remarkable the finding that several protease genes have been independently inactivated by different means in hominoids, as shown in Table 1. For example, the cysteine-protease **USP50** gene produces an active protease in mice and rats. By contrast, chimpanzee and human USP50 are devoid of peptidase activity due to the absence in both cases of the last catalytic residue of the protease, as a result of two different premature stop codons¹⁰⁵. Orangutan USP50 shows a frameshift after its third catalytic residue, which results in loss of a conserved sequence (Figure S13-3). Likewise, the tryptase **PRSS34** gene produces an active serine-protease in mice and rats and has been independently inactivated in orangutans and humans. While human PRSS34 exhibits a mutation at its catalytic serine residue, orangutan PRSS34 has a premature stop codon that prevents the translation of its catalytic serine residue. Finally, some genes have been independently inactivated in some primates and conserved in others. These are the above discussed cases of **PRSS33** which is lacking in chimpanzees and is pseudogenized in orangutans, but is active in humans, and **ADAM6** which is active in orangutan but is pseudogenized in human and chimpanzee.

Taken together, these data confirm that speciation in the hominoid lineage has been heavily influenced by the immune and reproductive systems. In addition, we have found several potential changes in the degradomes of orangutans, chimpanzees, and humans that may affect brain biology. Finally, the finding that several proteases have been independently inactivated in hominoids suggests that these inactivation events may have been important during the hominization process.

Technical comments

To perform this analysis, we used a previously generated set of manually curated human protease sequences (<http://www.uniovi.es/degradome>). We used these sequences to probe the 2.0.2 version of the orangutan assembly. Each resulting alignment file was then semi-automatically rebuilt into a gene. To speed up this process, we have generated two Perl scripts. The first script takes a *tblastn* file and attempts to

predict the most likely reconstruction of the gene based on the hits. The user can change that prediction manually. The second script allows the user to manually choose the intron/exon junctions of the gene. We plan to extend and improve these scripts to add flexibility in the hope that it may be useful in the manual annotation of genomes.

Table S13-1. Summary of differences between orangutan (O), chimpanzee (C), and human (H) degradomes. Bolts represent pseudogenization events.








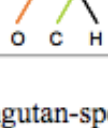







Gene	Events	Function
PRSS33		Host defense (?)
NAPB		Expressed in lymphoid tissues
USP6		HIV infection factor
CASP12		Host defense
HPP		Host defense (?)
TESSP2		Expressed in spermatogenic cells
ADAM6		Expressed in meiotic germ cells
ISP2		Embryo hatching and implantation
PA2G4-like	Orangutan-specific	Placental expression (?)
HTRA4		Placental expression (?)

Table S13-1. Continued.

PRSS12		Brain development
PRSS3		Brain homeostasis
CAPN7, CAPN14		Htt cleavage
MMP23A	Human-specific	Extracellular matrix remodelling
CTRL		Digestion
KLKBL1		Unknown
PSMB1-like	Orangutan-specific	Proteasome activity
C6.1A-like		Survival following DNA damage (?)

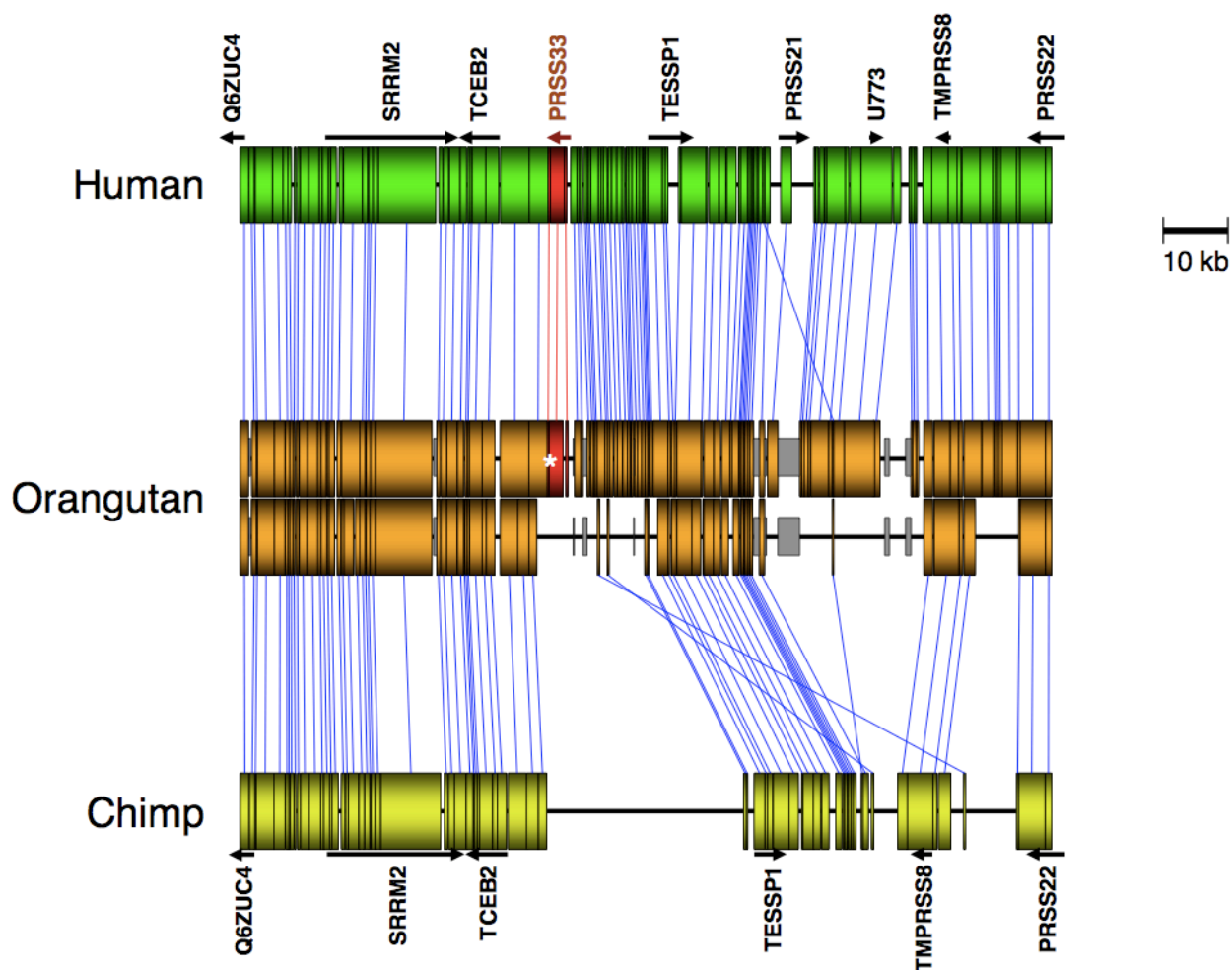


Figure S13-1. Independent inactivation of PRSS33 in the orangutan and chimpanzee. A region of the orangutan genome containing the PRSS33 pseudogene was compared to the orthologous regions of the human and chimpanzee genomes with the megablast algorithm. Each hit is represented by two boxes linked by a line. The hits containing the PRSS33 gene are in *red*. The premature stop codon in orangutan PRSS33 is marked with an *asterisk*. Regions of the orangutan genome with unknown sequence are displayed as *gray* boxes.

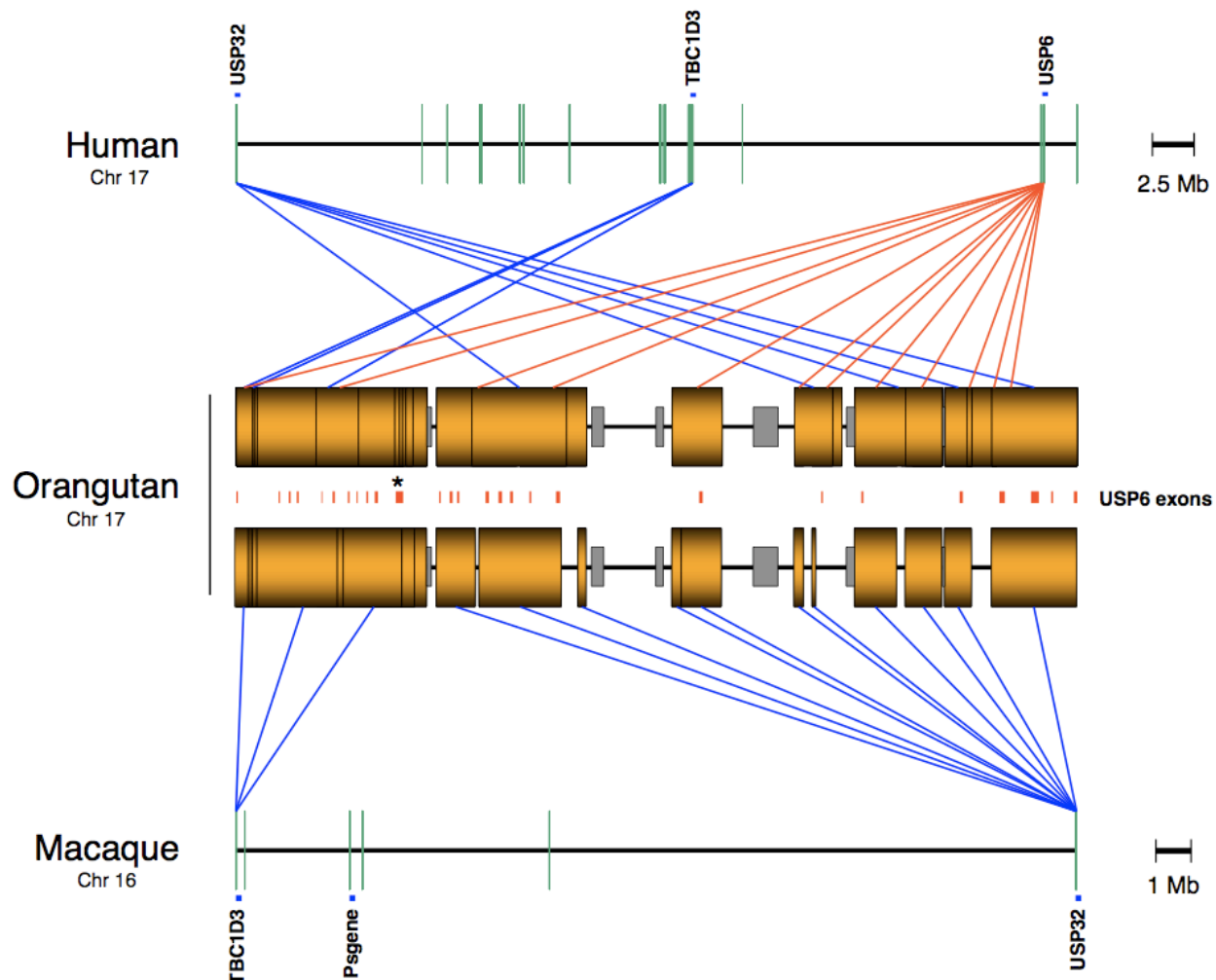


Figure S13-2. Inactivation of the hominoid-specific USP6 gene in orangutan. The orangutan USP6 pseudogene was compared to the human and macaque genomes with the megablast algorithm. Putative USP6 exons are represented by *red boxes*. Each hit is represented by two boxes linked by a line. The hits containing the USP6 gene are linked by *red lines*. The premature stop codon in the orangutan TBC-like domain of USP6 is marked with an *asterisk*. Regions of the orangutan genome with unknown sequence are displayed as *gray boxes*.

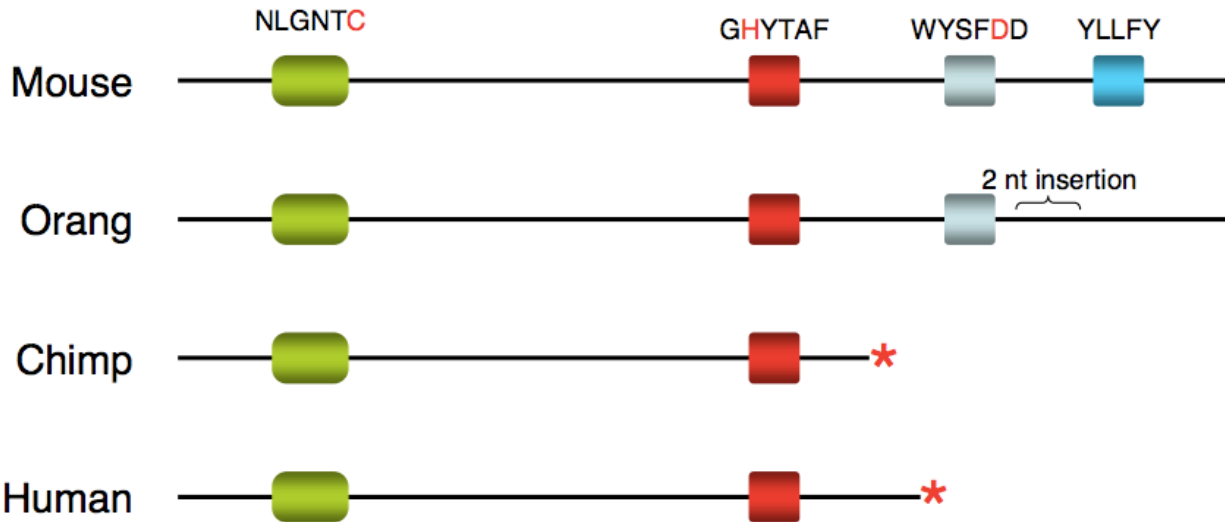


Figure S13-3. Catalytic domain of USP50 throughout evolution. Conserved sequences are represented by *boxes*. The consense sequences are shown on top, with catalytic triad residues in *red*. Premature stop codons are represented by *red asterisks*. A 2-bp insertion in orangutan USP50 is also indicated.

Supplemental Section S14 – Evolution of Orangutan Alpha and Theta Defensins

Defensins are short antimicrobial peptides that are expressed in epithelial cells and leukocytes, and are heavily involved in innate defense. The defensins are commonly divided into alpha, beta, and theta subfamilies, depending on the conformation of six conserved cysteine residues¹⁰⁶. While beta defensins are shared by all mammals, alpha defensins have likely evolved by a gene duplication from a beta defensin gene in a common ancestor of primates and rodents¹⁰⁷. The complex clusters of alpha defensin genes in rodents and primates arose independently in these lineages¹⁰⁸, probably due to their newly acquired ability to participate in antiviral defense^{109,110,111,112}.

To examine the recent evolutionary history of alpha defensins in primates, we analyzed BAC sequences covering the alpha defensin cluster in the orangutan (chromosome 8, AC206038.3, 236Kbp) and macaque (chromosome 8, AC204742.8 and AC202726.6, 312Kbp assembled) genomes, together with the finished human genome sequence. In the human genome, the alpha defensins are located in a subtelomeric region of chromosome 8 (8p23, 134Kbp). Figure S14-1 shows gene orders in all three reference sequences.

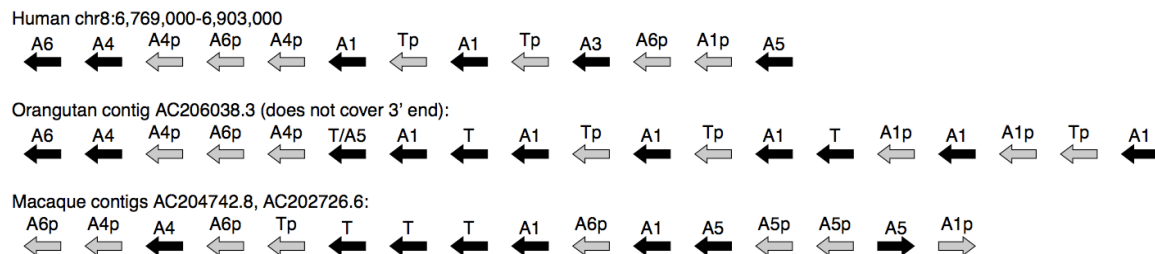


Figure S14-1. Gene orders in human, orangutan, and macaque alpha defensin gene clusters. Pseudogenes are shown in gray. The genes labeled A1/3, A4, A5, A6 are counterparts to the human genes DEFA1/3, DEFA4, DEFA5, and DEFA6 respectively. The genes labeled T are theta defensins.

We reconstructed the histories of these gene clusters using methods developed by Vinar et al. (2009)¹¹³. The inferred history for the alpha defensins shows remarkable differences among the three species (Figure S14-2). At the 5' end of the cluster, all three species show conserved order of the genes DEFA4 and DEFA6, however the orthologs of these genes in macaque are pseudogenized. At the 3' end, the human cluster contains a single copy of the DEFA5 gene, while the macaque genome has seen a recent expansion of the DEFA5 locus, resulting in two genes and three pseudogenes in the reference sequence. Since both DEFA5 and DEFA6 are highly expressed in intestinal Paneth cells¹¹⁴, this expansion may have compensated for the loss of function of DEFA6 gene in macaque. The human-macaque ancestor likely had two copies of DEFA4, one of them surviving on human-orangutan lineage (DEFA4), and one surviving on the macaque lineage (HA4_4p).

