Supplementary Information

Supplemental Section S1 – Genome sequencing and assembly

1.1 Genome sequencing

The *Nomascus leucogenys* whole genome shotgun (WGS) project is available from Genbank under the project accession ADFV00000000.1. Sequencing using Sanger methods was performed at The Genome Institute, Washington University School of Medicine (St. Louis, MO) and Human Genome Center, and Baylor College of Medicine (Houston, TX). DNA for both the BAC library and the WGS sequence was isolated from blood provided by Alan Mootnick (former director, Gibbon Conservation Center, Santa Clarita, California) from a female named Asia (international studbook #0098, ISIS # NLL605) housed at the Virginia Zoo in Norfolk. The CHORI-271 BAC Library (https://bacpac.chori.org/library.php?id=228) was constructed by Mr. Boudewijn ten Hallers and Dr. Baoli Zhu in Pieter deJong's laboratory (BACPAC Resources, Children's Hospital Oakland Research Institute).

1.2 Genome assembly (Nleu1.0)

The genomic sequence was first assembled by Sante Gnerre (Broad Institute of Harvard and MIT) using the ARACHNE genome assembler assisted with alignment data from the Human genome (NCBI build 35, UCSC hg18) using previously described methods[1]. Such assembly was called Nleu1.0. The mitochondrion was assembled by Yue Liu (BCM-HGSC) using Phrap (http://www.phrap.org); reads with similarity to mitochondrial sequences were assembled using Phrap with default parameters. The genome was sequenced to a depth of 5.7X in Q20 bases with the number of reads and coverage for each read type as seen in Table ST1.1 and the final assembly statistics shown in Table 1 in the main text. This assembly, named Nleu1.0/nomLeu1 was used for all the main analyses described in the genome paper.

Library	*No. of reads*	*trimmed read length mean ± s.d.*	*genomic coverage by trimmed read bases*	*genomic coverage by trimmed read bases of qual ≥20*
BCM-2266645	2,660,257	809±106	0.72X	0.67X
BCM-62020	3,130,851	815±100	0.85X	0.79X
BCM-938188323	1,080	956±124	0.00X	0.00X
BCM-GXZMP	2,355,108	810±115	0.64X	0.58X
BCM-MBTLP	1,372,914	817±99	0.37X	0.35X
WUGSC-BAC 173	289,746	708±165	0.07X	0.06X
WUGSC-CH271	38,096	617±197	0.01X	0.01X
WUGSC-Fosmids 40	2,189,628	634±192	0.46X	0.36X
WUGSC-Plasmids 1.75	24,430	601±238	0.00X	0.00X
WUGSC-Plasmids 3	340,938	631±207	0.07X	0.06X
WUGSC-Plasmids 4.25	15,245,577	639±177	3.25X	2.79X
Total	**27,648,625**	**699±178**	**6.44X**	**5.68X**

Library	*fraction paired (%)*	*fraction assembled (%)*	*fraction assembled with partner (%)*
BCM-2266645	4.2	93.6	3.8
BCM-62020	11.6	93.5	10.6
BCM-938188323	0.0	92.1	0
BCM-GXZMP	4.0	93.4	3.7
BCM-MBTLP	2.1	93.8	1.9
WUGSC-BAC 173	88.5	92.3	78.6
WUGSC-CH271	89.1	89.8	75.1
WUGSC-Fosmids 40	83.8	84.4	66.5
WUGSC-Plasmids 1.75	78.9	86.6	71.6
WUGSC-Plasmids 3	82.4	91.3	74.6
WUGSC-Plasmids 4.25	92.9	87.7	76.8
Total	**62.2**	**89.5**	**51.5**

Table ST1.1 Input Read Statistics for the gibbon assembly.

1.3 Creation of chromosomal "A Golden Path" (AGP) files

The assembly data were aligned against the human genome at UCSC utilizing BLASTZ[2] and non-repetitive gibbon regions were scored against the repeat-masked human sequence (GRCh37). Alignment chains were differentiated between orthologous and paralogous alignments[3] and only "reciprocal best" alignments were retained in the alignment set. The gibbon "A Golden Path" (AGP) files were generated from these alignments in a manner similar to that already described[4]. Documented

inversions based on FISH data (http://www.biologia.uniba.it/gibbon) and inversions suggested by the assembly and supported by additional mapping data (e.g. fosmid and BAC-end pairs) were also introduced. There were 81 human/gibbon breakpoints (not including centromeres) as defined by the initial maps (Rocchi, personal communication). Based on the assembly data and resulting re-examination of the FISH data we added two additional breaks, one within the very complex and large pericentromeric region of human chromosome 9 and the other, a smaller inversion near the telomere of human chromosome 17. Thus the final maps suggested 83 major (large enough to be detected by FISH) breakpoints and, of those, 64 were spanned by assembly scaffolds. Additional breakpoints were identified through manual inspection of the alignments, obtaining a final list of 96 breakpoints. A list of the human-gibbon synteny breakpoints used for the analyses reported in the main text is included in Supplementary File 3. In the cases in which the alignment with GRChr37 had large gaps, we listed the left and right sides of the breakpoints separately and labeled each side 'a' and 'b'. Additional smaller rearrangements are found fully contained within the assembly scaffolds. Centromeres were placed based on their identified positions from cytogenetic data (Rocchi, personal communication). In the final chromosomal AGP files there are 2.79 Gbp of sequence ordered and oriented along the 26 gibbon chromosomes, 52 Mbp placed on the *_random segments associated with those gibbon chromosomes, and an additional 114 Mbp unplaced.

1.4 Assembly quality assessment based on single-copy genes

Introduction

To assess both the quality of the assembly and the completeness of the annotation, we have devised a new strategy to compare the annotation of single-copy genes in mammalian genomes. The rationale is that many assembly errors like gaps or inversions will affect our ability to correctly predict the genes in the genome. Our approach is similar to CEGMA[5] in that we look for the presence of a core set of proteins. In our case, we restrict the analysis to mammalian species and focus on single-copy genes. In

summary, we define a set of genes that are expected to be in single-copy in most mammalian species and test whether we find them in the set of gene models that we could predict in the gibbon genome.

Methods

Starting from the Ensembl GeneTrees (Beal et al, in preparation) in release 67 (http://e67.ensembl.org), we extracted a set of genes present once and only once in a set of 11 high-quality genomes, namely human, chimpanzee, gorilla, orangutan, rhesus macaque, marmoset, mouse, rat, cow, dog and horse. To accommodate assembly and annotation artifacts in this initial set of genomes, we tolerate up to two duplications or losses among these species. These sets of genes are defined as a sub-family in the tree that corresponds to a placental mammal speciation event. In practical terms, the trees are traversed from the root up to all the branches until an eutherian speciation node is found. The corresponding sub-family is tested and kept if it represents a single-copy gene as described earlier. It is possible to extract more than one sub-family from the same tree in this process.