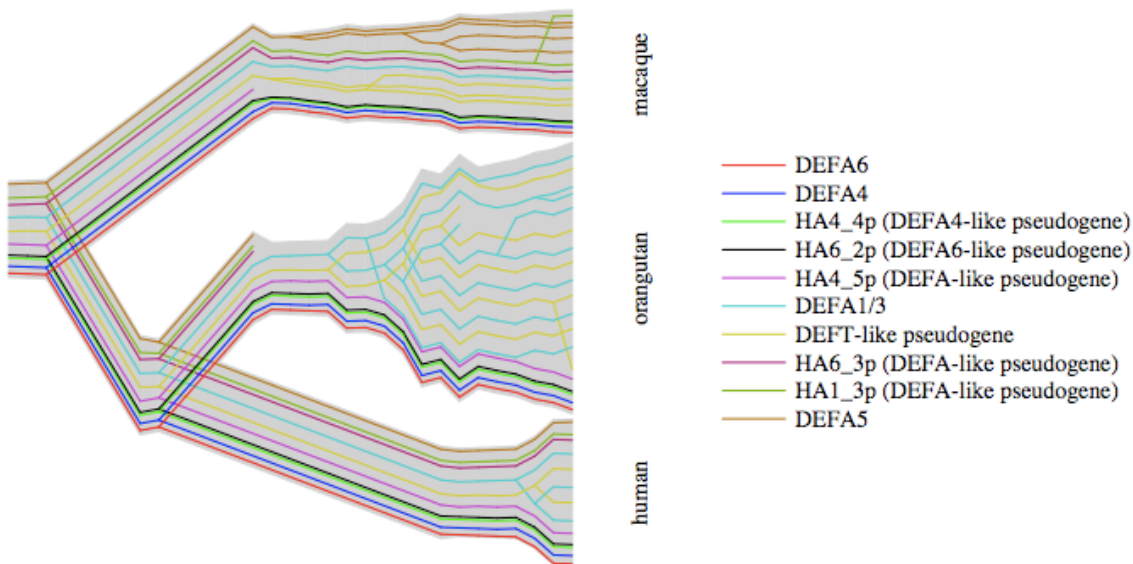


Figure S14-2. Evolutionary history of alpha and theta defensin cluster in human, orangutan, and macaque genome. The history of duplications was reconstructed by methods based on MCMC sampling by Vinar et al. (2009) and displayed in the form of a tube tree. The positions corresponding to the genes and pseudogenes in the human genome, and their orthologs in orangutan and macaque are shown in the context of corresponding gene trees. Individual genes or pseudogenes are shown in the order of their placement in extant and ancestral genomes. Note that the orangutan BAC does not cover region orthologous to DEFA6/DEFA5 genes in human.

Nearly identical DEFA1 and DEFA3 genes occupy the center of the gene cluster. The human reference sequence has three DEFA1/3 genes, while the BAC-assembled orangutan reference sequence shows eight DEFA1-like genes and pseudogenes. DEFA1-like genes may contribute to resistance to HIV-1 progression¹¹², which could explain these recent independent expansions. On the other hand, we do not observe a similar expansion in the macaque reference sequence.

Also located at the center of the alpha defensin cluster, the theta defensins (DEFT) evolved from alpha defensins by the emergence of a nonsense mutation that allowed the creation of a unique circular peptide. Although the theta defensins are transcribed in the human genome, all copies contain premature stop codon at amino acid 17 preventing subsequent functional translation¹¹⁵. However, the synthetic molecule retrocyclin-1, which is created by repairing this premature stop codon, protects human cell from HIV-1 infection in vitro¹¹⁶. The rhesus macaque contains three functional theta-defensins, which likely duplicated recently on the macaque lineage together with DEFA1/3 (Figure S14-2). The orangutan genome has also seen a large expansion of the theta defensins, with 6 copies in the reference sequence. However, not all copies are functional theta-defensins: one copy lacks the retrocyclin non-sense mutation and may function as an alpha-defensin, one copy is a pseudogene containing the same premature stop codon as the human theta defensins, two are pseudogenes, and two appear to be functional theta defensins.

The DEFA1/DEFA3/DEFT region in the human genome shows copy number variation ranging from 3 to 14 copies per diploid genome, with the presence of DEFA3 also being polymorphic¹¹⁷. To study copy number variation of alpha and theta defensins in the orangutan genome, we examined the next generation Illumina reads from 5 Bornean and 5 Sumatran individuals described in Section S4. The individual reads were mapped to the orangutan alpha defensin reference BAC (see above) by using SOAP2¹¹⁸ with default options. Due to apparently recent duplications within the reference contig, many reads were not uniquely mappable to the reference; these reads were reported only once, randomly at one of the best hit positions. The results (Figure S14-3) show copy number variation between individuals in the region containing DEFA1/3 and theta defensin genes. The orthologous region is also copy number variable in human genome¹¹⁷. Several Sumatran individuals show copy numbers similar to those of the reference genome (which is also Sumatran), while all Bornean individuals exhibit smaller numbers of theta defensin copies.

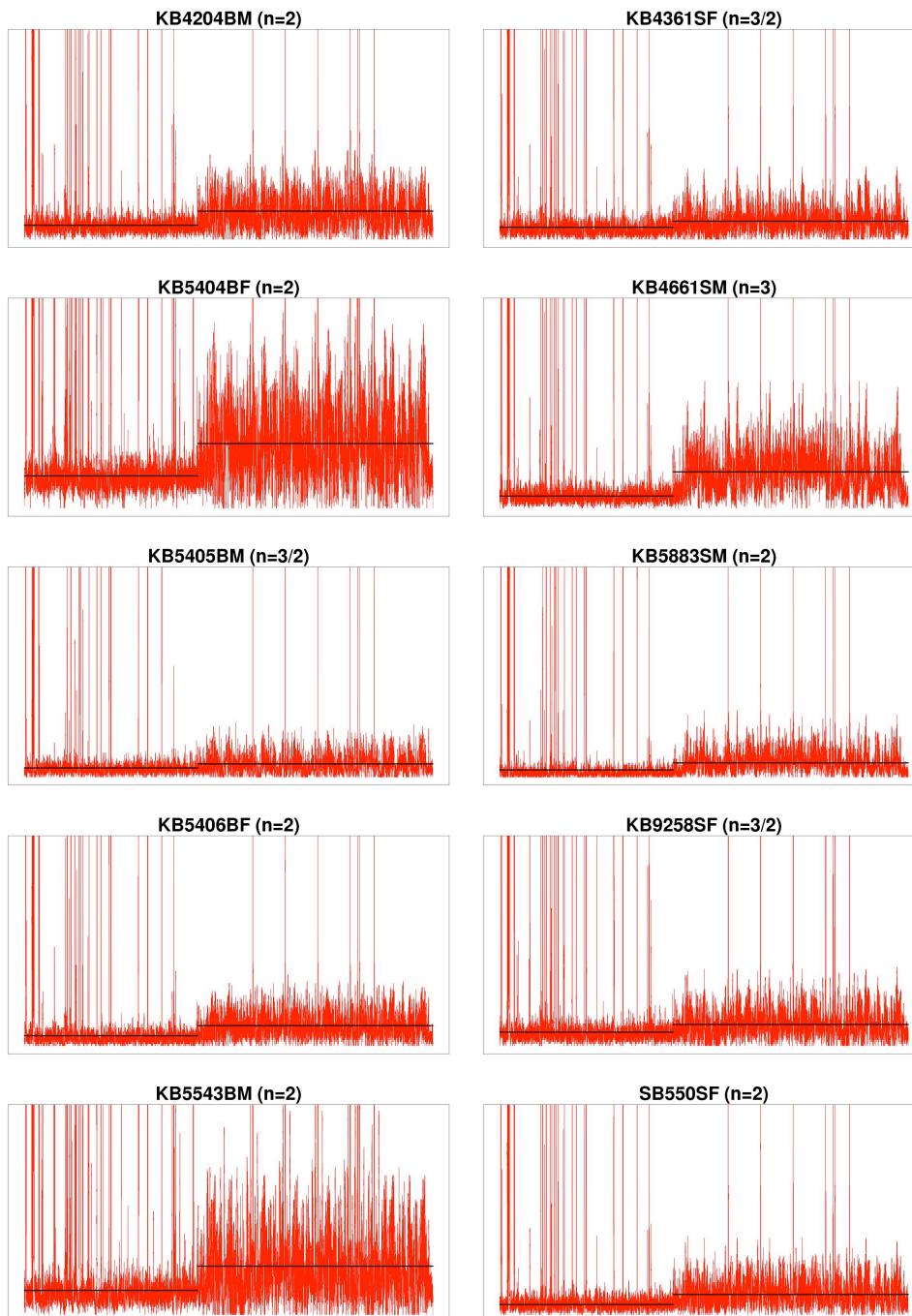


Figure S14-3. Alpha and theta defensin read-depth variation. The reads from 10 individuals (5 Bornean on the left, and 5 Sumatran on the right) were mapped on the reference contig of alpha and theta defensin cluster. Compared to the reference contig, all Bornean and 3 Sumatran individuals show lower copy numbers in the theta defensin region (3' half of the reference contig).

The enhanced inventory and copy number variation of theta defensins in the orangutan genome suggests a significantly more complex evolutionary history than that implied by the human and macaque sequences alone and presents an opportunity for further study of evolution and function of theta defensins.

Supplemental Section S15 – Genic Positive Selection

Identification of Orthologs

Using methods similar to those described by Kosiol et al. (2008), we identified 13,872 high confidence 1:1 orthologous genes in the human (hg18) and orangutan (ponAbe2) genomes, with additional genes included from the chimpanzee (panTro2), macaque (rheMac2), and dog (canFam2) genomes where possible¹¹⁹.

Briefly, we started with a permissive set of human genes consisting of the union of the RefSeq, knownGene, and VEGA collections (downloaded on May 26, 2008 from the UCSC genome browser). The overlapping transcripts were grouped into 20,435 clusters and mapped to the other genomes using the UCSC multiz primate alignments (10 species). These transcripts were then subjected to a series of rigorous tests: synteny (at least 80% of the coding sequence must be remapped via a single alignment chain), completeness ($\leq 10\%$ of coding sequence in sequencing gaps), no frameshifts (unless compensated within 15 bp), conserved gene structure (start sites, stop sites, and splice sites must be conserved), and no-recent-duplication (no duplication since the split from the human lineage). Due to frequent changes in start and stop site positions, the incomplete transcripts missing 10% of their length from both ends were also considered for remapping. Finally, from each of the clusters, we selected the transcript mapping to the highest number of species, with the length of the transcript as a secondary criterion.

Two additional measures were taken to improve the results of the above procedure. First, bases in the alignments with sequence quality score < 20 in the chimpanzee, orangutan, and macaque genomes were masked (changed to “N”s). Second, additional filtering was applied to syntenic nets in the orangutan genome, by removing all chains of size $< 31,700$ bp as such short chains were observed often to lead to spurious alignments. (This cutoff was determined empirically, based on a set of positive and negative examples.)

Of the 13,872 orthologous sets, 89% include chimpanzee, 83% include macaque, and 76% include dog genes, in addition to the required human and orangutan genes, while 60% include genes from all five species. A comparison of numbers of genes left after individual tests in the pipeline revealed that an unexpectedly large number of candidate genes were rejected due to apparent frameshift indels in orangutan (Figure S15-1). Further inspection indicated that a high percentage of these genes corresponded to regions of the orangutan assembly with low read coverage and low quality sequence. We expect that further sequencing and improved assembly quality will allow larger numbers of orthologous genes to be identified with high confidence.

We have also compared our orthologs to ones generated by ENSEMBL (Figure S15-2). While the ENSEMBL data contains more orthologs than ours (17491), 23% of their orthologous sets fail the quality filters in our pipeline, with frameshift indels as the single largest contributing factor. For the remaining ENSEMBL-only orthologs, there were no overlapping genes in the RefSeq, knownGenes, or VEGA catalogs. On the other hand,

our pipeline identified more than 1400 high-confidence orthologs that have no counterpart in ENSEMBL. We conclude from this comparison that our ortholog set, while perhaps overly conservative for some purposes, is more appropriate than the ENSEMBL set for an analysis of positive selection, which can be highly sensitive to annotation errors or differences between species in gene structure. Notably, for approximately 200 genes, the two sets differ in their orthology mapping, by identifying different target coordinates in orangutan genome.

Likelihood Ratio Tests For Positive Selection

Our likelihood ratio tests (LRTs) for positive selection are based on the widely used site or branch-site models of codon evolution developed by Nielsen, Yang, and colleagues^{120,121,122}. The LRT for selection on any branch of the phylogeny is essentially Nielsen and Yang's (1998) test of site models 2a versus 1a, and the lineage- and clade-specific LRTs are essentially instances of Yang and Nielsen's (2002) test 2. However, to reduce the number of parameters estimated per gene, the complete set of 13,872 genes was divided into eight equally sized classes by G+C content in third codon positions. The branch lengths and the transition-transversion rate ratio κ were estimated for each class under the null model, and these estimates were subsequently held fixed, in a G+C dependent way, for the LRTs. Instead of a complete set of branch lengths, a single scale parameter μ was estimated per gene. Thus, only the parameters μ , $\omega_0 < 1$ and p_0 for the null model, and the additional parameters $\omega_2 > 1$ and p_1 for the alternative model, were estimated per gene (see Nielsen and Yang, 1998 and Yang and Nielsen, 2002). For the LRT for selection on any branch, *P*-values were computed empirically, based on simulation experiments (see Kosiol et al., 2008 for details). For the lineage- and clade-specific LRTs, *P*-values were computed assuming the null distribution was a 50:50 mixture of a distribution and a point mass at zero. The method of Benjamini and Hochberg (1995) was used to estimate the appropriate *P*-value threshold for a false discovery rate (FDR) of < 0.05 ¹²³.

A website is available at <http://compgen.bscb.cornell.edu/~kosiol/orang-psg/> with definitions of the candidate genes (accession numbers, genomic coordinates, and descriptions), multiple alignments of orthologous gene sets, and detailed results of the LRTs.

Gene classification analysis

Each gene was assigned categories from the GO¹²⁴ and PANTHER¹²⁵ databases, based on the Uniprot identifiers of associated transcripts. To account for the hierarchical nature of these databases, each gene was also considered to belong to all parent categories of the ones to which it was directly assigned. The distributions of LRT *P*-values among the genes assigned to each category *C* and not assigned to *C* were compared by a (one-sided) Mann-Whitney *U* (MWU) test. Nominal *P*-values computed by the MWU tests were corrected for multiple comparisons using the method of Holm (1979)¹²⁶.

In addition to enrichments for genes related immunity and defense we found enrichments for two new GO categories: “visual perception” and “glycolipid metabolic processes”. Below we list PSGs (p -value <0.05) in these categories and relevant publications. For the genes involved in metabolic genes we also show their location in the sphingolipid metabolic pathway. Note that Figure 15-3 below shows the full pathway as given in the KEGG database, not only the parts shown in the main paper.

Vision

Three major visual signaling proteins, transducin (GNAT1,2), arrestin (ARR3, X-arrestin) and recoverin (RCVRN) undergo translocations between the outer segment and the inner compartments of rod photoreceptors in a light dependent manner (Artemyev, 2008). Two of these proteins, arrestin ($P=0.00665$) and recoverin ($P=0.00842$) show signatures of positive selection and appear to play roles in maintaining light sensitivity¹²⁷. Additionally, we find OPN1SW1 ($P=0.01996$), related to blue color vision, to be under positive selection. Primates rely heavily on vision to evaluate the world around them. However, there is much diversity among and within primate species in both visual acuity and color vision abilities. Monkey and ape retinas lack a tapetum lucidum, which enhances night vision, but have evolved a specialized region of tightly packed light-sensitive cells, the retinal fovea, allowing for increased visual acuity¹²⁸. Trichromatic color vision appears to have evolved after Old World and New World monkeys split, as all Old World monkeys and apes are trichromatic and only some New World monkeys are¹²⁹. This variation has been linked to foraging ecology, with the hypothesized selective pressure being the ability to distinguish ripe fruit from background foliage, particularly in times of food shortages¹³⁰.

Metabolic Disease

GAL3ST1 ($P=0.00448$) catalyzes the synthesis of galactosylceramide sulfate, a major lipid component of the myelin sheath, which is a protective coating around a nerve that acts as an “insulator” and aids in proper conduction of the nerve impulse¹³¹. HEXB ($P=0.00448$) catalyzes the degradation of the ganglioside GM2 and is associated with Sandhoff’s disease^{132,133}. B4GALNT1 ($P=0.0163$) is the enzyme involved in the biosynthesis of GM2. NEU3 ($P=0.0351$)¹³⁴ may play a role in modulating the ganglioside content of the lipid bilayer¹³⁵. PSAP ($P=0.00695$) is associated with Gaucher disease, Tay-Sachs disease, and metachromatic leukodystrophy. Patients with these disorders either do not produce enough of one of the enzymes needed to metabolize lipids or they produce enzymes that do not work properly. Over time, this excessive storage of fats can cause permanent cellular and tissue damage, particularly in the brain, but also in the liver, spleen, and bone marrow.

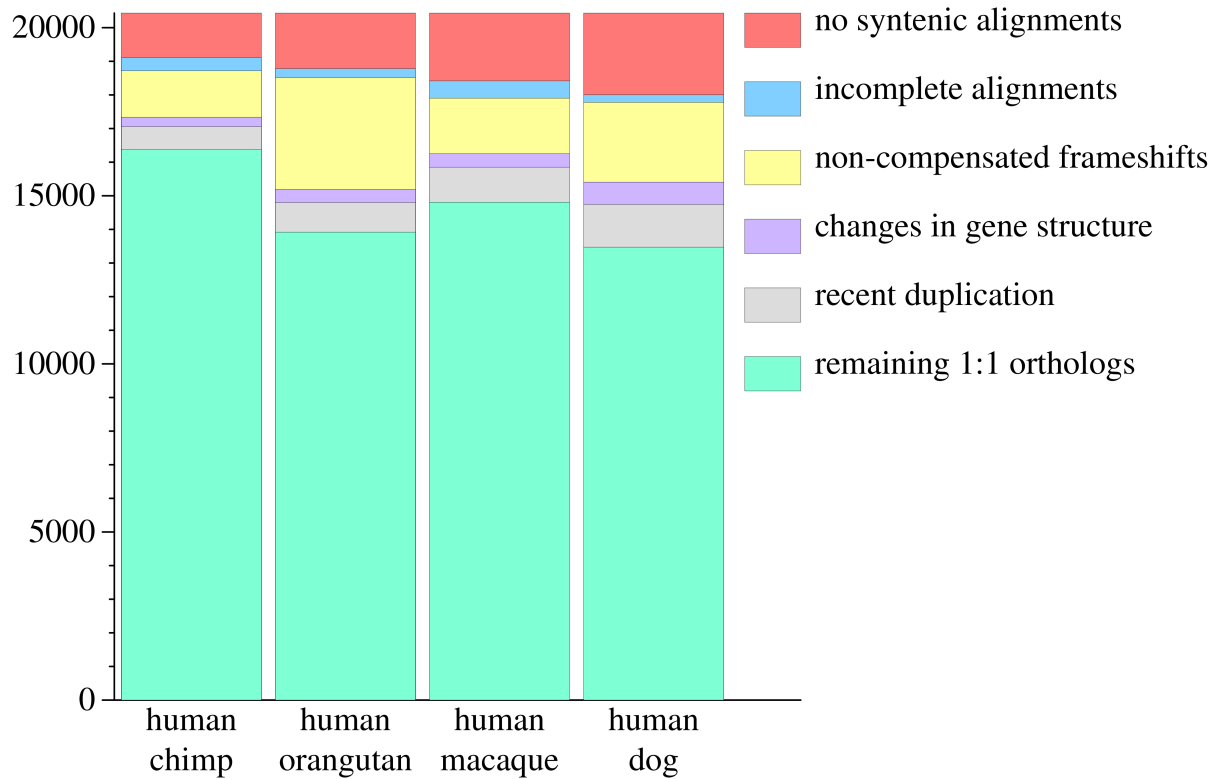
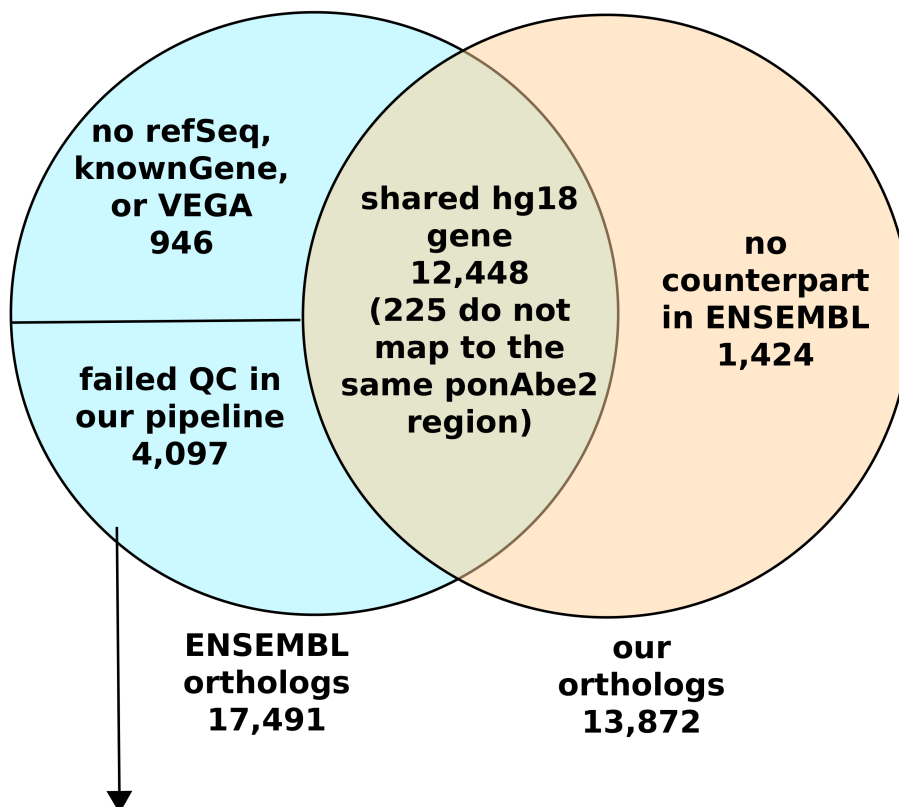


Figure S15-1. Numbers of human genes passing successive filters in the orthology analysis pipeline.



The most common quality control failures:

- 57% noncompensated frameshifts
- 19% synteny breakpoint within gene
- 11% recent duplications in orang or human lineages (unable to assign 1:1 orthology)
- 6% significant sequencing gaps in orang
- 4% likely gene structure changes

Figure S15-2. Comparison of our orthologs with Ensembl orthologs. Note the Venn diagram is not drawn to scale.

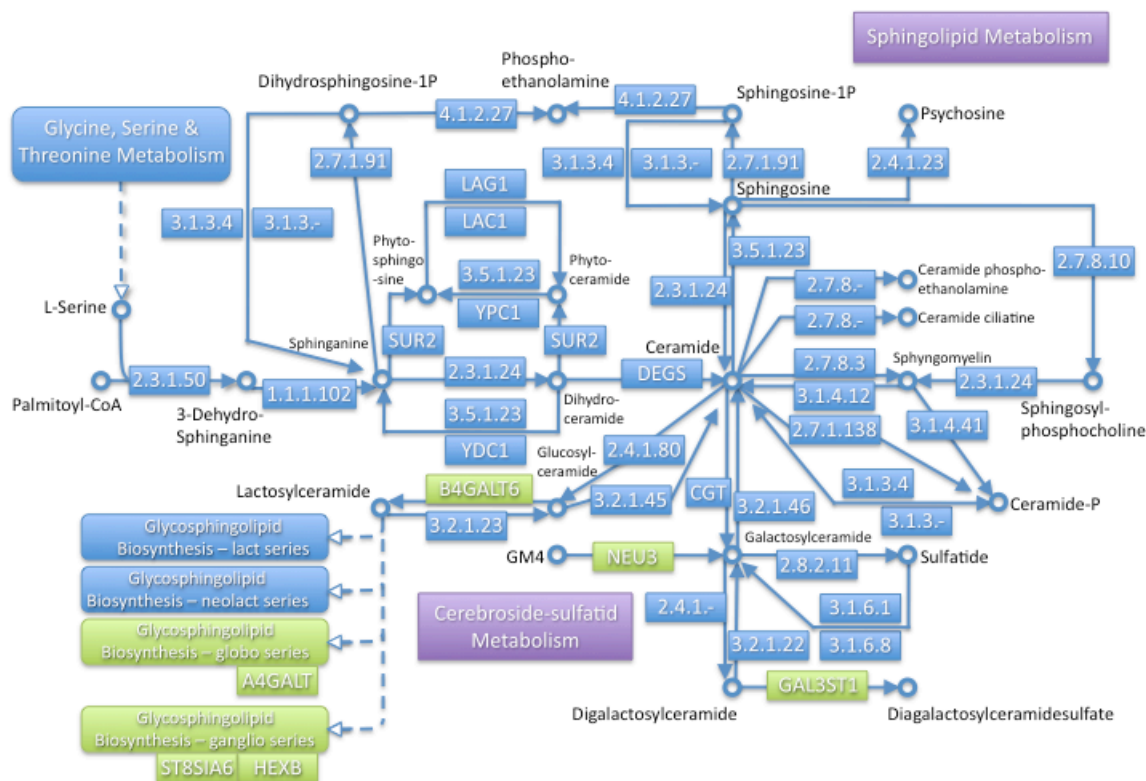


Figure S15-3. Full primate sphingolipid metabolic pathway as presented in the KEGG database. Putative PSGs (p -value < 0.05) are indicated by green boxes. All arrows, connectors and nodes are indicated graphically in accordance with KEGG guidelines.

Supplemental Section S16 – Bornean/Sumatran Divergence

We aligned next generation whole genome sequence reads (Supplemental Section S4) generated from the high-depth Bornean female individual (KB5404) to the Sumatran orangutan genome assembly (v2.0.2 a.k.a. ponAbe2) using mrFAST¹³⁶ to calculate the average divergence/identity (Table S16-1). We used 5 kb windows of non-RepeatMasked, gap-free and duplication-free sequence (PhredQ > 20 ; $n=223,192$ windows). We only considered variants supported by at least 2 reads. The distribution is depicted in Figure S16-1.

Table S16-1. Estimate of Bornean vs Sumatran single nucleotide diversity.

AVERAGE of % identity	MEDIAN of % identity	STD DEV
0.9968	0.9989	0.00968

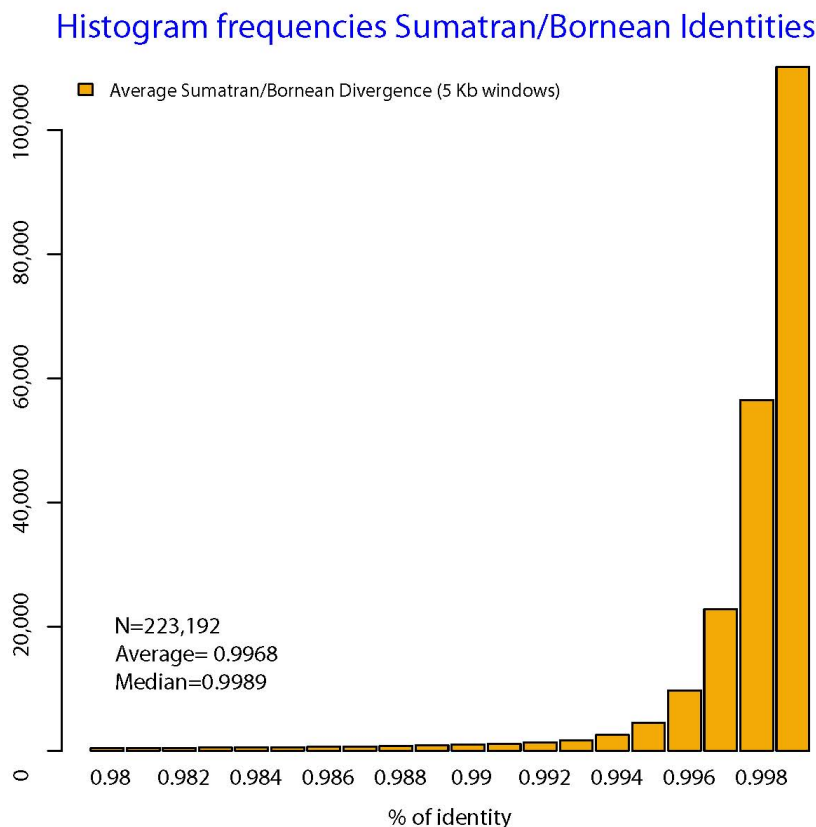


Figure S16-1. Frequency histogram of identities comparing Bornean orangutan next generation sequence reads mapped to the Sumatran reference genome.

Supplemental Section S17 – HMM Estimate of Bornean/Sumatran Divergence Time, Speciation Time and Effective Population Size

Data preprocessing

We obtained an alignment of the Sumatran and Bornean orangutan individuals by mapping a mix of 36- and 50-bp single-end and paired-end Illumina short reads from the Bornean individual to the reference assembly, constructed from capillary reads of the Sumatran individual (Supplemental Section S4). Blocks less than 100 nucleotides apart were pooled and all gaps replaced by 'N', resulting in a series of “chunks”. Chunks with less than 300 sites were removed and the remaining chunks were concatenated into segments of ~1 Mb. The result of the preprocessing steps is 2,689 segments that were each analyzed independently.

Model

We developed a hidden Markov model (available in the software at <http://gna.org/projects/coalhmm/>) that uses changes in coalescence time along the genome to estimate the speciation time, recombination rate and ancestral effective population size from pairs of sequences. Details of the method is described in the companion paper Mailund *et al.* **Estimating Speciation Time and Ancestral Effective Population Size of Bornean and Sumatran Orangutan Subspecies using a Coalescent Hidden Markov Model**. We validated the model using extensive coalescent simulations and generally found a good recovery of parameters.

Parameter estimates

We estimated speciation time, average recombination rate and ancestral effective population size in each ~1 Mb segment independently and discarded the results where 1) the speciation time was below 5 thousand years or above 1 million years, 2) the ancestral effective population size was below 5,000 or above 100,000, and 3) the recombination rate was below 0.1 cM/Mb or above 10 cM/Mb. In total 203 segments were discarded from the analysis (leaving 2,486 segments).

Assuming a per-basepair per-generation substitution rate of $2e-8$ we estimated the speciation time to be 334 ± 145 kya and the recombination rate to be 0.95 ± 0.72 cM/Mb. We estimated the ancestral effective population size on autosomes and the X chromosome independently. For the autosomes we found the effective population size to be $26,800 \pm 6,700$ and for the X chromosome $20,400 \pm 7,400$, consistent with the theoretical $\frac{3}{4}$ effective population size of X chromosomes.

Supplemental Section S18 – Bornean/Sumatran Duplication Comparison

The duplication map of both orangutan subspecies reveals that most of the duplications are shared (97% of SDs >20kb), with ~30% of the shared duplications showing copy number differences between species. The rate of both Sumatran and Bornean specific duplications is within the lower bound of previously published estimates (ranging from 1-3 Mb/Myr) (Marques-Bonet *et al.* 2009). For the duplication map of the Sumatran orangutan genome we used the WSSD-based methodology described above, based on the WGS read data. For the Bornean duplication map, we mapped approximately 20x coverage of next generation sequence data from a Bornean individual (KB5404) using mrFAST¹³⁶ (Supplemental Section S4). To reduce false positives from short read mapping, we excluded artifacts corresponding to smaller duplications that are not detected by the capillary sequence based approach. We also restricted our analysis to the autosomes. By these two methods we detected 53.7 Mb of duplication in the Sumatran individual and 61.4 Mb of duplication in the Bornean individual. We classified

segmental duplications into three categories: shared Bornean/Sumatran duplications, Sumatran specific duplications and Bornean specific duplications.

Array comparative genomic hybridization was used to confirm individual-specific duplications and to confirm copy-number differences for shared duplications. The Sumatran (“Susie”) and the Bornean reference genomes (“KB5404”) were hybridized in replicate with dye-flips between test and reference. Log₂ relative hybridization intensity was calculated for each probe. In this analysis, we restricted our analysis to those regions that were greater than 10 kb in length and contained at least 20 probes with a consistent log₂ in both experiments. We used a heuristic approach to calculate log₂ thresholds of significance for each comparison dynamically adjusting the thresholds for each hybridization to result in a false discovery rate of <1% in the control regions (Marques-Bonet et al. 2009). We validated a total of 1.2 Mb (> 10 Kb) and 630 Kb (> 20 Kb) of Sumatran-specific SDs and 1.1 Mb (> 10 Kb) and 624 Kb (> 20 Kb) of Bornean-specific SDs (Table S18-1). These copy-number polymorphic SD encompass in part or completely a total of 47 genes. However, the validation of specific SDs is very low (~10%), in large part, because of the excess of predicted Bornean-specific duplications using Illumina next-gen WGS sequence data. We suspect that differences in the platforms (Illumina versus Sanger sequences) as well as randomness of the libraries contribute to a higher rate of false negatives and positives for the short read WGS.

Table S18-1. Summary of array CGH validated sites of duplication comparative map between Sumatran and Bornean Orangutans. Numbers in Italics correspond to copy number correction.

	Fragments > 10 kb	Fragments > 20 kb
Sumatran SDs (not in Bornean)	1,197 kb (<i>2,057 kb</i>)	630 kb (<i>1,019 kb</i>)
Sumatran specific SDs	260 kb (<i>567 kb</i>)	96 kb (<i>195 kb</i>)
Bornean SDs (not in Sumatran)	1,122 kb (<i>3,215 kb</i>)	624 kb (<i>2,006 kb</i>)
Bornean specific SDs	367 kb (<i>1,657kb</i>)	176 kb (<i>859 kb</i>)
Shared SDs with more copies in Sumatran	6,161 kb	5,341 kb
Shared SDs with more copies in Bornean	3,173 kb	2,508 kb
Shared SDs with similar copy number	25,087 kb	19,798 kb

After removing all the intervals overlapping with previously known human, chimpanzee or macaque SDs, we found 260 Kb (567 Kb after copy number correction) and 366 Kb (1,657 Kb) of Sumatran and Bornean exclusive SDs respectively (> 10 Kb). Similarly, 9.2 Mb (> 10 kb) and 7.8 Mb (> 20 Kb) that were found copy number variant in shared duplications (6.1 Mb (> 10 Kb) and 5.3 Mb (> 20 Kb) of Sumatran SDs and 3.1 Mb (> 10 Kb) and 2.5 Mb (> 20 Kb) of Bornean SDs). 55 genes are included in those regions (20 had more copies in Bornean and the remaining 35 had more copies in Sumatran). In summary, the genomic architecture of the two subspecies is highly similar, yet contains several megabases of potential lineage-specific duplication (only 3% of the duplications are not shared). If one assumes 1.2-2 Myr as the separation of both species, the rate of Mb/Myr (for fragments > 20 kb) is in the lower bound of what was previously published (previously rates ranging from 1-3 Mb/Myr)⁵⁷. Figures S18-1 to S18-4 show examples of shared Bornean/Sumatran duplications, as well as Bornean-specific and Sumatran-specific duplications, all containing genic content.

Figure S18-1. An example genic duplication shared among Bornean and Sumatran orangutans.

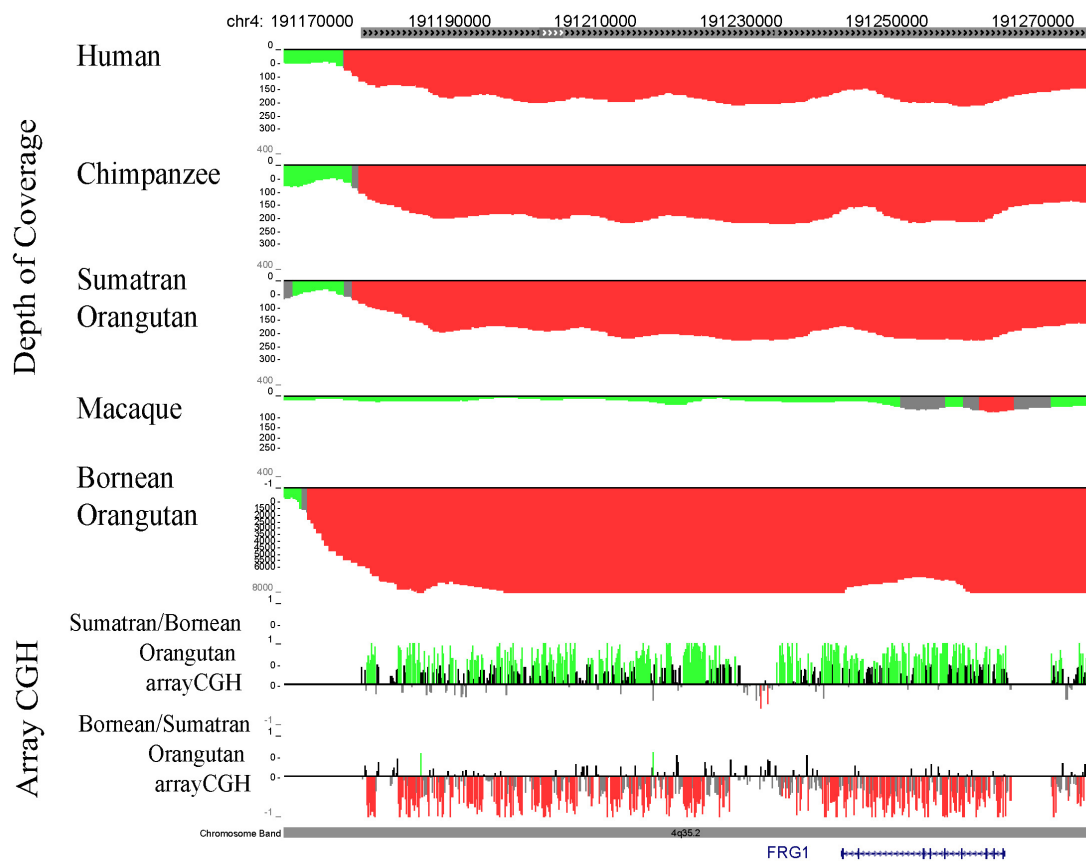


Figure S18-2. An example genic Sumatran-specific duplication.