For each selected sub-family, we extract the alignment from Ensembl and build a profile Hidden Markov Model (HHM) using HMMer 3.0[6]. For this stage we consider the sequences from every eutherian genome to obtain a more representative HMM model. The discriminative power of the HMM is then tested. We record the worst score of every protein in the original sub-family and the best score of any other protein. To account for partial genes in the original sub-family, we disregard low scores from partial proteins. These are defined as sequences that cover less that 80% of the HMM profile. We only consider HMMs if all the proteins in the original sub-family match the HMM with an e-value smaller than 1e-100 and these e-values are at least 10 orders of magnitude smaller than those of any other gene. These HMMs are considered a good representative of their sub-family and are used to assess the genomes.

Results

Using the strategy described above, we have defined a set of 14,319 sub-families. Out of these, 3,485 were discarded because the corresponding HMM did not match the required criteria. The final set of 10,734 HMMs was used to assess the completeness of the gibbon genome. To help interpreting the

results, the same analysis was done on all mammalian genomes available in Ensembl 70 (http://e70.ensembl.org) (Extended Data Fig. 1). With this methodology, the best annotated genomes appear to be human and mouse. This is not surprising as these are the two genomes that are being under constant improvement, both at the assembly level by the Genome Reference Consortium[7] and at the annotation level by manual curation[8].

The gibbon genome scores very similarly to orangutan and rhesus macaque. Of note, the approach described here only focuses on single-copy genes. While these represent a large portion of the genome, segmental duplications and repeats are ignored in this analysis. These are typically much more difficult to resolve in the assembly process. It is therefore advisable to be cautious when extrapolating these results to the whole genome. On the other hand, these results allow us to make a comparative assessment of the single-copy gene annotation and give us a broad idea of the relative quality of these assemblies. It also confirms that the gibbon genome assembly is within the range of what are commonly accepted as high-quality genomes.

1.5 Comparison of gibbon BAC sequences to the gibbon assembly

In order to evaluate the general quality of the gibbon assembly (both versions Nleu1.0 and Nleu3.0), we considered the sequence of 242 fully sequenced and previously published gibbon BACs[9,10]. First, to determine the correct location of these BACs compared to each assembly, we aligned the BAC sequences to the unmasked assembly using MEGABLAST (version 2.2.19, parameters -D 2 -v 5 -b 5 -e 1e-70 -p 89 -s 220 -W 12 -t 21 -F F). We took into account alignments with more than 94% of identity and greater than 1 Kbp for the identification of contiguous blocks (>30Kbps) in the assembly, formed by alignments that are less than 65 Kbp apart. Each BAC and its resulting blocks in the assembly were then realigned using the bl2seq tool of BLAST (version 2.2.19, parameters -p blastn -m T -W 12 -t 21 -F F -e 1.e-70). Finally, we kept only alignments higher than 98% of identity for our analysis.

Out of the 242 BACs, we identified 218 in Nleu1.0 and 219 in Nleu3.0 as a unique block in the assembly (Table ST1.2). Blocks of the remaining BACs that were not univocally determined are shown

in Supplementary Files 1-2. It is remarkable that only 3 BACs represent strong discrepancies between the two assemblies as they are anchored into the assembly as unique in one version and are fragmented in two different chromosomes or contigs in the other version (one example is shown in Fig. SF1.1).

BACs identified in assembly:	Nleu1.0	Nleu3.0	In both
All BACs	242	242	242
BACs with a unique block	218	219	217
BACs without translocations	167	167	166
BACs with translocations	51	52	51
BACs with inverted sequences	20	20	20

Table ST1.2 Summary statistics of anchored BACs into both gibbon assemblies. [Note: all BACs with inverted sequences are included in the BACs with translocations].

To obtain a conservative quantitative statistics about the correspondence of BACs in the assembly, we considered only those BACs having been identified as a unique block in the assembly (218 in Nleu1.0 and 219 in Nleu3.0). Alignments in the blocks were reduced to non-overlapping alignments. If two alignments overlap in the BAC sequence, then the overlap region was assigned only to the longer alignment and alignments completely included in larger alignments either in the assembly or in the BAC were removed; we treated in a similar way the overlaps in the assembly. Final blocks of non-overlapping alignments for each of these univocally identified BACs in Nleu1.0 and Nleu3.0 are plotted in Supplementary Files 2. Moreover, a complete description of the mapping in each assembly per each of the 218 and 219 BACs anchored uniquely in the respective assembly is shown in Supplementary File 1, separated into two sets, the ones not presenting any translocations and the rest.

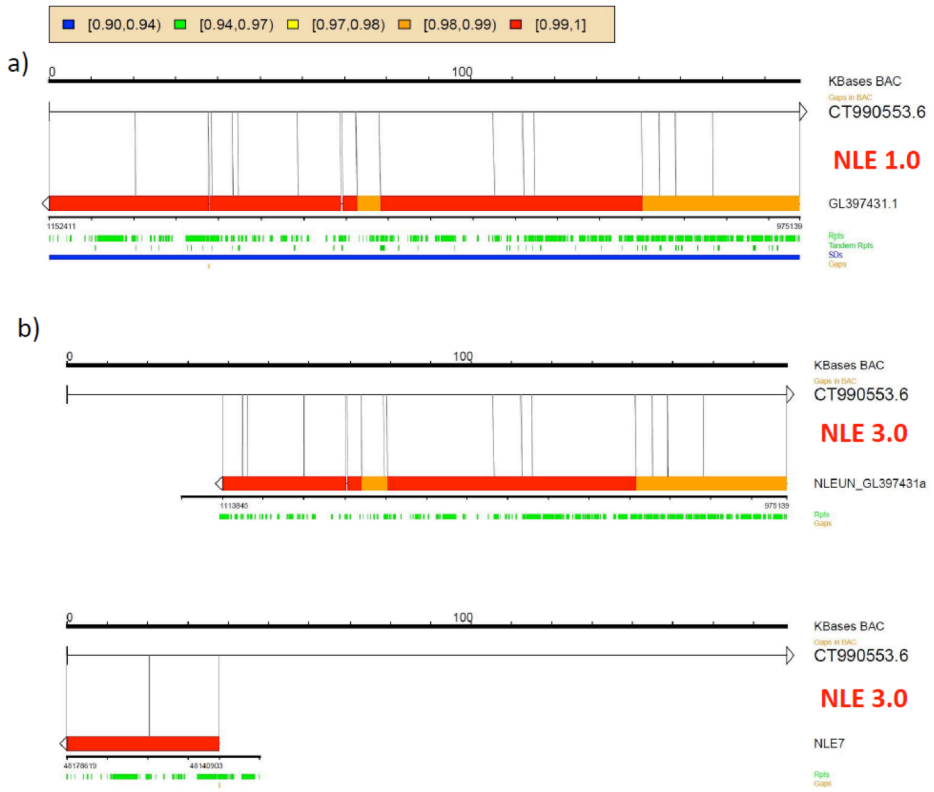


Figure SF1.1 BAC CT990553.6 aligned against a) Nleu1.0 and b) Nleu3.0 assemblies. BAC and assembly are plotted; lines between them in their respective coordinates mark the ends of the alignments. Alignments are colored depending on their similarity and the strand is shown with an arrow. Gaps are shown in orange. The BAC aligns to a contiguous block in Nleu1.0 while in Nleu3.0 most of it has been included in the unknown chromosome.

Considering the 166 BACs that are uniquely anchored to the assembly without any translocation in both versions of the gibbon assembly, the total amount of sequence evaluated is 28.70 Mbps (Table ST1.3). In both versions, Nleu1.0 and Nleu3.0, results are almost identical. Overall, there are 27.48 Mbps (96% of the BAC sequence excluding gaps) that correlate in both BAC and assembly sequences on intervals with high similarity (>99%, except for only 0.68% of the aligned sequence that has >98% and <99% of identity; notice that small indels are allowed). Most of the intervals that don't properly align correspond to gaps, from both the assembly and the BACs (Supplementary File 2). The non-aligned BAC sequence comprises 1.22 Mbp, of which 40.9% are in intervals where there is a gap in the BAC, being their counterpart in the assembly also plenty of gaps (75.50% of the corresponding sequence in the

assembly belong to intervals that contain gaps). Moreover, the remaining non-aligned intervals that do not have a gap in the BAC are mainly regions of gaps in the assembly (83% of 1.23 Mbp).

Finally, we considered 51 BACs that present at least one region that is translocated relative to the BAC sequence. Only one BAC, AC202649.3, is not included in this set of BACs for both assemblies. The 51 BACs comprise 9.26 Mbp, and on average 88.86% of the ungapped sequence of a BAC has a highly similar sequence in the corresponding block in the assembly. In these BACs, 6.64% of the ungapped sequence of the BAC is inverted in the assembly. Results are almost identical again for both versions of the gibbon assembly (Table ST1.4).

	BACs							
	Length	Gaps	Aligned bps in BACs			Percentage aligned of ungapped bps		
			Total	>99% id	>98% id <99%	Total	>99% id	>98% id <99%
Nleu1	28,702,239	64,352	27,480,862	27,285,300	195,562	95.960	95.277	0.683
Nleu3	28,702,239	64,352	27,480,850	27,285,293	195,557	95.960	95.277	0.683
	Assembly							
	Length	Gaps	Aligned bps in Assembly			Percentage aligned of ungapped bps		
			Total	>99% id	>98% id <99%	Total	>99% id	>98% id <99%
Nleu1	29,532,083	1,397,388	27,473,364	27,278,338	195,026	97.649	96.956	0.693
Nleu3	29,532,083	1,397,388	27,473,365	27,278,344	195,021	97.649	96.956	0.693
	Not Aligned bps							
	Total		gaps in BACs			no gaps in BACs		
	in BACs	in Assembly	BACs	Assembly	gaps in Assembly	BACs	Assembly	gaps in Assembly
Nleu1	1,221,377	2,058,719	499,271	832,336	628,081	722,106	1,226,383	1,024,503
Nleu3	1,221,389	2,058,718	499,268	832,333	628,078	722,121	1,226,385	1,024,507

Table ST1.3 Number of bps mapped and not mapped in 166 uniquely anchored BACs that do not show any translocation for both assemblies. The assembly regions are defined by the left and right most positions of the alignments with the BACs. Percentage mapped is computed as the total bps mapped divided by the total sequence minus the length of the gaps. The non-mapped BAC sequence was separated in contiguous regions including a BAC gap, and the rest. Then, we counted how many bps correspond to these intervals in the assembly, and how many belong to assembly gaps.

	BACs					
	Length	Gaps	Aligned bps in BACs			
			Total	>99% id	>98% id <99%	Inverted bps
Nleu1	9,264,048	37,180	8199297 (88.86%)	7990383 (86.60%)	208914 (2.26%)	612758 (6.64%)
Nleu3	9,264,048	37,180	8199271 (88.86%)	7986936 (86.56%)	212335 (2.30%)	612761 (6.64%)
	Assembly					
	Length	Gaps	Aligned bps in BACs			
			Total	>99% id	>98% id <99%	Inverted bps
Nleu1	9,604,800	781,320	8197062 (92.90%)	7988771 (83.17%)	208291 (2.17%)	612771 (6.94%)
Nleu3	9,604,800	781,320	8197071 (92.90%)	7988771 (83.17%)	211720 (2.20%)	612774 (6.94%)

Table ST1.4 Number of bps mapped and not mapped in 51 uniquely anchored BACs with at least one translocation in both assemblies. The assembly regions are defined by the left and right most positions of the alignments with the BACs. Percentage mapped (shown in parenthesis) is computed as the total bps mapped divided by the total sequence minus the length of the gaps. Inverted bps correspond to alignments that do not followed same orientation that the majority of alignments in the block.

In conclusion, the analysis results are nearly identical for both versions of the gibbon assembly. Overall, 96% of the 166 BAC sequences uniquely identified in both assemblies and not showing any translocation are well represented in the assembly, meaning that these regions align in the assembly contiguously and with more than 99% of similarity (except for 0.68% of the aligned sequence that has >98%) while small indels are allowed. The main source of error in the rest of the sequence derives from existing gaps on the assembly. On the remaining 51 BACs that map uniquely in the assembly and show at least a region translocated relative to the BAC sequence, the percentage of ungapped BAC sequence present on the assembly is 88.87%. Despite of the good correspondence between BAC and assembly sequences, integrating the BAC sequences to the assembly will still improve a fraction of regions in the assembly.

1.6 Comparison to finished BACs to assess substitution and indel error rates

We aligned 98 completely finished BACs totaling 16.69 Mb from the CHORI271 gibbon library to the gibbon assembly (Nleu1.0) using blastn[54] to identify the scaffolds representing those BACs. After potential scaffolds were identified, they were realigned with cross_match (P. Green, unpublished) using the following parameters: gap_init: -4, gap_ext: -3, ins_gap_ext: -3, del_gap_ext: -3, minmatch: 14,

maxmatch: 20, max_group_size: 0 (turned off), minscore: 30, near_minscore: 30, bandwidth: 14, indexwordsize: 12, indexwordsize2: 4, gap1_minscore: 17, gap1_dropoff:-12, minmargin: 0.5 and a score matrix set by a value of penalty -2 with the -tags and -discrep_lists options for ease of parsing. Alignments were retained for the best scoring alignment along the BAC. Because cross_match considers the phred base qualities (Ewing, Green, 1998), estimates of the substitutions and indel error rates could be confined to the high quality bases of the assembly (bases with a quality value ≥40, corresponding to an error rate of $\leq 10^{-4}$). This analysis revealed a high quality discrepancy rate of 11×10^{-4} substitutions and an indel error rate of $< 3 \times 10^{-4}$ in hiqh quality bases. When taking into account the estimated heterozygosity, the discrepancy rate is reasonably similar to that found in finished sequence confirming that this draft assembly provides an excellent resource for the analyses presented here.