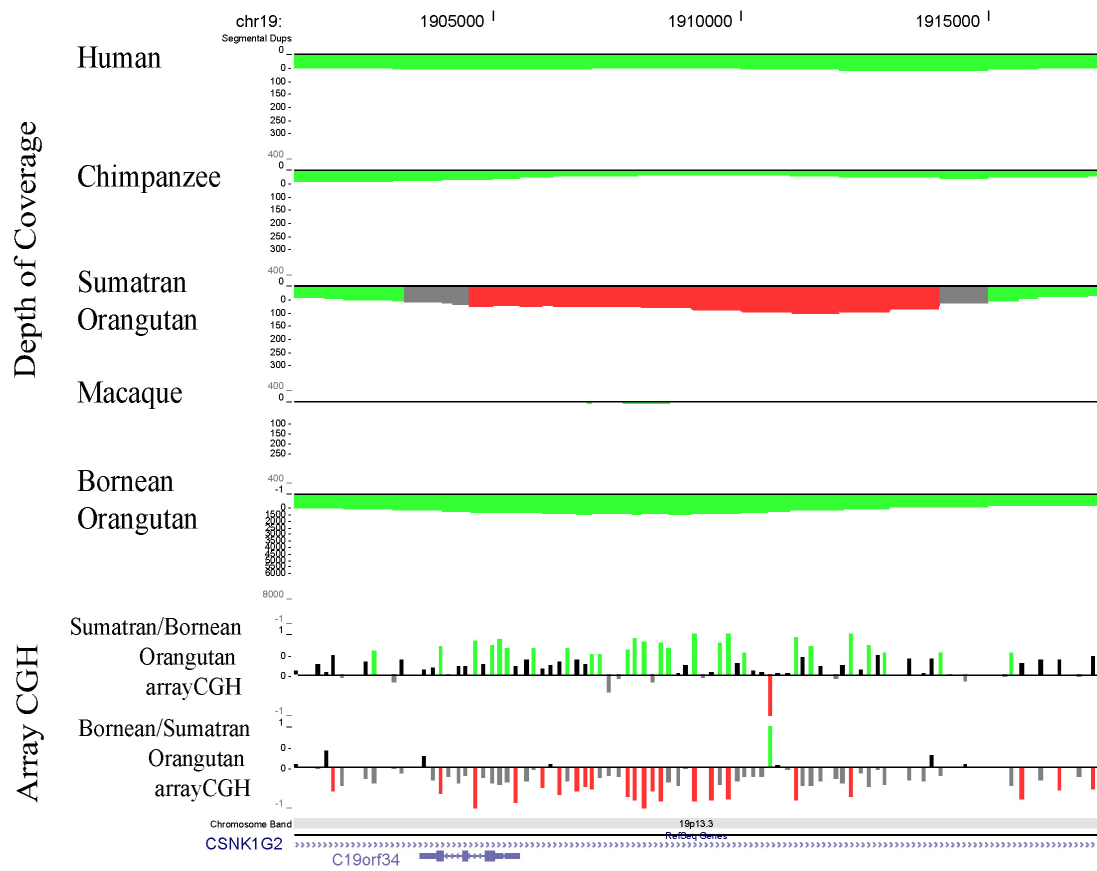


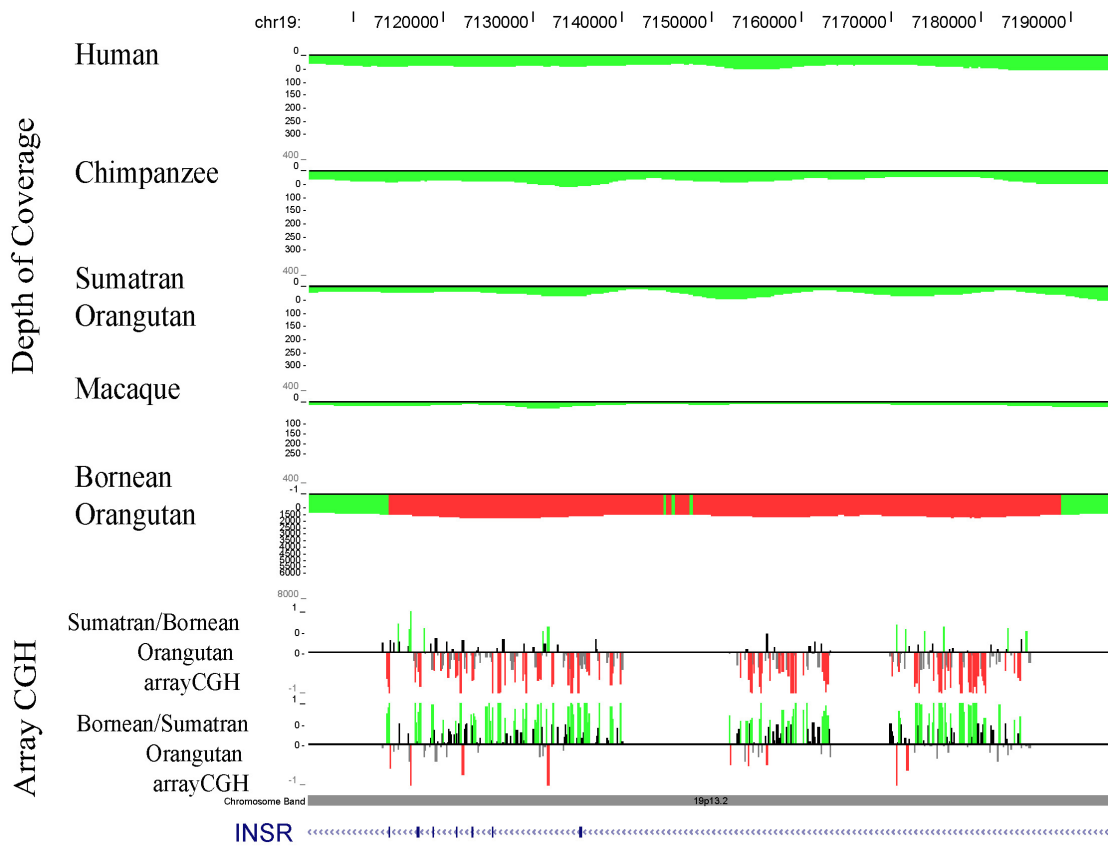
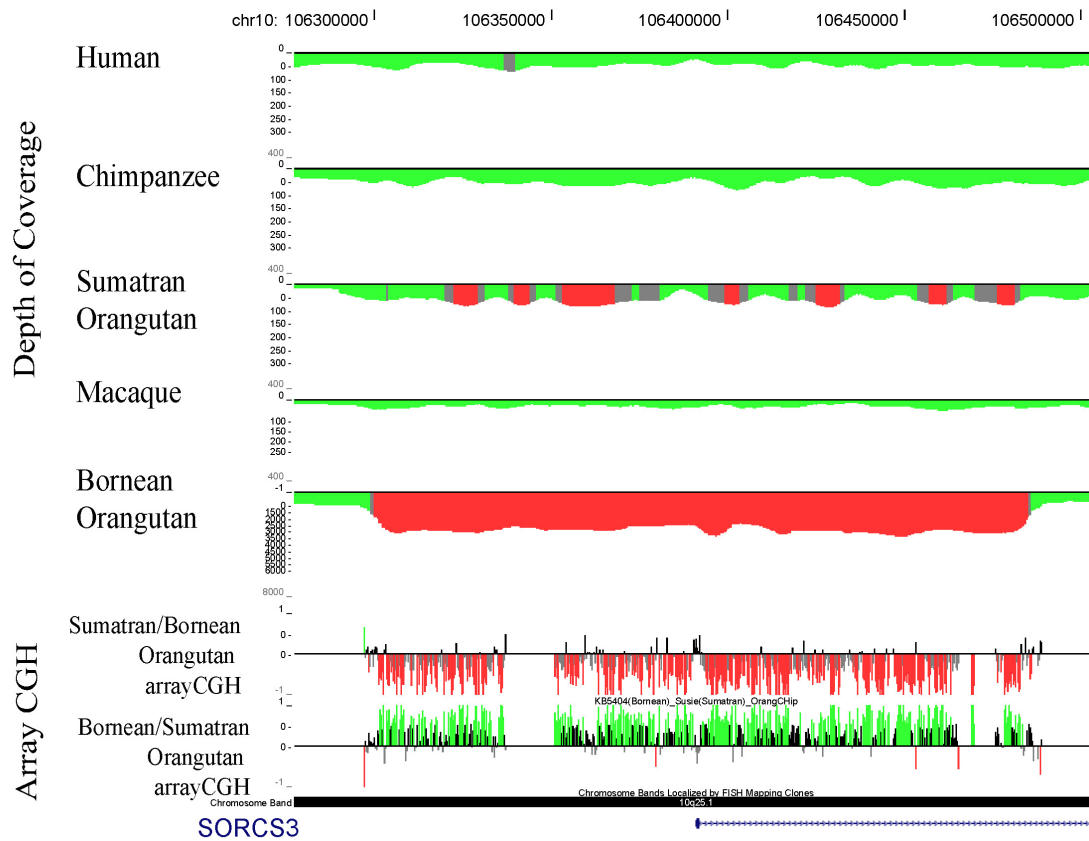
Figure S18-3. An example genic Bornean-specific duplication.

Figure S18-4. An example of a genic Bornean-specific duplication expansion.

Supplemental Section S19 – Retroelement Polymorphisms

Structure Analysis

We investigated the population structure within 37 orangutan individuals (Table S19-1) using the Structure software v.2.1^{137,138}. Simulations were performed under the admixture model. The admixture model assumes that individuals may have mixed ancestry, meaning that some retrotransposon insertions are inherited from ancestors in population *k*. The structure analyses were run on a desktop machine with 4 CPUs.

Using genotype data from unlinked markers this software performs a model-based clustering method to infer the population structure. Because the number of population clusters is unknown for a given dataset, initially *K* – where *K* equals the number of population clusters – was set from 1 to 6 to allow the software to determine the value of *K* with the highest likelihood. The initial burn-in period was set at 1,000,000 iterations and followed by a run-length of 1,000,000 steps and repeated at least five times. After determination of the value of *K* (here 3) 25 replications were run under identical burn-in and run-length settings. The information regarding the geographic origin was omitted.

For the structure analysis we selected 85 autosomal polymorphic retrotransposon insertions (PCR reactions, conditions, and primer sequences see Supplemental Section S9). Due to the relative quiescence of *Alu* retrotransposition in the orangutan lineage we also included polymorphic L1 and SVA insertions (15 *Alu*, 39 L1 and 31 SVA). To our knowledge this represents the first population genetic study that makes use of polymorphic SVA insertions. The majority of retrotransposon insertions were selected from PonAbe2. To reduce the common ascertainment bias we also included retrotransposon insertions identified from short sequence reads (Illumina) of a Bornean orangutan (KB 5404) through comparison against the *P. pygmaeus abelii* draft genome sequence.

Our structure analysis revealed clear evidence for population structure within the Sumatran orangutans apart from the clear distinction of Bornean and Sumatran orangutans. We found evidence for the existence population substructure within Sumatran orangutans. We identified two population clusters within the Sumatran orangutans. The different Sumatran orangutans show varying degrees of admixture with some individuals being clearly distinct from the Sumatran draft genome sequence (Figure S19-1).

Table S19-1.

	Species Names	Common Names	Origin	ID number
1	<i>Pongo pygmaeus abelii</i>	Sumatran Orangutan	Coriell ^a	PR01109 ("Susie")
2	<i>Pongo pygmaeus abelii</i>	Sumatran Orangutan	SDFZ ^b	KB4361
3	<i>Pongo pygmaeus abelii</i>	Sumatran Orangutan	SDFZ ^b	KB4503
4	<i>Pongo pygmaeus abelii</i>	Sumatran Orangutan	SDFZ ^b	KB4661
5	<i>Pongo pygmaeus abelii</i>	Sumatran Orangutan	SDFZ ^b	KB5370
6	<i>Pongo pygmaeus abelii</i>	Sumatran Orangutan	SDFZ ^b	KB5883
7	<i>Pongo pygmaeus abelii</i>	Sumatran Orangutan	SDFZ ^b	KB9258 (OR1447)
8	<i>Pongo pygmaeus abelii</i>	Sumatran Orangutan	SDFZ ^b	SB550
9	<i>Pongo pygmaeus abelii</i>	Sumatran orangutan	Coriell ^a	GM06213A
10	<i>Pongo pygmaeus abelii</i>	Sumatran orangutan	Coriell ^a	GM04272A
11	<i>Pongo pygmaeus abelii</i>	Sumatran orangutan	Coriell ^a	NG06209
12	<i>Pongo pygmaeus abelii</i>	Sumatran orangutan	Coriell ^a	NG12256
13	<i>Pongo pygmaeus abelii</i>	Sumatran orangutan	SDFZ ^b	OR823
14	<i>Pongo pygmaeus abelii</i>	Sumatran orangutan	NIH/NCI ^c	PPY16
15	<i>Pongo pygmaeus abelii</i>	Sumatran orangutan	NIH/NCI ^c	PPY17
16	<i>Pongo pygmaeus abelii</i>	Sumatran orangutan	NIH/NCI ^c	PPY29
17	<i>Pongo pygmaeus abelii</i>	Sumatran orangutan	NIH/NCI ^c	PPY30
18	<i>Pongo pygmaeus abelii</i>	Sumatran orangutan	NIH/NCI ^c	PPY36
19	<i>Pongo pygmaeus abelii</i>	Sumatran orangutan	NIH/NCI ^c	PPY44
20	<i>Pongo pygmaeus pygmaeus</i>	Bornean Orangutan	SDFZ ^b	KB4204
21	<i>Pongo pygmaeus pygmaeus</i>	Bornean Orangutan	SDFZ ^b	KB5404
22	<i>Pongo pygmaeus pygmaeus</i>	Bornean Orangutan	SDFZ ^b	KB5405
23	<i>Pongo pygmaeus pygmaeus</i>	Bornean Orangutan	SDFZ ^b	KB5406
24	<i>Pongo pygmaeus pygmaeus</i>	Bornean Orangutan	SDFZ ^b	KB5418
25	<i>Pongo pygmaeus pygmaeus</i>	Bornean Orangutan	SDFZ ^b	KB5419
26	<i>Pongo pygmaeus pygmaeus</i>	Bornean Orangutan	SDFZ ^b	KB5482
27	<i>Pongo pygmaeus pygmaeus</i>	Bornean Orangutan	SDFZ ^b	KB5543
28	<i>Pongo pygmaeus pygmaeus</i>	Bornean Orangutan	SDFZ ^b	KB6109
29	<i>Pongo pygmaeus pygmaeus</i>	Bornean Orangutan	SDFZ ^b	SB664
30	<i>Pongo pygmaeus pygmaeus</i>	Bornean orangutan	SDFZ ^b	OR315
31	<i>Pongo pygmaeus pygmaeus</i>	Bornean Orangutan	Coriell ^a	AG05252A
32	<i>Pongo pygmaeus pygmaeus</i>	Bornean orangutan	NIH/NCI ^c	PPY23
33	<i>Pongo pygmaeus pygmaeus</i>	Bornean orangutan	NIH/NCI ^c	PPY28
34	<i>Pongo pygmaeus pygmaeus</i>	Bornean orangutan	NIH/NCI ^c	PPY35
35	<i>Pongo pygmaeus pygmaeus</i>	Bornean orangutan	NIH/NCI ^c	PPY104
36	<i>Pongo pygmaeus pygmaeus</i>	Bornean orangutan	NIH/NCI ^c	PPY105
37	<i>Pongo pygmaeus pygmaeus</i>	Bornean orangutan	NIH/NCI ^c	PPY106

^a Coriell Institute for Medical Research, 403 Haddon Avenue, Camden NJ 08103, USA

^b San Diego Frozen Zoo, Conservation and Research for Endangered Species (CRES)

^c NIH/National Cancer Institute, Laboratory of Genomic Diversity, Frederick, MD 21702

^d DNA PR01109 in conjunction with the Orangutan Genome Sequencing Project, The Genome Center at Washington University, St. Louis, MO 63108