1.7 Assessing large-scale rearrangements in the gibbon genome

We wanted to compare the level of synteny between the human genome and the genomes of other primates (great apes, gibbon, old world monkeys, new world monkeys) in order to test the possibility that the gibbon genome bears substantially more large-scale rearrangements than the other species, while maintaining a similar number of smaller scale rearrangements. In order to do so, Repeat-masked genomic sequences for human, chimpanzee, gorilla, orangutan, macaque and marmoset were downloaded from Ensembl (release 73, ftp://ftp.ensembl.org/pub/release-73/fasta/)[11]. In addition, the repeat-masked sequence for the gibbon genome (Nleu3.0) was obtained from the UCSC Genome Browser (http://hgdownload.soe.ucsc.edu/goldenPath/nomLeu3/bigZips/)[12]. The gibbon and human genomes were compared with LASTZ (http://www.bx.psu.edu/miller_lab/dist/README.lastz-1.02.00/README.lastz-1.02.00a.html) by aligning each gibbon chromosome to each human one. Very stringent parameters (--step=30 --seed=match12 --exact=50 --matchcount=1000 --masking=3) are used to obtain only the most relevant alignments. These alignments are further summarized in chains using axtChain[3]. These represent groups of alignments in the same order and orientation. In Fig. 2-A in the main text, the resulting chains were split if they contain any gap longer than 10 Mbp. Any chain

shorter than 10 Mbp was ignored. Each chain represents a collinear block between the gibbon and the human genome. The same approach was used for the other primates. The alignments, the chains and the parsing is performed with an eHive[13] pipeline (available on request).

In parallel, the chains were analyzed with the chainNet software to find rearrangements at different scales (Fig.2-B). Each net represents a region of co-linearity between both genomes. Importantly these nets can be nested, such that a local rearrangement within a long co-linear block will be noted as one additional net, but will not break the longer net. We used the following thresholds to study the rearrangements at different scales: 10 Kbp, 20 Kbp, 30 Kbp, 40 Kbp, 50 Kbp, 60 Kbp, 70 Kbp, 80 Kbp, 90 Kbp, 100 Kbp, 150 Kbp, 200 Kbp, 300 Kbp, 400 Kbp, 500 Kbp, 600 Kbp, 700 Kbp, 800 Kbp, 900 Kbp, 1 Mbp, 1.5 Mbp, 2 Mbp, 3 Mbp, 4 Mbp, 5 Mbp and 10 Mbp. Fig. 2-B in the main text shows the total number of nets at each threshold for each species, independently of whether they are nested or not.

Supplemental Section S2 – Next-generation sequencing datasets

2.1 The diversity panel: whole-genome sequences

In order to examine diversity at the whole genome level we performed next-generation sequencing on two individuals (one male and one female) from each of the four genera (Table ST2.1).

Genus	Species	Species (common name)	Code	Name	ISIS #	Sex	Origin
Nomascus (2n=52)	*N. leucogenys leucogenys*	Northen white-cheeked gibbon	**NLE**	Vok	NLL600	M	CB, parents WB
				Asteriks	NLL607	F	CB, parents WB
Hylobates (2n=44)	*H. moloch*	Javan gibbon	**HMO**	Madena	HMO892	M	sire WB, dam CB
	H. pileatus	Pileated gibbon	**HPI**	Domino	HP120	M	CB, parents WB
Symphalangus (2n=50)	*S. syndactylus*	Siamang	**SSY**	Karenina	SS901	F	sire WB, dam CB
				Monty*	SS910	M	CB, parents WB
Bunopithecus (2n=38)	*Hoolock leuconedys*	Eastern hoolock	**HLE**	Maung	HH308	M	WB
				Drew	HL307	F	WB

Table ST2.1 Description of the gibbon samples used for whole-genome sequencing. CB = captive born; WB = wild born; F = female; M = male (*deceased

All the blood samples were received from the gibbon conservation center and were obtained during routine veterinarian check-up visits. Blood and tissues were obtained in agreement with protocols reviewed and approved by the Gibbon Conservation Center. High molecular weight DNA was extracted from blood using the Gentra Puregene kit (Qiagen). About 1 μg of DNA from each sample was individually fragmented by sonication (Bioruptor, Diagenode) and Illumina libraries were prepared in agreement with manufacturer's instructions. Libraries were sequenced on the HiSeq 2000 platform, generating 2x100 bp reads. Four different sequencing centers contributed to the sequencing of these samples: the Oregon Health & Science University Massively Parallel Sequencing Shared Resource (MPSSR) (Portland, OR), National Center of Genomic Analyses (CNAG) (Barcelona, Spain), University of Arizona Genetics Core (UAGC) (Tucson, AZ), and the UCSF sequencing core (San Francisco, CA). Multiple runs were performed to generate a mean coverage ranging from 11.5X to 19.5X. Details about number of reads and coverage are summarized on Table ST2.2. All reads have been submitted to the NCBI Short Read Archive (http://www.ncbi.nlm.nih.gov/sites/sra).

Code	Name	Raw data (Gb)	unniquely mapped reads	Mean Coverage
NLE	Vok	499,719,188	414,235,422	13.78
	Asteriks	416,267,412	348,252,851	11.50
HMO	Madena	445,595,278	387,402,657	12.96
HPI	Domino	485,796,436	432,480,633	14.33
SSY	Karenina	783,123,982	587,916,397	19.53
	Monty	518,935,362	382,427,125	12.80
HLE	Maung	783,942,998	596,742,872	19.15
	Drew	738,381,714	451,702,370	14.36

Table ST2.2 Next-generation sequencing data summary

2.2 Exome sequencing

Exome capture using the TruSeq Exome Enrichment Kit (Illumina) was performed on one NLE sample (Vok, 116x coverage) and one SSY sample (Monty, 64x coverage) by the University of Arizona Genetics Core (UAGC) (Table ST2.3). All reads have been submitted to the NCBI Short Read Archive.

Code	Name	reads for alignment	PCR duplicates	% duplicate	unniquely mapped reads	Mean Coverage
NLE	Vok	190,619,344	109,225,969	57.30	80,027,758	116.00
SSY	Monty	140,806,954	93,850,771	66.65	44,796,909	63.81

Table ST2.3 Exome sequencing data summary

2.3 RNA sequencing

Total RNA was extracted from an EBV-transformed lymphoblastoid cell line established for the individual used for the reference genome (Asia). Both polyA and directional RNA-seq were performed by the CNAG. .All reads have been submitted to the NCBI Short Read Archive.

Directional mRNA sequencing library preparation

RNA quality was assessed using a Nanodrop 2000c (Thermo Scientific) and a 2100 Bioanalyzer (Agilent Technologies, CA, USA). The library was prepared using the ScriptSeqTM Complete Gold Kit (Human/Mouse/Rat) (Epicentre Biotechnologies, WI, USA, #BG1224), according to manufacturer's protocol. Briefly, 3 µg of total RNA was used for removal of both, cytoplasmic and mitochondrial rRNAs,

using the Ribo-Zero™ Gold rRNA Removal Reagents. The total rRNA depletion of the samples was confirmed by the 2100 Bioanalyzer RNA 6000 Pico Chip. Up to 50 ng of Ribo-Zero-treated RNA was used to perform the library preparation, using the ScriptSeqTM v2 RNA-Seq Library Preparation Kit. This includes random-priming, first-strand cDNA synthesis and incorporation of Illumina platform-specific 3' sequencing tag. The multiplexing index was added through 12 cycles of PCR were performed using the FailSafeTM PCR Enzyme Mix (Epicentre Biotechnologies, #FSE51100) followed by AMPure XP Purification (Agencourt, Beckman Coulter).

Illumina TruSeq™ RNA sequencing library preparation

The library was prepared using the TruSeq™ RNA sample preparation kit (Illumina Inc.) according to manufacturer's protocol. Briefly, 0.5 µg of total RNA was used for poly-A based mRNA enrichment selection using oligo-dT magnetic beads followed by fragmentation by divalent cations at elevated temperature resulting into fragments of 80-250 nt, with the major peak at 130 nt. First strand cDNA synthesis by random hexamers and reverse transcriptase was followed by the second strand cDNA synthesis performed using RNAseH and DNA Pol I. Double stranded cDNA was end repaired, 3´adenylated and the 3´-"T" nucleotide at the Illumina adaptor was used for the adaptor ligation. The ligation product was amplified with 15 cycles of PCR.

Sequencing

Both, the directional mRNA and the TruSeq™ RNA libraries, were sequenced using TruSeq™ SBS Kit v3-HS, in paired end mode, 2x76 bp, each in a fraction of a lane of a HiSeq sequencing system (Illumina, Inc) following the manufacturer's protocol, generating minimally 150 million paired end reads for each sample. Images from the instrument were processed using the manufacturer's software to generate FASTQ sequence files.

Protocol	Sequence Name	Raw reads	Read type
RNA-seq (polyA)	C0ET8ACXX 421-422	164,537,257	76 bp paired end
dirRNA-seq	C0ET8ACXX 411-412	46,513,623	76 bp paired end
	D0MD0ACXX 711-712	116,229,377	76 bp paired end

Table ST2.4 Summary of RNA-seq experiments

Supplemental Section S3 – Analysis of gibbon duplications

3.1 Segmental duplications in Nleu1.0 / nomLeu1

We applied two *in silico* methods to discover segmental duplications in the *Nomascus leucogenys* reference assembly (Nleu1.0). Whole-genome assembly comparison (WGAC) compares repeat-free sequence of the assembly to itself to identify duplicated sequences greater than 1 kbp and with higher than 90% identity[14]. Whole-genome shotgun sequence detection (WSSD) aligns whole-genome shotgun (WGS) reads to the assembly and identifies large, highly identical regions (>10 kbp, >94% identity) with a higher read depth compared to known unique regions[15]. Here, to match previous studies done on great-ape genomes, we modified this pipeline by increasing the size threshold to 20 kbp[16], filtering reads with >85% overlap with common repeats or >75% overlap with tandem repeats[9,16], and masking of all satellite and L1P repeats in the reference, as they would otherwise increase the number of false positives in duplication calls. Using these settings, we mapped 25,757,713 gibbon WGS reads from the female *Nomascus leucogenys* individual "Asia" to Nleu1.0. These reads were sequenced with Sanger technology and are the ones used to assemble the genome reference. We called duplications in 5 Kbp windows with a read depth >81 (threshold determined by the unique regions), >200 bp of unmasked sequence, and >200 bp of sequence with a Phred quality >30.

In the assembly-based analysis (WGAC) of Nleu1.0, we discovered 17,924 pairwise alignments corresponding to 6.98 Mbp (0.25% of the genome sequence excluding gaps) of non-redundant duplications (Table ST3.1). The majority of these (6.89 Mbp) were inter-contig duplications with only ~0.17 Mbp of intra-contig duplications (Fig. SF3.1-a). The distribution of segmental duplications by similarity was bimodal for inter-contig duplications with one mode near 92% identity and the other at 97.5% (Fig. SF3.1-b). This bimodal distribution is seen in the human reference sequence, however, typically the first mode belongs to interchromosomal duplications and the second mode to intrachromosomal duplications. Due to the relatively fragmented nature of the Nleu1.0 assembly (there are 2,916 scaffolds greater than 10 kbp, that comprise 2.86 Gbp (97.5% of the genome)), segmental

duplications from the same chromosome are likely to be split into separate contigs which would partially explain the higher ratio of inter-contig to intra-contig duplications we found.

A total of 41.58 Mbp (1.51% of the genome sequence excluding gaps) of duplicated sequences (>20 kbp, >94% identity) were detected with WSSD prior to copy number correction (Supplementary File 4) and 120.59 Mbp (4.4%) after copy number correction (Table ST3.1) a similar amount compared to the human genome assembly[15].

Category	
Total genome length	2.94 Gbp
Chrom length (genome without gaps)	2.76 Gbp
Number of WGAC pairs	17,924
Number of inter contig	17,770
Number of intra contig	154
nr length	6.98 Mbp
nr length of inter contig	6.89 Mbp
nr length of intra contig	0.17 Mbp
WSSD	41.58 Mbp
Copy-number corrected WSSD	120.59 Mbp

Table ST3.1 Summary of duplications in Nomascus leucogenys reference assembly by WGAC and WSSD analyses

By comparing the duplicated sequence identified by both methods, focusing on regions greater than 20 kbp and with higher identity than 94%, we can assess the quality of the reference in terms of

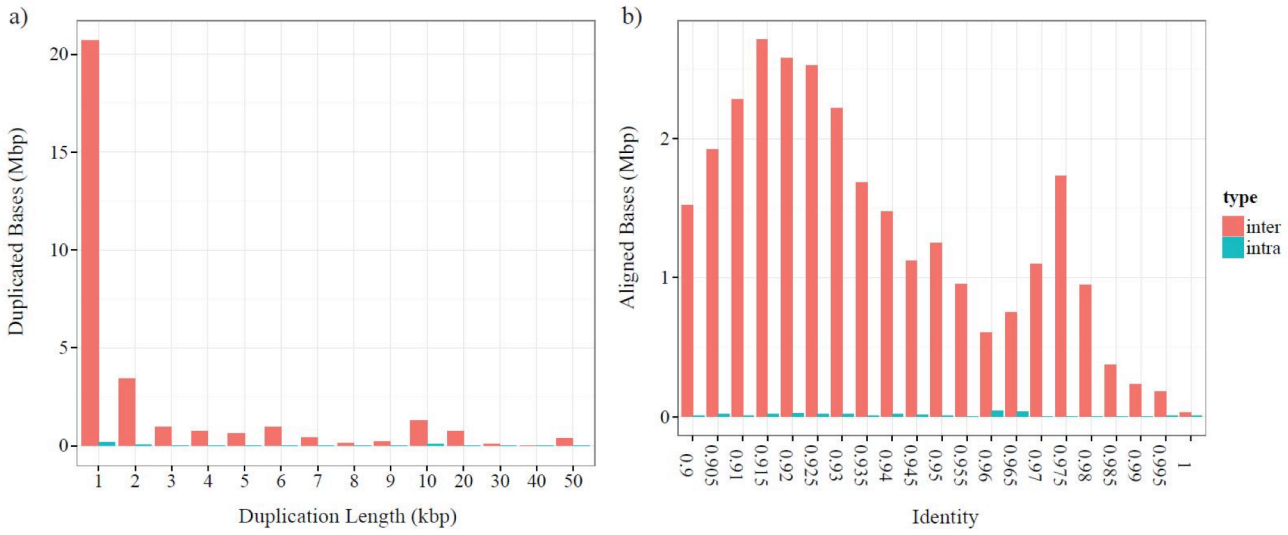


Figure SF3.1 a) Total number of base pairs for inter-contig and intra-contig WGAC duplications with sizes between 1 kbp and 50 kbp. b) Total number of aligned base pairs for intercontig and intracontig WGAC duplication events based on sequence identities between 0.9 and 1 in intervals of 0.005

duplication content: regions detected by WGAC but not by WSSD are potential artifact duplications while, regions determined by WSSD and not by WGAC are duplications potentially collapsed in the reference. Of the total 1.19Mbp detected by WGAC with regions greater than 20 kbp, 0.82 Mbp of the duplications are common to both methods of detection (69% of the WGAC duplications that are being compared), while 0.37 Mbp represent potential artifact duplications (only predicted by WGAC), and 41.21 Mbp of duplicated sequence is potentially collapsed in the assembly (only detected by WSSD) (Table ST3.2).