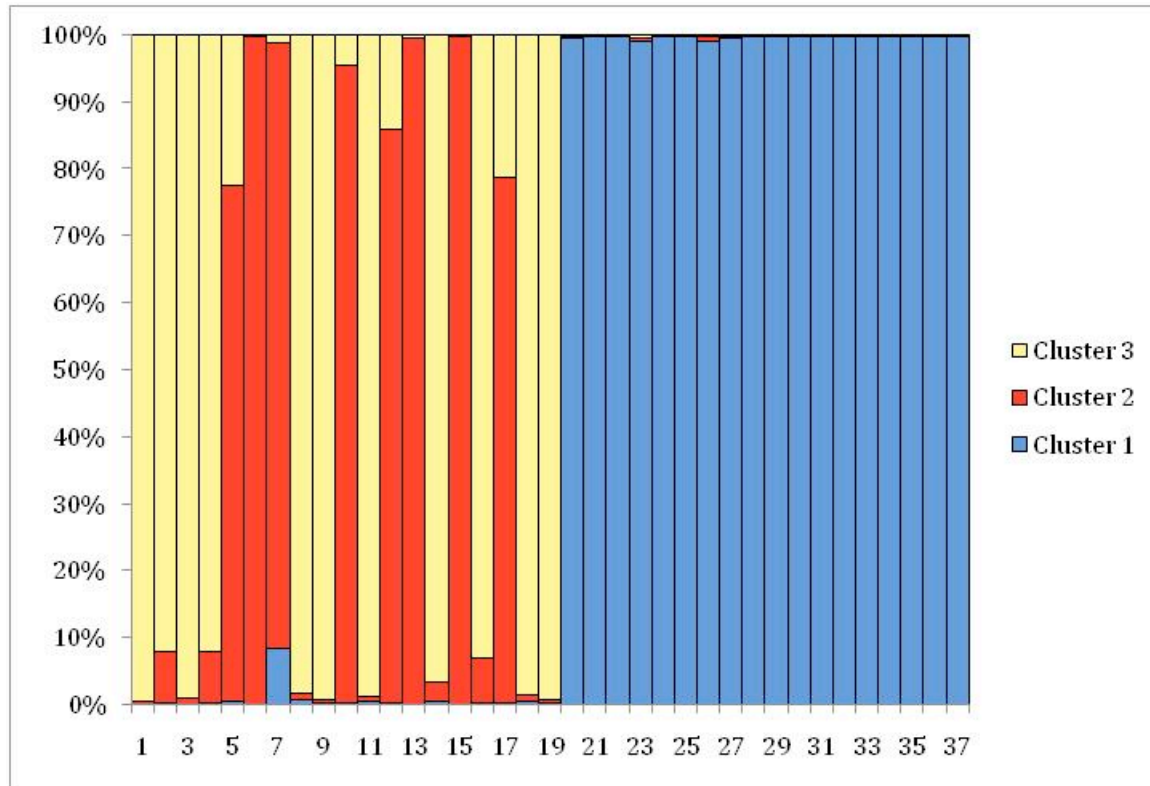


Figure S19-1. Population structure of 37 orangutans on the basis of 85 polymorphic retrotransposon markers. Samples 1-19 are of Sumatran origin with sample 1 being the reference genome individual, “Susie”; samples 20-37 are Bornean orangutans. The 85 polymorphic retrotransposon markers were identified from either the orangutan reference genome (Sumatran) or short paired-end Illumina sequence reads of a Bornean orangutan. Information regarding the geographic origin of the 37 samples was omitted from the Structure analysis runs to allow the program to infer the most likely number of populations. Results of the Structure analysis show that the most likely number of clusters (populations) was three ($K=3$) with cluster 1 (blue) representing the Bornean orangutans and clusters 2 (red) and 3 (yellow) representing population structure within the Sumatran sample set.

Supplemental Section S20 – SNP Calling and Ancestral Base Reconstruction

Alignment of Next-generation Sequence Data

Illumina sequence reads (see Supplemental Section S4 for details) from all 10 donor individuals were aligned against the Sumatran reference genome (v2.0.2) using Novoalign (www.novocraft.com). Only reads with less than 1 mismatch per 17 bp of “effective sequence” (i.e., excluding ambiguous base calls and 2 bp 5’ and 3’ of the read ends) were retained for SNP calling. Furthermore, we required that both reads from a mate pair align to this threshold and fall within the bounds of a log-normal estimated distribution for insert size in order to retain either read from the pair. As we see in Figure S20-1, for three libraries from individual KB5404, a Bornean donor individual

sequenced to ~20X coverage, the log-normal distribution does an excellent job of modeling the dispersion around the modal insert size

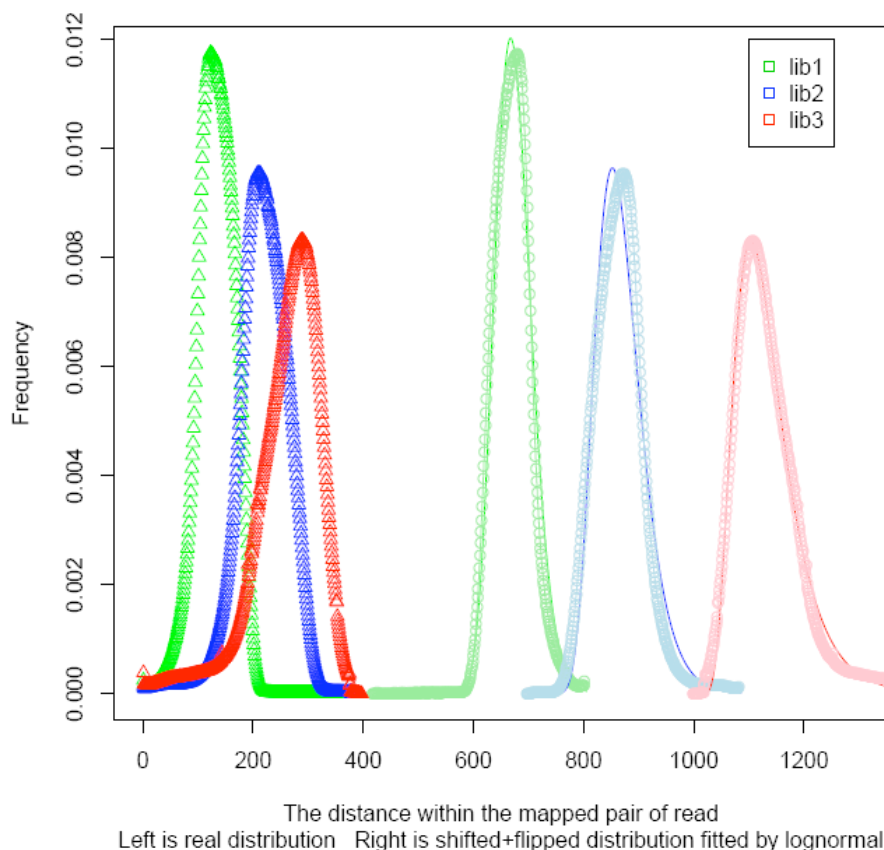


Figure S20-1: Distribution of pair span for three libraries sequenced for the high coverage (20x) Bornean individual (KB5404).

Recalibration of Illumina Read Error Rates

To account for potential biases in the assignment of Illumina quality scores, we first used Dohm et al.'s recalibration method¹³⁹. Specifically, for reads of length L (e.g., $L = 36$ or $L = 72$), at a given position i ($1 \leq i \leq n$) for individual j ($1 \leq j \leq 10$), the empirical quality score function e_{ijL} is:

$$e_{ijL} = (\# \text{ of non-reference alleles at position } i \text{ for reads from individual } j \text{ of length } n) / (\text{total } \# \text{ of reads that map to position } i \text{ for reads from individual } j \text{ of length } n).$$

We then used the lower of the two scores (raw and recalibrated) in the SNP calling algorithm to reduce the influence of sequencing error on SNP calls, which potentially increased the false negative rate. We therefore assessed the impact of this scheme on the false negative rate in our simulation study (see below).

SNP Calling Algorithm

SNPs for each population were called separately using a Bayesian population genomic approach that pools information regarding allele frequency among individuals within the same species when calling genotypes, i.e. it leverages the short-read sequence data for all individuals of given species aligned to the reference genome. The model utilizes a prior distribution on allele frequencies for variable sites as well as Hardy-Weinberg assumptions at the species level within species to derive a posterior distribution on genotype for each individual at each SNP based on the sequence data. We also applied a series of post- and pre-calling filters to reduce the possibility of errors including the following. First, we removed positions with an ambiguous base (N) in the reference genome, along with 5 bp 5' and 3' of that position, since the presence of even a single ambiguous base is an effective indicator of low-quality sequence¹⁴⁰. SNPs were only called where the reference genome had a consensus quality score greater than 90 (on a scale of 1-97, based on the phred scores of underlying whole-genome shotgun reads). Regions of known segmental duplication were also excluded. Furthermore, we required all potential SNP sites to have at least 7 individuals with greater than 2X coverage at that locus in order to be considered in the population genetic analysis. Finally, we did not allow for SNPs within 5 bp of each other, indels within 10 bp, or more than 8 individuals to be classified as heterozygous, in order to minimize the rate of false positives caused by recent segmental duplications.

The details of our calling algorithm are as follows. We wish to estimate the genotype for a given individual by jointly considering the reads for that individual and the estimate of the frequency for the allele in the population given the genotype calls. Under standard population genetic theory the allele frequency distribution for a single allele within a population follows a beta distribution^{141,142}. Therefore, our model assumes a vague prior distribution with a skew towards rare alleles for the minor allele frequency within each of the two populations by utilizing a beta distribution with parameters $\left[\begin{array}{c} \text{---} \\ \text{---} \end{array} \right]$. Denote the 10 individuals as $\left[\begin{array}{c} \text{---} \\ \text{---} \end{array} \right]$. For a particular site on the genome, let A and a be, respectively, the major and minor allele. Let \hat{p} represent the minor allele frequency for a specific population at this site, where the prior $P(\hat{p}) = \text{Beta}(\alpha = 0.01, \beta = 0.09)$; let N_i represent the total number of alleles observed for individual i ; let r_{ij} be the type of the j^{th} allele copy among these N_i allele copies where $j = 1 \dots N_i$; let e_{ij} be the corresponding error probability determined as above by either the recalibrated or raw quality score. For a particular site on the reference genome, we have that:

$$\begin{aligned} \mathbb{P}(\text{Genotype}|\text{Data}) &\sim \mathbb{P}(\text{Data}|\text{Genotype}) \cdot \mathbb{P}(\text{Genotype}) \\ &\sim \mathbb{P}(\text{Data}|\text{Genotype}) \cdot \int \mathbb{P}(\text{Genotype}|\hat{p}) \cdot p(\hat{p}) d\hat{p} \\ &\sim \mathbb{P}(\text{Data}|\text{Genotype}) \cdot \int \mathbb{P}(\text{Genotype}|\hat{p}) \cdot \text{Beta}(\alpha, \beta) d\hat{p} \end{aligned}$$

Therefore, for individual i , the posterior distribution on the three possible genotypes (AA, Aa, aa) are:

$$\begin{aligned} \mathbb{P}_i(AA|Data) &\sim \mathbb{P}(Data|AA) \cdot \int (1 - \hat{p})^2 \cdot Beta(\alpha, \beta) d\hat{p} \\ &\sim \frac{\prod_j^{N_i} (1 - e_{ij})^{1_{(r_{ij}=A)}} \cdot e_{ij}^{1_{(r_{ij}=a)}} \cdot \beta(\beta + 1)}{(\alpha + \beta)(\alpha + \beta + 1)} \end{aligned}$$

$$\begin{aligned} \mathbb{P}_i(Aa|Data) &\sim \mathbb{P}(Data|Aa) \cdot \int 2\hat{p} \cdot (1 - \hat{p}) \cdot Beta(\alpha, \beta) d\hat{p} \\ &\sim \frac{(0.5)^{N_i} \cdot 2\alpha\beta}{(\alpha + \beta)(\alpha + \beta + 1)} \end{aligned}$$

$$\begin{aligned} \mathbb{P}_i(aa|Data) &\sim \mathbb{P}(Data|aa) \cdot \int \hat{p}^2 \cdot Beta(\alpha, \beta) d\hat{p} \\ &\sim \frac{\prod_j^{N_i} (1 - e_{ij})^{1_{(r_{ij}=a)}} \cdot e_{ij}^{1_{(r_{ij}=A)}} \cdot \alpha(\alpha + 1)}{(\alpha + \beta)(\alpha + \beta + 1)} \end{aligned}$$

Based on the posterior probability for each one of the three potential genotypes for individuals within the same population, the genotype frequencies can be estimated for the population and the parameter of the allele frequency distribution, alpha and beta, can be updated with respect to the population level using the following method:

$$\begin{aligned} \hat{\alpha} &= \sum_i^K (\mathbb{P}_i(aa|Data) * 2 + \mathbb{P}_i(Aa|Data) * 1 + \mathbb{P}_i(AA|Data) * 0) \\ \hat{\beta} &= \sum_i^K (\mathbb{P}_i(AA|Data) * 2 + \mathbb{P}_i(Aa|Data) * 1 + \mathbb{P}_i(aa|Data) * 0) \end{aligned}$$

Where K includes the indexes of individuals which belongs to the same population and the iteration will continue until the two consecutive updated parameters differ less than 0.001. In the above posterior probability formulae, $1_{(r_{ij}=A)}$ denotes the indicator function, which is 1 when $r_{ij} = A$ and 0 otherwise. Based on the posterior probabilities we classified the site as a variate site or nonvariate site exclusively.

Simulation Study

To estimate our SNP calling sensitivity and false discovery rate, we adapted a modified simulation protocol from the 1000 Genomes project. The dataset, which consists of 500 individuals, was simulated under a split model with the time of separation for Bornean and Sumatran populations of approximately 1M years ago. 10 individuals were randomly chosen and 36 basepair paired-end reads were generated by ART (<http://biomedempire.org/>) at the exact observed coverage in the real data for each individual, in order to be comparable to the real dataset. A single individual in the Sumatran group was randomly chosen as the reference genome and all the short reads were aligned against it. The simulation results are listed below in Table S20-1:

Table S20-1.

Individual	Coverage	Correctly called	Incorrectly called	Missed	Sensitivity	FDR
1. 5504	20X	33819	175	194	98.9%	0.6%
2. 5503	10X	33640	589	904	95.8%	2.6%
3. 9258	9X	23240	984	1204	91.4%	4.7%
4. 5406	7X	21693	1036	1537	89.4%	6.3%
5. 5405	8X	23161	1154	1342	90.3%	5.2%
6. 4202	8X	33267	740	1318	94.2%	3.7%
7. 550	7X	33024	847	1515	93.3%	4.3%
8. 5883	8X	33074	748	1363	94.0%	3.9%
9. 4361	6X	22532	1666	1806	86.6%	6.9%
10. 4661	6X	22211	1721	1831	86.2%	7.1%

SNP Validation

We validated a subset of SNPs by PCR-based re-sequencing on the 3730 platform. SNPs were selected from arbitrarily chosen regions of the orangutan genome, sampling from chromosomes 1 and 3-11, with the sole requirement that predicted genotypes were available for all 10 sequenced orangutan individuals at each site. Several categories of SNP were selected, including singletons, doubletons and higher frequency SNPs with three or more alleles observed among the 10 individuals we sequenced. Among doubletons and higher frequency SNPs, both heterozygous and homozygous sites were selected, as well as sites with a combination of heterozygous and homozygous alleles in the higher frequency category. Overall, the set is biased toward singleton SNPs (63 out of 108 sites, see below) to assess the ability of the SNP caller to successfully detect such sites with 8-10x coverage of short read sequence alignments.

From an initial set of 114 sites, 108 amplicons were successfully designed and sequenced. Manual genotype calls were then made at sites with sufficient Sanger data quality, which allowed 87.0% (940/1,080) of all possible genotypes to be called (Table S20-2). Overall genotyping accuracy, defined as the concordance between the

predicted genotype and the Sanger data was very high, with 99.0% (931/940) of sites confirming computational predictions.

Genotyping accuracy for the singleton pool was high (98.9%), but this figure includes validated sites that were homozygous with respect to the reference genome for 9/10 individuals. Specifically among the individuals bearing a heterozygous singleton SNP we found 4 false positives out of 51 sites where validation data was available (a 7.8% false positive rate). Of the four false positive heterozygous singleton calls, one was a mis-called homozygous variant; the other three were homozygous wildtype alleles according to the Sanger data. We also found 2 false negative calls among the singleton pool (489 calls total) for a false negative rate of 0.4%. For the 11 doubleton SNPs in our validation set, a full 100% (110/110) were concordant between the Sanger data and the predicted genotypes. For higher frequency SNPs the overall concordance was 99.0% (291/294), and the validation rate of non-reference allele genotypes was 98.0% (149/152) with one false negative (0.7%). Overall, the high rate of concordance between the genotypes predicted by the SNP caller and the Sanger-based sequence data suggests a high level of accuracy among the large pool of SNPs detected across the orangutan genome using this methodology. These results should provide confidence in the use of these SNPs in downstream analyses and applications.

Table S20-2.

SNP Category	Sites Assessed	Resequencing Success Rate Across All 10 Individuals	Overall Genotype Accuracy	Non-reference Allele Accuracy	False Negatives
Singletons (1 allele observed)	63	540/630 (85.7%)	534/540 (98.9%)	47/51 (92.2%)	2/489 (0.4%)
Doubletons (2 alleles)	11	106/110 (96.4%)	106/106 (100%)	15/15 (100%)	0/91 (0.0%)
High Frequency (3+ alleles)	34	294/340 (86.5%)	291/294 (99.0%)	149/152 (98.0%)	1/142 (0.7%)
Total	108	940/1,080 (87.0%)	931/940 (99.0%)	211/218 (96.8%)	3/722 (0.4%)

Principal Component Analysis

To quantify patterns of population substructure using the 13 million SNPs discovered using our algorithm, we use a modified Principal Component Approach. Specifically, we encoded each SNP for each individual as the posterior expected # of copies of the alternate (i.e., non-reference) allele for sites with coverage of at least 2X per individual:

$$E(\# \text{ of copies of "a" | Data}) = 2 * \Pr(aa | \text{Data}) + 1 * \Pr(Aa | \text{Data})$$

This approach integrates out over uncertainty in assignment of the heterozygous vs. homozygous state. We also utilized PCA on the genotype matrix based on a maximum *a posteriori* approach that assigned each individual the genotype with highest posterior probability. The results are qualitatively very similar and show the extremely high quality of the data. Namely, the first principal component (PC 1) separates the Bornean from Sumatran samples and explains approximately 35.8% of the variance. The second PC identifies one Sumatran individual (9258) as distinct from the other four. This observation is consistent with the higher overall genetic diversity of Sumatran orangutans (see main paper).

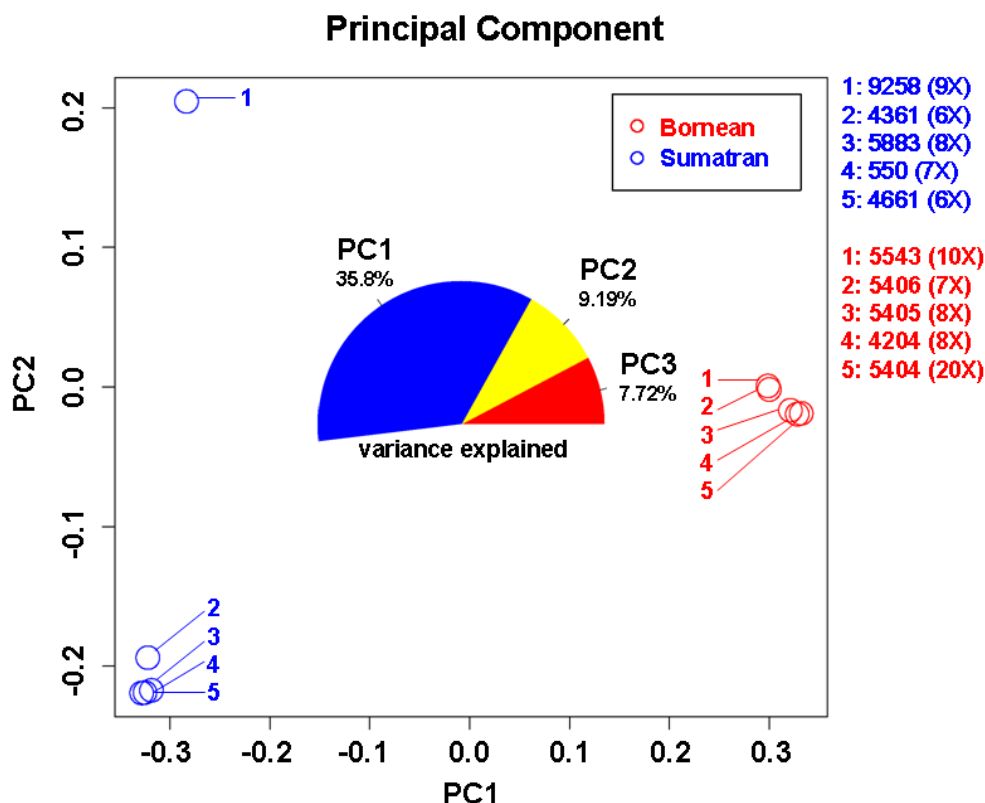


Figure S20-2. Principal component analysis of the genomic SNP data found by resequencing 10 Bornean and Sumatran donor individuals to at least 6X coverage.

Ancestral Base Reconstruction

The ancestral base reconstruction of SNPs from the Bornean and Sumatran orangutan population sequence data made use of the human (hg18), rhesus macaque (rheMac2) and chimpanzee (panTro2) assemblies. Alignments for these three species were extracted from the 44-way multi-species alignments hosted at the University of California, Santa Cruz¹⁴³. Autosomal genome-wide estimates of branch length were obtained from four-fold degenerate sites in these alignments using PhyloFit¹⁴⁴ under the general time-reversible (GTR) substitution model (see Pollard et al., 2010¹⁴⁵ for further details). An approximate estimate for the average branch lengths leading to Bornean and Sumatran orangutans was obtained from alignments of two individuals (one from each subspecies) for chromosome 1. Marginal posterior distributions over the four bases in the most recent common ancestor of the Bornean and Sumatran orangutans were then computed for each polymorphic site by the sum-product algorithm, using prequel (<http://compgen.bscb.cornell.edu/phast/>). Here, the Sumatran individuals were summarized with one sequence and the Bornean individuals with another sequence, and IUPAC ambiguity characters were used to represent polymorphic sites within these populations. Prequel integrates over the possible bases associated with each ambiguity

character in its calculations. The maximal value of the computed posterior distribution of ancestral bases, per position, is over 0.9 in 93% of cases.

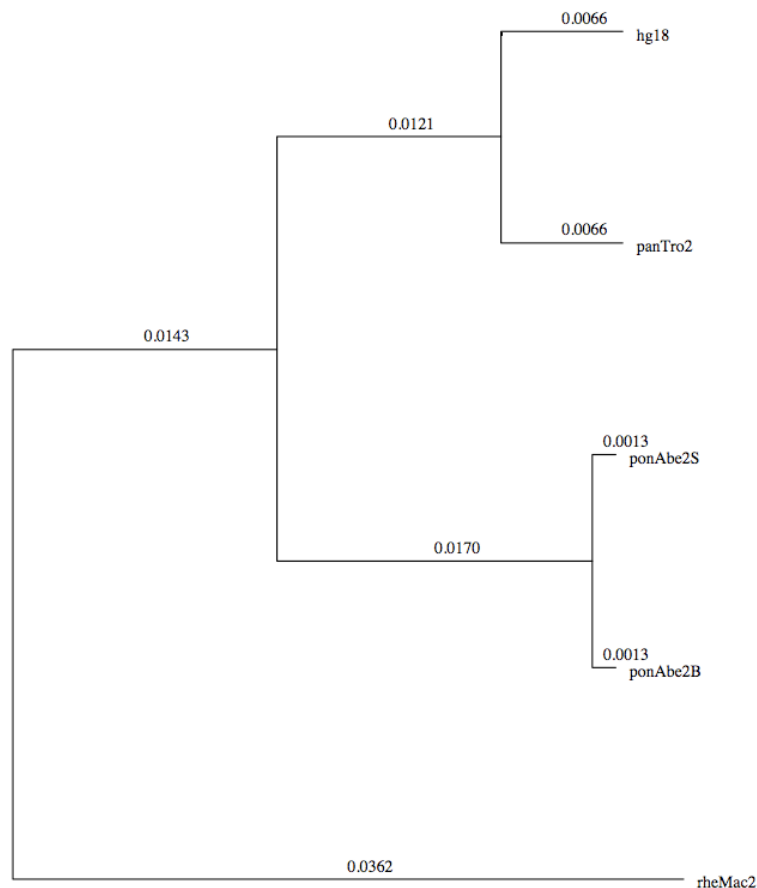


Figure S20-3. Branch length estimates based on alignment of macaque, chimpanzee, human, Bornean orangutan and Sumatran orangutan sequence data. Shown are branch length estimates based on alignments of Bornean and Sumatran orangutan sequence data to chromosome 1, with 92,232 informative sites over 21,450,994 bases of total alignment space. Branch lengths for the other species were autosomal whole-genome estimates derived from the 44-way alignments hosted at the UCSC Genome Browser.

Supplemental Section S21 – Demographic Inference Using DaDi

The data consist of SNPs detected in 1.96 Gb of genome sequenced in each of 5 Bornean and 5 Sumatran individuals. To account for occasional missing data, the frequency spectrum was projected down to 8 chromosome samples per population. The resulting cutoff of having 4 or more individuals called in each population yielded 12.74 million usable SNPs in the frequency spectrum.

Two analyses were performed, differing in how SNPs were polarized. We first worked with the folded spectrum, which ignores ancestral state information and considers only minor allele frequencies. The upper-left panel of **Figure S21-1** shows the resulting spectrum. We found that this spectrum very poorly constrained the split time in models with migration, so we also worked with a polarized spectrum. For the polarized spectrum, we used the ancestral state inferred using the algorithm discussed in **Section S20**. That algorithm assigns a probability for each possible ancestral state for each SNP, and the renormalized probabilities of the two segregating states were used for each SNP in the data set. One limitation of the current ancestral state algorithm is that if allele A is fixed in one population sample and the second sample has alleles A and G segregating, the ancestral state is always called as A. To compensate for this limitation, the data and model spectra were partially folded, ignoring ancestral state information for entries in which one allele is fixed in a population. The upper-left panel of **Figure S21-2** shows the resulting spectrum.

We fit two increasingly complex series of nested models to the two frequency spectra. First, we fit a series of models in which the two derived population sizes are held constant since the time of the split, as in **Figure S21-3 a**). We then fit a series of models with potential exponential growth since the time of split, as in **Figure S21-3 b**). **Table S21-1** and **Table S21-2** report the results of these fits to the two spectra. In the tables, each pair of rows corresponds to a different model. The first row contains the parameters in genetic units, while the second has converted them to physical units using a per generation mutation rate of 2.0×10^{-8} and a generation time of 20 years. In each case, the parameters allowed to vary are highlighted.

Considering the maximum-likelihood model parameters, we see that the Sumatran current effective population size is always estimated to be substantially larger than the Bornean. In models fit to the folded spectrum, we found evidence for a ridge in the likelihood surface that traded off longer split times for higher migration (data not shown). As a result, the split times inferred from that analysis are unreliable, with larger uncertainties. In the fits to the polarized spectrum, we recover split times of roughly 400 thousand years ago. Also, in all cases we find evidence for low levels of migration between the two populations.

For the most complex model with growth and migration, **Figure S21-1** and **Figure S21-2** compare the maximum likelihood model spectra with the data. In both analyses, we see that the models are underestimating the number of shared low-frequency polymorphisms. This may indicate that a more complex time-dependent migration model might be desirable. We also see that the model fit to the polarized spectrum underestimates the number of observed high-frequency shared mutations. This may indicate additional polarization difficulties, as it is difficult to concoct scenarios in which there are more shared high-frequency polymorphisms than shared mid-frequency polymorphisms. Nevertheless, the overall consistency of results among the different models suggests lends them credence.

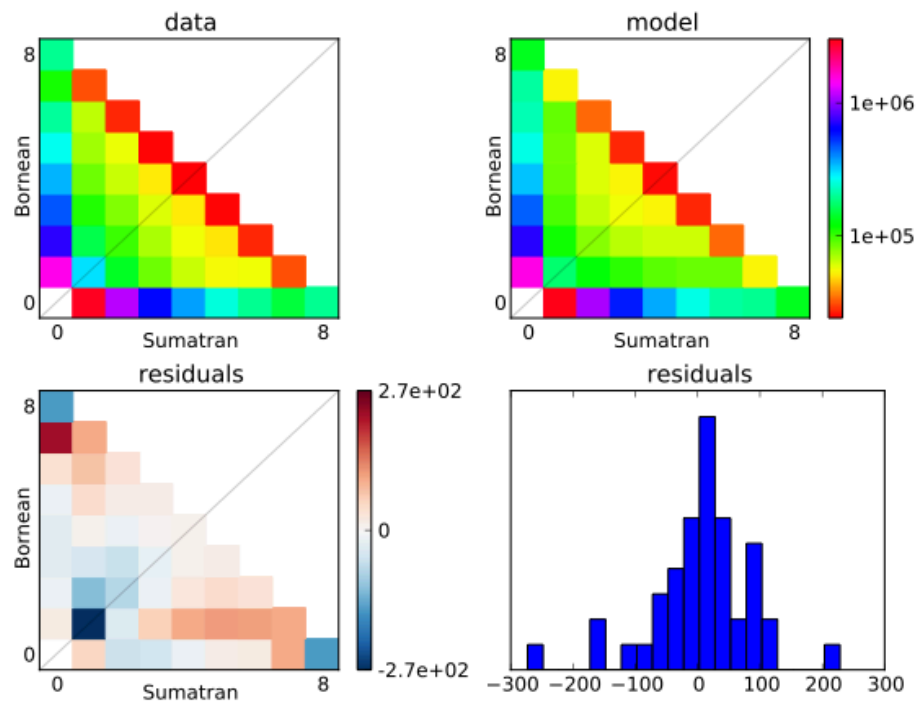


Figure S21-1: Analysis results for completely folded spectrum

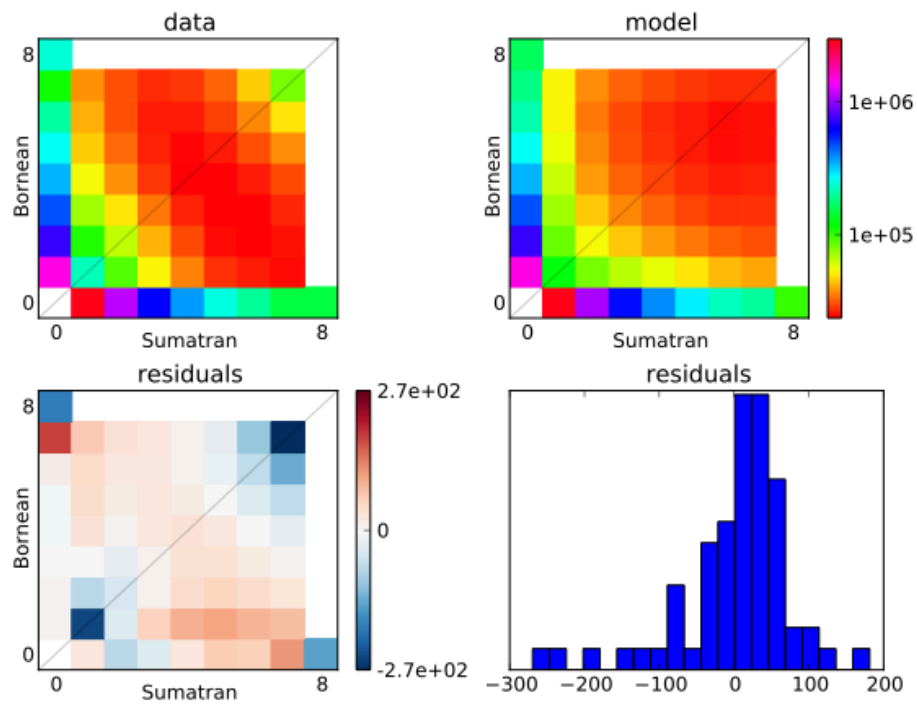
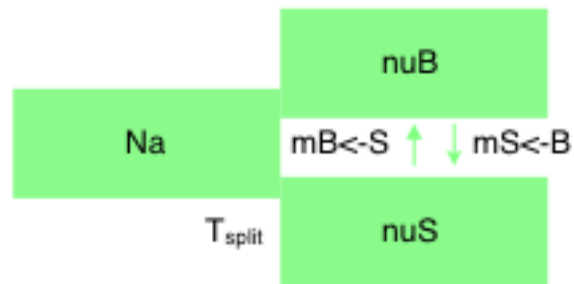


Figure S21-2: Analysis results for polarized spectrum

Figure S21-3: Demographic models considered

a) Fixed derived population sizes



b) Exponential growth

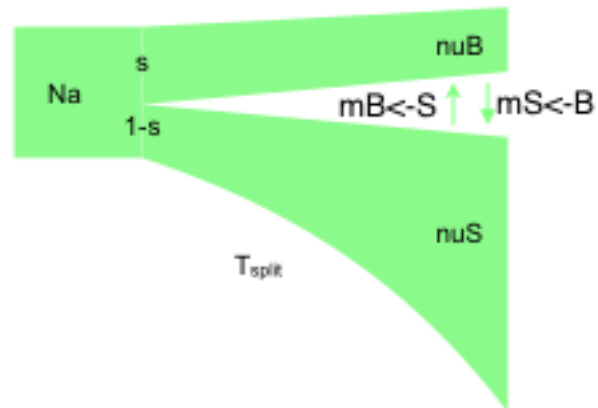


Table S21-1: Maximum-likelihood parameters for models fit to completely folded data.

log(likelihood)	theta & Na	T_{split}	s	nuB	nuS	mB<-S	mS<-B
-6.22E+05	2.59E+06	0.488		1	1	0	0
	16,518	322,429		16,518	16,518		
-3.07E+05	2.97E+06	0.353		0.516	1.008	0	0
	18,954	267,632		9,780	19,098		
-1.73E+05	4.54E+06	10.000		3.592	6.716	0.071	0.071
	28,948	11,579,082		103,980	194,413		
-1.36E+05	4.55E+06	10.000		2.932	7.477	0.118	0.040
	28,986	11,594,388		84,987	216,728		
-6.64E+05	3.45E+06	0.256	0.5	0.5	0.5	0	0
	22,015	225,437		11,008	11,008		
-3.95E+05	3.51E+06	0.234	0.344	0.344	0.656	0	0
	22,353	209,227		7,690	14,664		
-2.82E+05	3.05E+06	0.326	0.505	0.422	2.403	0	0
	19,452	253,648		8,209	46,742		
-1.61E+05	2.68E+06	0.689	0.618	0.569	2.067	0.343	0.343
	17,073	470,524		9,714	35,289		
-1.50E+05	2.46E+06	0.908	0.503	0.581	2.071	0.546	0.229
	15,702	570,280		9,123	32,518		

Table S21-2: Maximum-likelihood parameters for models fit to data with probabilistic ancestry and fixed+seg sites folded.

log(likelihood)	theta & Na	T _{split}	s	nuB	nuS	mB<-S	mS<-B
-6.35E+05	2.59E+06	0.487		1	1	0	0
	16,524	321,892		16,524	16,524		
-3.09E+05	2.98E+06	0.353		0.518	1.066	0	0
	18,973	267,902		9,828	20,225		
-2.21E+05	2.64E+06	0.626		0.608	1.157	0.269	0.269
	16,830	421,432		10,233	19,473		
-2.08E+05	2.59E+06	0.662		0.560	1.254	0.447	0.169
	16,537	437,899		9,261	20,737		
-6.82E+05	3.45E+06	0.256	0.5	0.5	0.5	0	0
	22,003	225,306		11,001	11,001		
-3.96E+05	3.51E+06	0.234	0.345	0.344	0.655	0	0
	22,353	209,227		7,690	14,641		
-2.84E+05	3.05E+06	0.326	0.601	0.423	2.389	0	0
	19,445	253,565		8,225	46,454		
-1.87E+05	2.84E+06	0.536	0.633	0.498	2.148	0.294	0.294
	18,087	387,780		9,007	38,850		
-1.84E+05	2.81E+06	0.562	0.592	0.491	2.100	0.395	0.239
	17,934	403,149		8,805	37,661		

Short Read Coverage Effects

A potential concern is that our low-coverage data may bias our estimation of SNP frequencies and thus our model fitting, although our SNP calling algorithm is explicitly designed to correctly work with low-coverage data. To test for potential biases, we compared results from the full data set to results from a data set filtered based on read depth. For this “strict” data set, we included only SNPs for which each individual had read depth between the mean read depth for that individual and twice that value. We thus restricted ourselves to SNPs with relatively high coverage, but we exclude SNPs with very high coverage that may indicate mis-mapping of copy-number variation. This strict filtering left us with 5280 SNPs for analysis.