WGAC-identity	WGAC	WSSD	Shared	WGAC-only	WSSD-only
≥94%	1.19Mb	41.58Mbp	0.82Mbp	0.37Mbp	41.21Mbp

Table ST3.2 Comparison of duplications (>20Kbps) in Nleu1.0 predicted by WGAC and WSSD analyses.

Based on WSSD duplications and the latest Ensembl gene set for Nleu1.0 (e70), this draft genome contains 273 genes with 95% of their exons overlapping segmental duplications (Supplementary File 4) and 428 genes with at least one exon overlapping a duplication.

Validation of WGAC duplications by fluorescent in situ hybridization (FISH)

We selected 9 human fosmid clones to test duplicated regions bigger than 20 kbp identified by WGAC method (regions that were mapped to GRCh37 to determine their human concordant sequences). Fosmids were used as probes in FISH assays on slides with gibbon interphases and metaphase spreads and were co-hybridized with control probes (human fosmids selected on gibbon single-copy regions). Metaphases were obtained from gibbon lymphoblastoid cell lines. DNA extraction[17] and FISH experiments[18] were performed as previously described. For each slide, we observed at least 20 nuclei: when 95-98% of the nuclei showed more intense control probe signals, the region was assigned as duplicated. We observed multiple signals either by examination of interphase or metaphase FISH for all probes (9/9), confirming their duplication status (100% validation). Seven out of 9 probes (77.78%) showed signals on non-homologous chromosomes (interchromosomal duplications) while the remaining probes showed evidence of duplicated signals that were locally clustered (intrachromosomal duplication) (Table ST3.3).

Chr	Start	End	Length	Genes	Fosmid name	NLE (Nomascus Leucogenys)
chr1	148,190,450	148,221,841	31,392	PPIAL4B	ABC8_000041056000_C5	interchromosomal
chr3	41,916,871	41,939,388	22,518	ULK4	ABC8_000002114640_G4	interchromosomal
chr3	76,020,896	76,042,784	21,889	-	ABC8_000041036300_P14	interchromosomal
chr3	76,065,709	76,087,918	22,210	-	ABC8_000042165200_L2	interchromosomal
chr4	176,002,721	176,028,247	25,527	-	ABC8_000002132540_P16	interchromosomal
chr5	49,787,235	49,811,255	24,021	-	ABC8_000041009200_I22	interchromosomal
chr10	75,433,721	75,490,148	56,428	AGAP5, BMS1P4	ABC8_000005704649_A15	interspersed intrachromosomal
chr12	34,116,099	34,136,300	20,202	-	ABC8_000002114140_J11	interchromosomal
chrX	50,740,664	50,770,421	29,758	-	ABC8_000040982900_O11	intrachromosomal tandem duplication

Table ST3.3 Duplication status via FISH of WGAC predicted duplications

3.2 Lineage-specific duplications in the ape lineage

We estimated lineage-specific duplications for human, chimpanzee, orangutan, and gibbon lineages using the WSSD method described by Marques-Bonet et al. 2009[16]. Sanger capillary reads from the gorilla genome project[19] did not reach a coverage comparable to the one of the other species and we therefore excluded this species from our analysis. Lineage-specific and shared duplications were calculated by mapping WGS reads from each species to a common reference (GRCh37) and identifying regions of excessive read depth. Copy number calls for each non-human species were scaled by the corresponding human copy number in the reference genome assembly to correct for reference bias. We found lineage-specific duplications for human, chimpanzee, and orangutan that closely match those in original analysis by Marques-Bonet et al. 2009 after copy number correction[16]. In addition to the great-ape duplications, we found 17 Mbp (~6 Mbp before copy-number correction) of gibbon-specific duplications and 5 Mbp of common (ancestral) duplications between gibbon and great-ape lineages (Extended Data Fig.2 and Supplementary File 4). These gibbon-specific duplications

contained 84 duplicated genes many of which are weakly enriched for olfactory receptors (enrichment score: 2.68) and sensory (enrichment score: 2.54) functions based on a DAVID functional classification[20] (Supplementary File 4). Particularly interesting duplicated genes include CHAD a protein involved in cartilage production, BZRAP1 a benzodiazapine receptor-associated protein, and IFT74 a protein involved in capillary morphogenesis.

Validation of duplications by FISH and array-comparative genomic hybridization (array-CGH)

We tested the results from the WSSD method on GRCh37 by performing FISH experiments with 109 human fosmids selected on regions bigger than 17 kbp (Supplementary Files 4 and 5). Among the selected fosmids, 34 fosmids contained duplications shared with other primates (human, chimpanzee, gorilla and orangutan) while 75 fosmids were selected to validate gibbon lineage-specific duplicated regions, with or without genes (36 and 39 fosmids respectively) (Fig. SF3.2). Duplication status in each species was determined as described in this section, with the exception of gorilla for which we used estimates from the Illumina-based WSSD method and the human for which we used the copy numbers from Sudmant et al. 2010[21] to infer the duplication status. Relative to the 34 shared duplications with other primate species, all the regions were confirmed as duplicated in gibbon genome except for 5 fosmids, whose duplication state could not be determined by FISH analyses because of the presence of high background noise (Table ST3.4a, Supplementary File 4). We estimated our confirmation rate by FISH of gibbon specific duplications as 100% for regions with genes and 94% for regions without genes (Table S3.4b-c and Supplementary File 4).

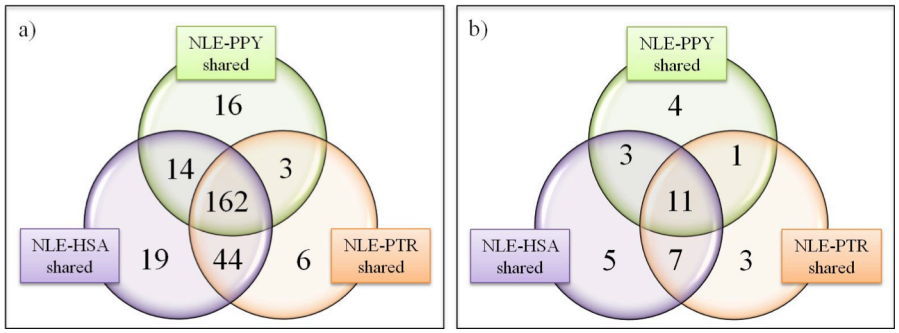


Figure SF3.2 a) Total shared duplications detected by WSSD strategy. b) Distribution of shared duplicated regions detected by WSSD and validated by FISH.