We first checked whether the frequency spectrum (FS) from the “strict” data set differed statistically from the full data FS. The upper-left panel of **Figure S21-4** shows the FS from the full data set, while upper right panel shows the FS from the “strict” data set. Qualitatively, they are very similar.

To more quantitatively test for deviations between these the full and strict data sets, we compared our strict FS with 10,000 spectra resulting from randomly sampling 5280

SNPs from the full data set. We then calculated a p-value for each entry in the FS, based on where the “strict” data set was within the distribution of values derived from the bootstraps of the full data set. These p-values are plotted in the lower-left panel of **Figure S21-4**. We see little correlation in the p-values, again qualitatively suggesting that our SNP calling is not biased. Furthermore, no entries in the strict FS differ significantly (at the 5% level) from the entries in the full data FS.

Finally, we refit all the models we considered to the strict data set, to check whether the model fitting is biased. The resulting maximum-likelihood parameter values are shown in **Table S21-3**. Comparing with our fits to the full data set (**Table S21-2**), we see that in all cases the inferred parameter values are very similar. Thus using even a very strictly filtered data set does not change any of our conclusions.

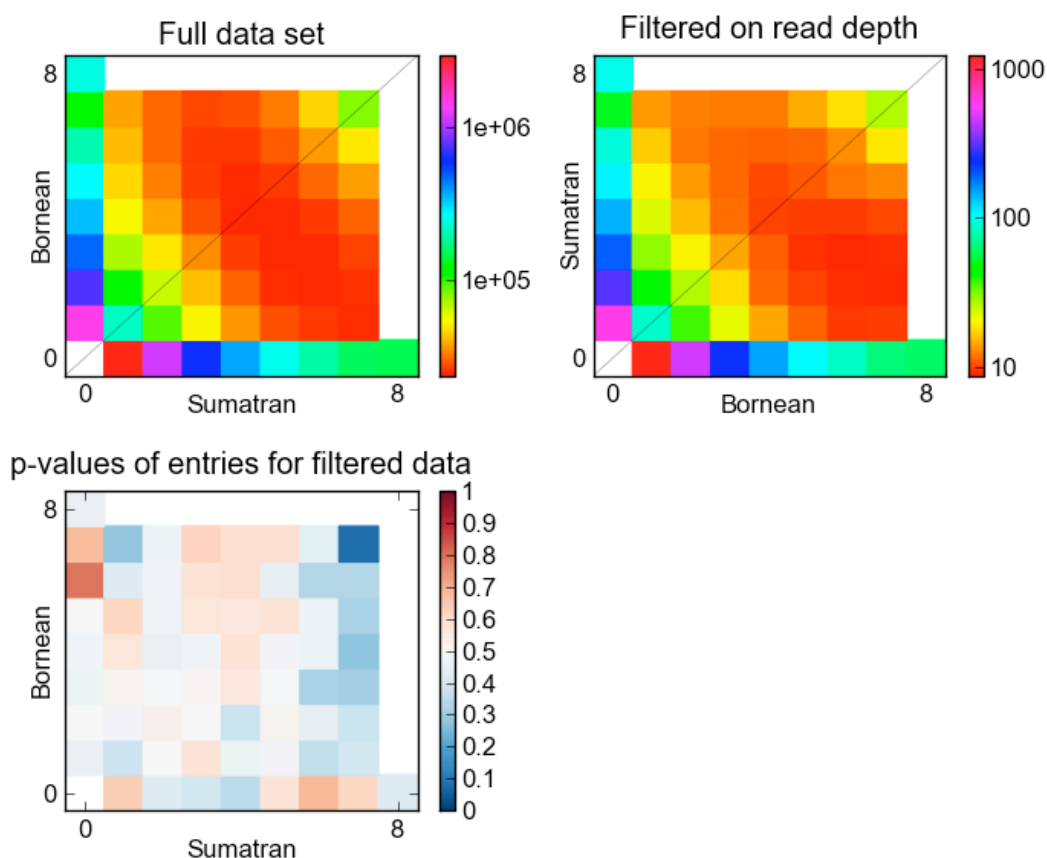


Figure S21-4: Comparison of spectra for full data set and data set filtered on read depth.

Table S21-3: Maximum-likelihood parameters for models fit to data with probabilistic ancestry and fixed+seg sites folded. In these fits, only SNPs which passed our read depth criteria were used.

log(likelihood)	Tsplit	s	nuB	nuS	mB<-S	mS<-B
-421.4	0.497		1	1	0	0
-281.5	0.357		0.509	1.07	0	0
-245.7	0.650		0.608	1.174	0.268	0.268
-239.0	0.704		0.554	1.292	0.459	0.151
-438.4	0.260	0.5	0.5	0.5	0	0
-318.1	0.237	0.340	0.340	0.660	0	0
-269.0	0.319	0.656	0.344	2.814	0	0
-235.9	0.620	0.605	0.478	1.756	0.357	0.357
-229.6	0.569	0.594	0.453	2.135	0.436	0.224

Supplemental Section S22 – References

- ¹ Huang, X., Wang, J., Aluru, S., Yang, S. P. & Hillier, L. PCAP: a whole-genome assembly program. *Genome Res* **13**, 2164-2170, doi:10.1101/gr.1390403 13/9/2164 [pii] (2003).
- ² Mewes, H. W. *et al.* MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res* **32**, D41-44, doi:10.1093/nar/gkh092 32/suppl_1/D41 [pii] (2004).
- ³ Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664, doi:10.1101/gr.229202. Article published online before March 2002 (2002).
- ⁴ Chiaromonte, F., Yap, V. B. & Miller, W. Scoring pairwise genomic sequence alignments. *Pac Symp Biocomput*, 115-126 (2002).
- ⁵ Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res* **13**, 103-107, doi:10.1101/gr.809403 (2003).
- ⁶ Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A* **100**, 11484-11489, doi:10.1073/pnas.1932072100 1932072100 [pii] (2003).
- ⁷ Lunter, G., Ponting, C. P. & Hein, J. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput Biol* **2**, e5, doi:10.1371/journal.pcbi.0020005 (2006).
- ⁸ Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851-1858, doi:gr.078212.108 [pii] 10.1101/gr.078212.108 (2008).

- ⁹ Gordon, D. Viewing and editing assembled sequences using Consed. *Curr Protoc Bioinformatics* **Chapter 11**, Unit11 12, doi:10.1002/0471250953.bi1102s02 (2003).
- ¹⁰ Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87, doi:nature04072 [pii] 10.1038/nature04072 (2005).
- ¹¹ Meader, S., Hillier, L. W., Locke, D., Ponting, C. P. & Lunter, G. Genome assembly quality: assessment and improvement using the neutral indel model. *Genome Res* **20**, 675-684, doi:gr.096966.109 [pii] 10.1101/gr.096966.109.
- ¹² Curwen, V. *et al.* The Ensembl automatic gene annotation system. *Genome Res* **14**, 942-950, doi:10.1101/gr.1858004 14/5/942 [pii] (2004).
- ¹³ Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31, doi:1471-2105-6-31 [pii] 10.1186/1471-2105-6-31 (2005).
- ¹⁴ Warren, W. C. *et al.* Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453**, 175-183, doi:nature06936 [pii] 10.1038/nature06936 (2008).
- ¹⁵ Ma, J. *et al.* Reconstructing contiguous regions of an ancestral genome. *Genome Res* **16**, 1557-1565, doi:gr.5383506 [pii] 10.1101/gr.5383506 (2006).
- ¹⁶ Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res* **18**, 1814-1828, doi:gr.076554.108 [pii] 10.1101/gr.076554.108 (2008).
- ¹⁷ Ma, J. *et al.* The infinite sites model of genome evolution. *Proc Natl Acad Sci U S A* **105**, 14254-14261, doi:0805217105 [pii] 10.1073/pnas.0805217105 (2008).
- ¹⁸ Conant, G. C. & Wagner, A. Asymmetric sequence divergence of duplicate genes. *Genome Res* **13**, 2052-2058, doi:10.1101/gr.1252603 13/9/2052 [pii] (2003).
- ¹⁹ Marques-Bonet, T., Cheng, Z., She, X., Eichler, E. E. & Navarro, A. The genomic distribution of intraspecific and interspecific sequence divergence of human segmental duplications relative to human/chimpanzee chromosomal rearrangements. *BMC Genomics* **9**, 384, doi:1471-2164-9-384 [pii] 10.1186/1471-2164-9-384 (2008).
- ²⁰ Marques-Bonet, T. *et al.* On the association between chromosomal rearrangements and genic evolution in humans and chimpanzees. *Genome Biol* **8**, R230, doi:gb-2007-8-10-r230 [pii] 10.1186/gb-2007-8-10-r230 (2007).
- ²¹ Navarro, A. & Barton, N. H. Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *Evolution* **57**, 447-459 (2003).
- ²² Locke, D. P. *et al.* Refinement of a chimpanzee pericentric inversion breakpoint to a segmental duplication cluster. *Genome Biol* **4**, R50, doi:10.1186/gb-2003-4-8-r50 (2003).

- ²³ Weise, A. *et al.* New aspects of chromosomal evolution in the gorilla and the orangutan. *Int J Mol Med* **19**, 437-443 (2007).
- ²⁴ Szamalek, J. M., Cooper, D. N., Hoegel, J., Hameister, H. & Kehrer-Sawatzki, H. Chromosomal speciation of humans and chimpanzees revisited: studies of DNA divergence within inverted regions. *Cytogenet Genome Res* **116**, 53-60, doi:000097417 [pii] 10.1159/000097417 (2007).
- ²⁵ Szamalek, J. M. *et al.* Molecular characterisation of the pericentric inversion that distinguishes human chromosome 5 from the homologous chimpanzee chromosome. *Hum Genet* **117**, 168-176, doi:10.1007/s00439-005-1287-y (2005).
- ²⁶ Szamalek, J. M., Goidts, V., Cooper, D. N., Hameister, H. & Kehrer-Sawatzki, H. Characterization of the human lineage-specific pericentric inversion that distinguishes human chromosome 1 from the homologous chromosomes of the great apes. *Hum Genet* **120**, 126-138, doi:10.1007/s00439-006-0209-y (2006).
- ²⁷ Goidts, V. *et al.* Independent intrachromosomal recombination events underlie the pericentric inversions of chimpanzee and gorilla chromosomes homologous to human chromosome 16. *Genome Res* **15**, 1232-1242, doi:15/9/1232 [pii] 10.1101/gr.3732505 (2005).
- ²⁸ Goidts, V., Szamalek, J. M., Hameister, H. & Kehrer-Sawatzki, H. Segmental duplication associated with the human-specific inversion of chromosome 18: a further example of the impact of segmental duplications on karyotype and genome evolution in primates. *Hum Genet* **115**, 116-122, doi:10.1007/s00439-004-1120-z (2004).
- ²⁹ Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**, 555-556 (1997).
- ³⁰ Stanyon, R. *et al.* Primate chromosome evolution: ancestral karyotypes, marker order and neocentromeres. *Chromosome Res* **16**, 17-39, doi:10.1007/s10577-007-1209-z (2008).
- ³¹ Seuanez, H., Fletcher, J., Evans, H. J. & Martin, D. E. A chromosome rearrangement in orangutan studied with Q-, C-, and G-banding techniques. *Cytogenet Cell Genet* **17**, 26-34 (1976).
- ³² Carroll, C. W. & Straight, A. F. Centromere formation: from epigenetics to self-assembly. *Trends Cell Biol* **16**, 70-78, doi:S0962-8924(05)00332-6 [pii] 10.1016/j.tcb.2005.12.008 (2006).
- ³³ Umlauf, D., Goto, Y. & Feil, R. Site-specific analysis of histone methylation and acetylation. *Methods Mol Biol* **287**, 99-120, doi:1-59259-828-5:099 [pii] 10.1385/1-59259-828-5:099 (2004).
- ³⁴ Wells, J. & Farnham, P. J. Characterizing transcription factor binding sites using formaldehyde crosslinking and immunoprecipitation. *Methods* **26**, 48-56, doi:10.1016/S1046-2023(02)00007-5 S1046-2023(02)00007-5 [pii] (2002).
- ³⁵ Trazzi, S. *et al.* The C-terminal domain of CENP-C displays multiple and critical functions for mammalian centromere formation. *PLoS One* **4**, e5832, doi:10.1371/journal.pone.0005832 (2009).

- 36 Bieda, M., Xu, X., Singer, M. A., Green, R. & Farnham, P. J. Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Res* **16**, 595-605, doi:gr.4887606 [pii]
10.1101/gr.4887606 (2006).
- 37 Gosden, J. & Lawson, D. Rapid chromosome identification by oligonucleotide-primed in situ DNA synthesis (PRINS). *Hum Mol Genet* **3**, 931-936 (1994).
- 38 Lo, A. W., Liao, G. C., Rocchi, M. & Choo, K. H. Extreme reduction of chromosome-specific alpha-satellite array is unusually common in human chromosome 21. *Genome Res* **9**, 895-908 (1999).
- 39 Roy, A. M. *et al.* Recently integrated human Alu repeats: finding needles in the haystack. *Genetica* **107**, 149-161 (1999).
- 40 Ray, D. A., Han, K., Walker, J. A. & Batzer, M. A. in *Methods in Molecular Biology - Genetic Variation Edition* eds M. R. Barnes & G. Breen) (Humana Press Inc, In Press).
- 41 Ray, D. A. *et al.* Alu insertion loci and platyrrhine primate phylogeny. *Mol Phylogenet Evol* **35**, 117-126, doi:S1055-7903(04)00333-1 [pii]
10.1016/j.ympev.2004.10.023 (2005).
- 42 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).
- 43 Gibbs, R. A. *et al.* Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**, 222-234, doi:316/5822/222 [pii]
10.1126/science.1139247 (2007).
- 44 Wang, H. *et al.* SVA elements: a hominid-specific retroposon family. *J Mol Biol* **354**, 994-1007, doi:S0022-2836(05)01203-9 [pii]
10.1016/j.jmb.2005.09.085 (2005).
- 45 Chesnokov, I. & Schmid, C. W. Flanking sequences of an Alu source stimulate transcription in vitro by interacting with sequence-specific transcription factors. *J Mol Evol* **42**, 30-36 (1996).
- 46 Roy, A. M. *et al.* Upstream flanking sequences and transcription of SINEs. *J Mol Biol* **302**, 17-25, doi:10.1006/jmbi.2000.4027
S0022-2836(00)94027-0 [pii] (2000).
- 47 Roy-Engel, A. M. *et al.* Active Alu element "A-tails": size does matter. *Genome Res* **12**, 1333-1344, doi:10.1101/gr.384802 (2002).
- 48 Bennett, E. A. *et al.* Active Alu retrotransposons in the human genome. *Genome Res* **18**, 1875-1883, doi:gr.081737.108 [pii]
10.1101/gr.081737.108 (2008).
- 49 Comeaux, M. S., Roy-Engel, A. M., Hedges, D. J. & Deininger, P. L. Diverse cis factors controlling Alu retrotransposition: what causes Alu elements to die? *Genome Res* **19**, 545-555, doi:gr.089789.108 [pii]
10.1101/gr.089789.108 (2009).
- 50 Zhang, Z. *et al.* PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* **22**, 1437-1439, doi:btl116 [pii]
10.1093/bioinformatics/btl116 (2006).
- 51 Kaessmann, H., Vinckenbosch, N. & Long, M. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet* **10**, 19-31, doi:nrg2487 [pii]
10.1038/nrg2487 (2009).

- 52 Emerson, J. J., Kaessmann, H., Betran, E. & Long, M. Extensive gene traffic on the mammalian X chromosome. *Science* **303**, 537-540, doi:10.1126/science.1090042 303/5657/537 [pii] (2004).
- 53 Bai, Y., Casola, C., Feschotte, C. & Betran, E. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in *Drosophila*. *Genome Biol* **8**, R11, doi:gb-2007-8-1-r11 [pii] 10.1186/gb-2007-8-1-r11 (2007).
- 54 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, doi:10.1006/jmbi.1990.9999 S002283680799990 [pii] (1990).
- 55 Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res* **11**, 1005-1017, doi:10.1101/gr.187101 (2001).
- 56 Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003-1007, doi:10.1126/science.1072047 297/5583/1003 [pii] (2002).
- 57 Marques-Bonet, T. *et al.* A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**, 877-881, doi:nature07744 [pii] 10.1038/nature07744 (2009).
- 58 Newman, T. L. *et al.* A genome-wide survey of structural variation between human and chimpanzee. *Genome Res* **15**, 1344-1356, doi:gr.4338005 [pii] 10.1101/gr.4338005 (2005).
- 59 Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat Genet* **37**, 727-732, doi:ng1562 [pii] 10.1038/ng1562 (2005).
- 60 Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-1584 (2002).
- 61 Demuth, J. P., De Bie, T., Stajich, J. E., Cristianini, N. & Hahn, M. W. The evolution of mammalian gene families. *PLoS One* **1**, e85, doi:10.1371/journal.pone.0000085 (2006).
- 62 Hahn, M. W., De Bie, T., Stajich, J. E., Nguyen, C. & Cristianini, N. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res* **15**, 1153-1160, doi:15/8/1153 [pii] 10.1101/gr.3567505 (2005).
- 63 De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269-1271, doi:bt1097 [pii] 10.1093/bioinformatics/bt1097 (2006).
- 64 Hahn, M. W., Demuth, J. P. & Han, S. G. Accelerated rate of gene gain and loss in primates. *Genetics* **177**, 1941-1949, doi:genetics.107.080077 [pii] 10.1534/genetics.107.080077 (2007).
- 65 Quesada, V., Ordonez, G. R., Sanchez, L. M., Puente, X. S. & Lopez-Otin, C. The Degradome database: mammalian proteases and diseases of proteolysis. *Nucleic Acids Res* **37**, D239-243, doi:gkn570 [pii] 10.1093/nar/gkn570 (2009).