		Gibbon	HSA	PTR	GGO	PPY
a) WSSD shared regions with other primates	All	34	34	34	34	34
	not defined	5	4	4	5	5
	unconcordant	0	2	5	2	10
	concordant	29	28	25	27	19
	concordance	**1.00**	**0.93**	**0.83**	**0.93**	**0.66**
b) gibbon-specific regions that harbor genes	All	36				
	not defined	5				
	unconcordant	0				
	concordant	31				
	concordance	**1.00**				
c) gibbon-specific regions that don't harbor genes	All	39				
	not defined	7				
	unconcordant	2				
	concordant	30				
	concordance	**0.94**				
d) gibbon-specific regions not previously confirmed by aCGH	All	45				
	not defined	5				
	unconcordant	4				
	concordant	36				
	concordance	**0.90**				
d) shared regions not previously confirmed by aCGH	All	2	2			
	not defined	0	0			
	unconcordant	0	1			
	concordant	2	1			
	concordance	**1.00**	**0.5**			

Table ST3.4 Summary of FISH results. (Abbreviations: HSA= Homo sapiens, PTR=Pan troglodytes, GGO= Gorilla gorilla, PPY= Pongo pygmaeus)

In the gibbon genome, only two of the selected regions scored negative for this assay, while 98% (90/92) of the regions were confirmed duplication positive by this assay. Seventeen regions remained unclear due to background noise in the experiment. Noteworthy, 65/90 showed evidence of duplicated signals that had a multichromosomal distribution pattern as opposed to a clustered intrachromosomal configuration. Conversely to the mouse[22], dog[23] and cattle genomes[24], all FISH results in this project demonstrate that interchromosomal duplications predominate in the gibbon genome (Fig. SF3.3).

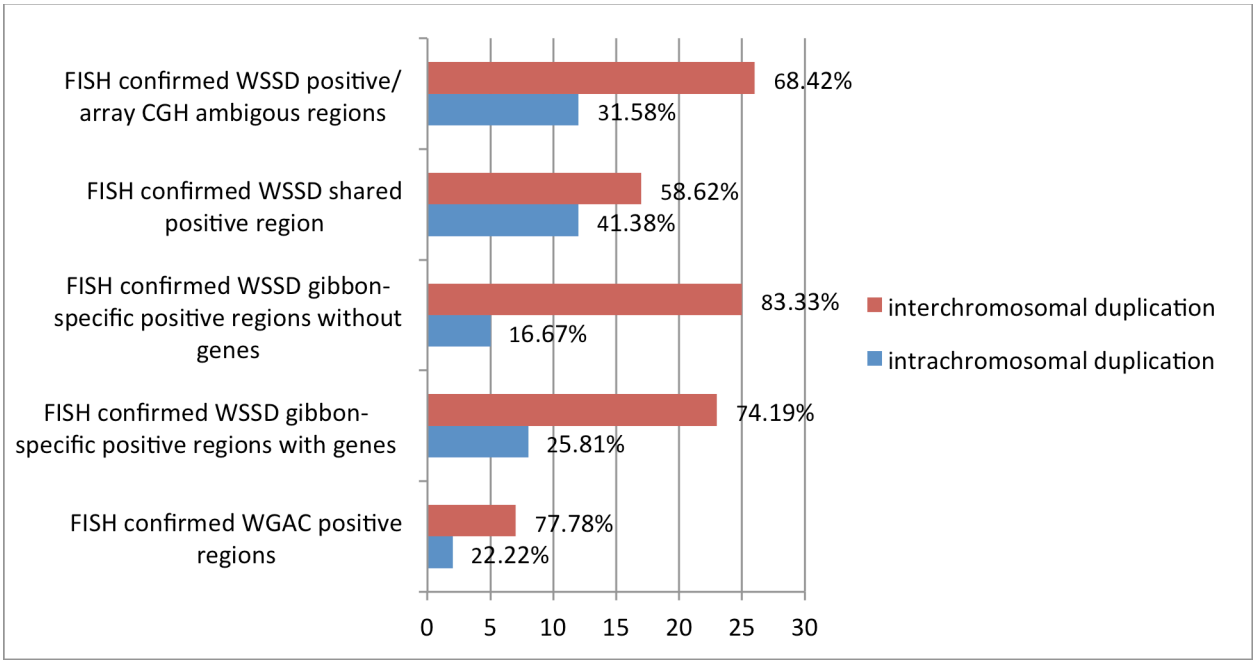


Figure SF3.3 Distribution of intra and interchromosomal duplications confirmed by FISH in the gibbon genome. Percentage of each category in each analysis shown.

Additionally, we tested gibbon-only duplications with a dye-swap array CGH experiment using Nimblegen 2.1 arrays with probes for a gibbon female (Asia) and a human female (G248). A hidden Markov model (HMM) was used to call duplications from log2 values and support 4.16 Mbp (71%) of the gibbon-only duplications (Table ST3.5, Fig. SF3.4), lower than in previous assays[16]. We validated by FISH experiments regions bigger than 17 Kbp that resulted WSSD positive but ambiguous by array CGH (Table ST3.4d-e and Supplementary File 4). In particular, we tested 47 fosmids on mixed slides (gibbon and human metaphases and interphases): 45 fosmids were selected to validate gibbon-specific duplications, while 2 fosmids were used for gibbon and human shared duplications.

Call Type	Regions	% Validated	Total Size (Mbp)	% Validated
Sanger WSSD	143	-	6.1	-
Array CGH	101	71%	4.9	81%
Illumina WSSD	114	79%	4.8	78%
All support	81	56%	3.7	60%
Any support	130	90%	5.7	93%

Table ST3.5 Array CGH and FISH validation of WSSD duplications. For both validation methods, the total number of validated duplications is shown along with the corresponding percentage of the original 143 duplications that validated by that method. Also shown are the total size of the validated duplications and percentage of the original size. In summary are the total number of regions and total sizes of duplications that validated by both methods and by at least one method.