- ⁶⁶ Puente, X. S., Sanchez, L. M., Overall, C. M. & Lopez-Otin, C. Human and mouse proteases: a comparative genomic approach. *Nat Rev Genet* **4**, 544-558, doi:10.1038/nrg1111
nrg1111 [pii] (2003).
- ⁶⁷ Lopez-Otin, C. & Overall, C. M. Protease degradomics: a new challenge for proteomics. *Nat Rev Mol Cell Biol* **3**, 509-519, doi:10.1038/nrm858
nrm858 [pii] (2002).
- ⁶⁸ Puente, X. S. & Lopez-Otin, C. A genomic analysis of rat proteases and protease inhibitors. *Genome Res* **14**, 609-622, doi:10.1101/gr.1946304
14/4/609 [pii] (2004).
- ⁶⁹ Puente, X. S., Gutierrez-Fernandez, A., Ordonez, G. R., Hillier, L. W. & Lopez-Otin, C. Comparative genomic analysis of human and chimpanzee proteases. *Genomics* **86**, 638-647, doi:S0888-7543(05)00211-9 [pii]
10.1016/j.ygeno.2005.07.009 (2005).
- ⁷⁰ Ordonez, G. R. *et al.* Loss of genes implicated in gastric function during platypus evolution. *Genome Biol* **9**, R81, doi:gb-2008-9-5-r81 [pii]
10.1186/gb-2008-9-5-r81 (2008).
- ⁷¹ Chen, C., Darrow, A. L., Qi, J. S., D'Andrea, M. R. & Andrade-Gordon, P. A novel serine protease predominately expressed in macrophages. *Biochem J* **374**, 97-107, doi:10.1042/BJ20030242
BJ20030242 [pii] (2003).
- ⁷² Johnson, M. E. *et al.* Recurrent duplication-driven transposition of DNA during hominoid evolution. *Proc Natl Acad Sci U S A* **103**, 17626-17631, doi:0605426103 [pii]
10.1073/pnas.0605426103 (2006).
- ⁷³ Cook, M., Buhling, F., Ansorge, S., Tatnell, P. J. & Kay, J. Pronapsin A and B gene expression in normal and malignant human lung and mononuclear blood cells. *Biochim Biophys Acta* **1577**, 10-16, doi:S0167478102004001 [pii] (2002).
- ⁷⁴ Paulding, C. A., Ruvolo, M. & Haber, D. A. The Tre2 (USP6) oncogene is a hominoid-specific gene. *Proc Natl Acad Sci U S A* **100**, 2507-2511, doi:10.1073/pnas.0437015100
0437015100 [pii] (2003).
- ⁷⁵ Panagopoulos, I., Mertens, F., Lofvenberg, R. & Mandahl, N. Fusion of the COL1A1 and USP6 genes in a benign bone tumor. *Cancer Genet Cytogenet* **180**, 70-73, doi:S0165-4608(07)00596-1 [pii]
10.1016/j.cancergencyto.2007.09.017 (2008).
- ⁷⁶ Oliveira, A. M. *et al.* USP6 (Tre2) fusion oncogenes in aneurysmal bone cyst. *Cancer Res* **64**, 1920-1923 (2004).
- ⁷⁷ Brass, A. L. *et al.* Identification of host proteins required for HIV infection through a functional genomic screen. *Science* **319**, 921-926, doi:1152725 [pii]
10.1126/science.1152725 (2008).
- ⁷⁸ Xue, Y. *et al.* Spread of an inactive form of caspase-12 in humans is due to recent positive selection. *Am J Hum Genet* **78**, 659-670, doi:S0002-9297(07)63703-3 [pii]
10.1086/503116 (2006).

- ⁷⁹ Roy, S. *et al.* Confinement of caspase-12 proteolytic activity to autoprocessing. *Proc Natl Acad Sci U S A* **105**, 4133-4138, doi:0706658105 [pii] 10.1073/pnas.0706658105 (2008).
- ⁸⁰ Nguyen, D. H., Hurtado-Ziola, N., Gagneux, P. & Varki, A. Loss of Siglec expression on T lymphocytes during human evolution. *Proc Natl Acad Sci U S A* **103**, 7765-7770, doi:0510484103 [pii] 10.1073/pnas.0510484103 (2006).
- ⁸¹ McEvoy, S. M. & Maeda, N. Complex events in the evolution of the haptoglobin gene cluster in primates. *J Biol Chem* **263**, 15740-15747 (1988).
- ⁸² Erickson, L. M. & Maeda, N. Parallel evolutionary events in the haptoglobin gene clusters of rhesus monkey and human. *Genomics* **22**, 579-589, doi:S0888-7543(84)71431-5 [pii] 10.1006/geno.1994.1431 (1994).
- ⁸³ Vanhollebeke, B. *et al.* A haptoglobin-hemoglobin receptor conveys innate immunity to *Trypanosoma brucei* in humans. *Science* **320**, 677-681, doi:320/5876/677 [pii] 10.1126/science.1156296 (2008).
- ⁸⁴ Caughey, G. H. Mast cell tryptases and chymases in inflammation and host defense. *Immunol Rev* **217**, 141-154, doi:IMR509 [pii] 10.1111/j.1600-065X.2007.00509.x (2007).
- ⁸⁵ Trivedi, N. N., Raymond, W. W. & Caughey, G. H. Chimerism, point mutation, and truncation dramatically transformed mast cell delta-tryptases during primate evolution. *J Allergy Clin Immunol* **121**, 1262-1268, doi:S0091-6749(08)00151-6 [pii] 10.1016/j.jaci.2008.01.019 (2008).
- ⁸⁶ Gallwitz, M., Reimer, J. M. & Hellman, L. Expansion of the mast cell chymase locus over the past 200 million years of mammalian evolution. *Immunogenetics* **58**, 655-669, doi:10.1007/s00251-006-0126-1 (2006).
- ⁸⁷ Schultz, N., Hamra, F. K. & Garbers, D. L. A multitude of genes expressed solely in meiotic or postmeiotic spermatogenic cells offers a myriad of contraceptive targets. *Proc Natl Acad Sci U S A* **100**, 12201-12206, doi:10.1073/pnas.1635054100 1635054100 [pii] (2003).
- ⁸⁸ Schlecht, U. *et al.* Expression profiling of mammalian male meiosis and gametogenesis identifies novel candidate genes for roles in the regulation of fertility. *Mol Biol Cell* **15**, 1031-1043, doi:10.1091/mbc.E03-10-0762 E03-10-0762 [pii] (2004).
- ⁸⁹ Sharma, N. *et al.* Implantation Serine Proteinases heterodimerize and are critical in hatching and implantation. *BMC Dev Biol* **6**, 61, doi:1471-213X-6-61 [pii] 10.1186/1471-213X-6-61 (2006).
- ⁹⁰ Quesada, V., Sanchez, L. M., Alvarez, J. & Lopez-Otin, C. Identification and characterization of human and mouse ovastacin: a novel metalloproteinase similar to hatching enzymes from arthropods, birds, amphibians, and fish. *J Biol Chem* **279**, 26627-26634, doi:10.1074/jbc.M401588200 M401588200 [pii] (2004).

- 91 Honda, A., Okamoto, T. & Ishihama, A. Host factor Ebp1: selective inhibitor of influenza virus transcriptase. *Genes Cells* **12**, 133-142, doi:GTC1047 [pii] 10.1111/j.1365-2443.2007.01047.x (2007).
- 92 Okada, M., Jang, S. W. & Ye, K. Ebp1 association with nucleophosmin/B23 is essential for regulating cell proliferation and suppressing apoptosis. *J Biol Chem* **282**, 36744-36754, doi:M706169200 [pii] 10.1074/jbc.M706169200 (2007).
- 93 Lamartine, J. *et al.* Molecular cloning and mapping of a human cDNA (PA2G4) that encodes a protein highly homologous to the mouse cell cycle protein p38-2G4. *Cytogenet Cell Genet* **78**, 31-35 (1997).
- 94 Clausen, T., Southan, C. & Ehrmann, M. The HtrA family of proteases: implications for protein composition and cell fate. *Mol Cell* **10**, 443-455, doi:S1097276502006585 [pii] (2002).
- 95 Nie, G. Y. *et al.* A novel serine protease of the mammalian HtrA family is up-regulated in mouse uterus coinciding with placentation. *Mol Hum Reprod* **9**, 279-290 (2003).
- 96 Jensen-Seaman, M. I. & Li, W. H. Evolution of the hominoid semenogelin genes, the major proteins of ejaculated semen. *J Mol Evol* **57**, 261-270, doi:10.1007/s00239-003-2474-x (2003).
- 97 Molinari, F. *et al.* Truncating neurotrypsin mutation in autosomal recessive nonsyndromic mental retardation. *Science* **298**, 1779-1781, doi:10.1126/science.1076521 298/5599/1779 [pii] (2002).
- 98 Didelot, G. *et al.* Tequila, a neurotrypsin ortholog, regulates long-term memory formation in *Drosophila*. *Science* **313**, 851-853, doi:313/5788/851 [pii] 10.1126/science.1127215 (2006).
- 99 Wang, Y. *et al.* Mesotrypsin, a brain trypsin, activates selectively proteinase-activated receptor-1, but not proteinase-activated receptor-2, in rat astrocytes. *J Neurochem* **99**, 759-769, doi:JNC4105 [pii] 10.1111/j.1471-4159.2006.04105.x (2006).
- 100 Gafni, J. *et al.* Inhibition of calpain cleavage of huntingtin reduces toxicity: accumulation of calpain/caspase fragments in the nucleus. *J Biol Chem* **279**, 20211-20220, doi:10.1074/jbc.M401267200 M401267200 [pii] (2004).
- 101 Reseland, J. E. *et al.* A novel human chymotrypsin-like digestive enzyme. *J Biol Chem* **272**, 8099-8104 (1997).
- 102 Morel, S. *et al.* Processing of some antigens by the standard proteasome but not by the immunoproteasome results in poor presentation by dendritic cells. *Immunity* **12**, 107-117, doi:S1074-7613(00)80163-6 [pii] (2000).
- 103 Dong, Y. *et al.* Regulation of BRCC, a holoenzyme complex containing BRCA1 and BRCA2, by a signalosome-like subunit and its role in DNA repair. *Mol Cell* **12**, 1087-1099, doi:S1097276503004246 [pii] (2003).
- 104 Pfutzer, R. *et al.* Novel cationic trypsinogen (PRSS1) N29T and R122C mutations cause autosomal dominant hereditary pancreatitis. *Gut* **50**, 271-272 (2002).

- 105 Quesada, V. *et al.* Cloning and enzymatic analysis of 22 novel human ubiquitin-specific proteases. *Biochem Biophys Res Commun* **314**, 54-62, doi:S0006291X03026457 [pii] (2004).
- 106 Ganz, T. Defensins: antimicrobial peptides of innate immunity. *Nat Rev Immunol* **3**, 710-720, doi:10.1038/nri1180 nri1180 [pii] (2003).
- 107 Bevins, C. L., Jones, D. E., Dutra, A., Schaffzin, J. & Muenke, M. Human enteric defensin genes: chromosomal map position and a model for possible evolutionary relationships. *Genomics* **31**, 95-106, doi:S0888-7543(96)90014-2 [pii] 10.1006/geno.1996.0014 (1996).
- 108 Patil, A., Hughes, A. L. & Zhang, G. Rapid evolution and diversification of mammalian alpha-defensins as revealed by comparative analysis of rodent and primate genes. *Physiol Genomics* **20**, 1-11, doi:00150.2004 [pii] 10.1152/physiolgenomics.00150.2004 (2004).
- 109 Bastian, A. & Schafer, H. Human alpha-defensin 1 (HNP-1) inhibits adenoviral infection in vitro. *Regul Pept* **101**, 157-161, doi:S0167011501002828 [pii] (2001).
- 110 Daher, K. A., Selsted, M. E. & Lehrer, R. I. Direct inactivation of viruses by human granulocyte defensins. *J Virol* **60**, 1068-1074 (1986).
- 111 Sinha, S., Cheshenko, N., Lehrer, R. I. & Herold, B. C. NP-1, a rabbit alpha-defensin, prevents the entry and intercellular spread of herpes simplex virus type 2. *Antimicrob Agents Chemother* **47**, 494-500 (2003).
- 112 Zhang, G. H. & Zheng, Y. T. [Anti-HIV-1 effect of compound K3 from flower of Japanese pagoda tree in vitro]. *Zhong Yao Cai* **29**, 355-358 (2006).
- 113 Vinar, T., Brejova, B., Song, G. & Siepel, A. in *Comparative Genomics, International Workshop (RECOMB-CG)* Vol. 5817 150-163 (Springer, Budapest, Hungary, 2009).
- 114 Ouellette, A. J. & Bevins, C. L. Paneth cell defensins and innate immunity of the small bowel. *Inflamm Bowel Dis* **7**, 43-50 (2001).
- 115 Nguyen, T. X., Cole, A. M. & Lehrer, R. I. Evolution of primate theta-defensins: a serpentine path to a sweet tooth. *Peptides* **24**, 1647-1654, doi:10.1016/j.peptides.2003.07.023 S0196978103003395 [pii] (2003).
- 116 Cole, A. M. *et al.* Retrocyclin: a primate peptide that protects cells from infection by T- and M-tropic strains of HIV-1. *Proc Natl Acad Sci U S A* **99**, 1813-1818, doi:10.1073/pnas.052706399 99/4/1813 [pii] (2002).
- 117 Linzmeier, R. M. & Ganz, T. Human defensin gene copy number polymorphisms: comprehensive analysis of independent variation in alpha- and beta-defensin regions at 8p22-p23. *Genomics* **86**, 423-430, doi:S0888-7543(05)00157-6 [pii] 10.1016/j.ygeno.2005.06.003 (2005).
- 118 Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966-1967, doi:btp336 [pii] 10.1093/bioinformatics/btp336 (2009).
- 119 Kosiol, C. *et al.* Patterns of positive selection in six Mammalian genomes. *PLoS Genet* **4**, e1000144, doi:10.1371/journal.pgen.1000144 (2008).

- 120 Nielsen, R. & Yang, Z. Likelihood models for detecting positively selected amino
acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929-936
(1998).
- 121 Yang, Z. & Nielsen, R. Codon-substitution models for detecting molecular
adaptation at individual sites along specific lineages. *Mol Biol Evol* **19**, 908-917
(2002).
- 122 Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an improved branch-site
likelihood method for detecting positive selection at the molecular level. *Mol Biol
Evol* **22**, 2472-2479, doi:msi237 [pii]
10.1093/molbev/msi237 (2005).
- 123 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and
powerful approach to multiple testing. *J R Stat Soc B*, 289-300 (1995).
- 124 Thomas, P. D. *et al.* PANTHER: a browsable database of gene products
organized by biological function, using curated protein family and subfamily
classification. *Nucleic Acids Res* **31**, 334-341 (2003).
- 125 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene
Ontology Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).
- 126 Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian
Journal of Statistics* **6**, 65-70 (1979).
- 127 Makino, C. L. *et al.* Recoverin regulates light-dependent phosphodiesterase
activity in retinal rods. *J Gen Physiol* **123**, 729-741, doi:10.1085/jgp.200308994
jgp.200308994 [pii] (2004).
- 128 Kirk, E. C. Comparative morphology of the eye in primates. *Anat Rec A Discov
Mol Cell Evol Biol* **281**, 1095-1103, doi:10.1002/ar.a.20115 (2004).
- 129 Jacobs, G. H., Neitz, M., Deegan, J. F. & Neitz, J. Trichromatic colour vision in
New World monkeys. *Nature* **382**, 156-158, doi:10.1038/382156a0 (1996).
- 130 Lucas, P. W. *et al.* Evolution and function of routine trichromatic vision in
primates. *Evolution* **57**, 2636-2643 (2003).
- 131 Boggs, J. M., Gao, W. & Hirahara, Y. Myelin glycosphingolipids,
galactosylceramide and sulfatide, participate in carbohydrate-carbohydrate
interactions between apposed membranes and may form glycosynapses
between oligodendrocyte and/or myelin membranes. *Biochim Biophys Acta* **1780**,
445-455, doi:S0304-4165(07)00260-7 [pii]
10.1016/j.bbagen.2007.10.015 (2008).
- 132 Juneja, S. C. Development of infertility at young adult age in a mouse model of
human Sandhoff disease. *Reprod Fertil Dev* **14**, 407-412, doi:RD02060 [pii]
(2002).
- 133 Sandhoff, R. *et al.* Novel class of glycosphingolipids involved in male fertility. *J
Biol Chem* **280**, 27310-27318, doi:M502775200 [pii]
10.1074/jbc.M502775200 (2005).
- 134 Kato, K. *et al.* Plasma-membrane-associated sialidase (NEU3) differentially
regulates integrin-mediated cell proliferation through laminin- and fibronectin-
derived signalling. *Biochem J* **394**, 647-656, doi:BJ20050737 [pii]
10.1042/BJ20050737 (2006).

- 135 Miyagi, T., Wada, T. & Yamaguchi, K. Roles of plasma membrane-associated sialidase NEU3 in human cancers. *Biochim Biophys Acta* **1780**, 532-537, doi:S0304-4165(07)00222-X [pii]
10.1016/j.bbagen.2007.09.016 (2008).
- 136 Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41**, 1061-1067, doi:ng.437 [pii]
10.1038/ng.437 (2009).
- 137 Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959 (2000).
- 138 Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567-1587 (2003).
- 139 Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36**, e105, doi:gkn425 [pii]
10.1093/nar/gkn425 (2008).
- 140 Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L. & Welch, D. M. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* **8**, R143, doi:gb-2007-8-7-r143 [pii]
10.1186/gb-2007-8-7-r143 (2007).
- 141 Crow, J. F., Kimura M. *An Introduction to Population Genetics Theory*. (Harper and Row, 1970).
- 142 Ewens, W. J. *Mathematical Population Genetics*. (Springer, 1979).
- 143 Rhead, B. *et al.* The UCSC Genome Browser database: update 2010. *Nucleic Acids Res* **38**, D613-619, doi:gkp939 [pii]
10.1093/nar/gkp939 (2010).
- 144 Siepel, A. & Haussler, D. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol* **21**, 468-488, doi:10.1093/molbev/msh039
msh039 [pii] (2004).
- 145 Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110-121, doi:gr.097857.109 [pii]
10.1101/gr.097857.109 (2010).