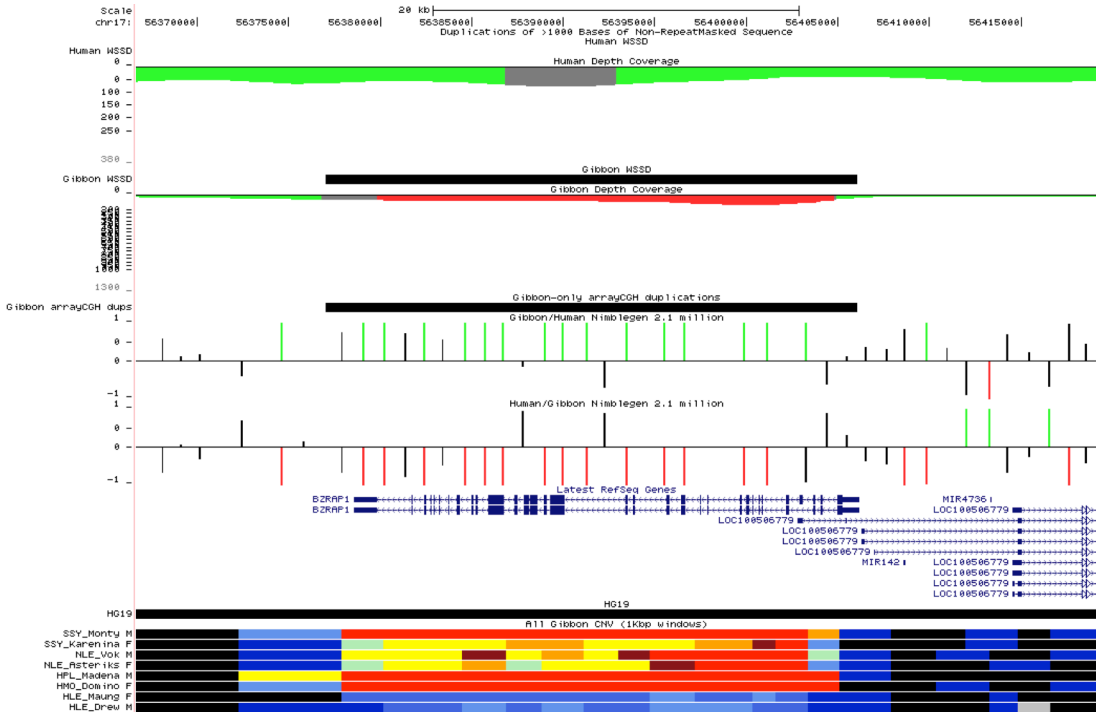


Figure SF3.4 Representative gibbon-only WSSD call by read depth validated by array CGH showing complete duplication of the gene BZRAP1. Illumina copy-number heatmaps show strong support for the same call in all eight individuals from five species.

Our experiments revealed that 38 out of 47 investigated regions (90%) were clearly confirmed as duplicated in gibbon, while 4 probes showed single signals. These results suggest that WSSD

predictions were correct and reliable, but the array-CGH confirmation had a lower validation rate, likelybecause of the divergence of the human probes used.

3.3 Assessing levels of variation among gibbon genera

Calling of gibbon duplications using the human assembly

We wanted to assess the variation of duplications for the eight gibbon individuals from the diversity panel (Table ST2.1) by mapping Illumina reads to GRCh37 using methods previously described in Sudmant et al. 2010[21]. We defined duplicated regions as those with copy number ≥3, at least 10 Kbp of non-repeat sequence, and no overlap with known artifacts of GRCh37 (Table ST3.6 and Supplementary File 4). After applying these filters, we performed an initial validation of the Illumina calls from the *Nomascus* female individual "Asteriks" by intersecting these calls with the Sanger WSSD calls which were based on sequence from Asia (the reference), an individual of the same species (also a female). Illumina calls overlapped with 15.2 Mbp (81%) of the original 18.7 Mbp from Sanger excluding chrX, chrY, and random and unplaced chromosomes.

Genus	Sample	Bps without Gaps	Bps copy-number corrected	% Genome	% Genome (CN Corr)
Hoolock	Drew	38,602,811	87,004,943	1.332	3.002
	Maung	39,820,263	86,048,337	1.374	2.969
Siamang	Karenina	42,166,336	146,511,755	1.455	5.056
	Monty	43,234,360	159,690,705	1.492	5.511
Hylobates	Domino	39,666,740	80,393,608	1.369	2.774
	Madena	49,297,162	117,349,702	1.701	4.05
Nomascus	Vok	45,708,217	107,489,385	1.577	3.709
	Asteriks	43,334,661	103,702,239	1.495	3.579

Table ST3.6 Duplications detected in each sample from WSSD against GRCh37

Calling of gibbon duplications using Nleu1.0

To account for potential human biases derived from using the human reference assembly, we also applied the WSSD strategy mapping the Illumina reads from each sample against Nleu1.0. We used a masked version of the reference using RepeatMasker (http://www.repeatmasker.org) and TandemRepeatFinder[25]. In addition, we masked 36-bp kmers that are present more than 20 times in the assembly. To identify the reference regions with these over-represented kmers, we partitioned the scaffolds of the assembly into 36-bp kmers (with adjacent kmers overlapping 5 bps) and mapped them against Nleu1.0 using mrsFast[26].

To estimate the read depth in non-repetitive 1 kbp sliding windows in the genome we followed the same steps previously described in Alkan et al. 2009[27]. This is, firstly we fragmented reads into smaller portions of 36 bp (keeping positions 10–45 and 46–81 of each read, so we excluded the lower-quality ends) and we mapped the resulting reads with mrFAST[27] against the masked assembly Nleu1.0. We then defined sliding windows of 1 kbp of non-repetitive sequence, and we counted the number of reads that map in each window. Finally we corrected the read depth by GC content and calculated the copy number of each window considering a set of single copy regions of the assembly.

We defined the control regions from BACs that were determined as single-copy via FISH[9] or being described as without any segment of duplicated sequence[10]. From those BACs, there were 21 that we additionally confirmed as unambiguously unique by read depth (WSSD with gibbon WGS Sanger reads from Asteriks sample). To determine their location on the gibbon reference we aligned these 21 BACs to the unmasked gibbon assembly using MEGABLAST (version 2.2.19, parameters –D 2 -v 5 -b 5 -e 1e-70 -p 89 -s 220 -W 12 -t 21 -F F) and identified contiguous blocks of alignments. We were able to unambiguously determine the exact position of each BAC in the assembly as we got a unique block per BAC except for one BAC (CT954298.5), which we conservatively removed from our list of control regions (Supplementary File 4). Thus, we got 20 control regions in 16 scaffolds spanning 3,548,976 bp (Table ST3.7).

BAC id	BAC start	BAC end	Scaffold	Scaffold start	Scaffold end
AC198097.2	1	177,271	GL397370.1	395,344	573,022
AC198099.1	1	107,387	GL397261.1	11,251,401	11,358,723
AC198101.2	1	201,469	GL397267.1	11,644,597	11,847,069
AC198102.2	1	196,265	GL397305.1	429,759	639,270
AC198103.2	1	166,521	GL397263.1	11,732,663	11,899,504
AC198144.2	1	182,039	GL397303.1	19,942,889	20,123,058
AC198146.2	1	189,226	GL397300.1	1,622,826	1,814,749
AC198147.2	1	212,311	GL397352.1	6,538,389	6,749,905
AC198150.2	1	181,383	GL397280.1	10,490,861	10,679,449
AC198151.2	1	175,544	GL397265.1	20,732,890	20,910,964
AC198152.2	1	139,965	GL397275.1	31,940,434	32,079,964
AC198154.2	1	131,717	GL397330.1	2,745,882	2,882,143
AC198526.1	1	192,813	GL397298.1	22,644,029	22,841,197
AC198875.2	1	182,505	GL397261.1	47,981,724	48,163,923
AC198945.2	1	203,114	GL397271.1	22,336,767	22,539,792
AC225372.3	1	208,921	GL397269.1	40,916,029	41,124,834
CT954300.6	1	159,125	GL397298.1	22,562,223	22,721,516
CT954301.7	1	186,267	GL397275.1	31,976,284	32,163,199
CT954310.8	1	190,981	GL397399.1	1,497,855	1,663,599
CT954321.3	1	150,097	GL397261.1	11,308,384	11,464,326

Table ST3.7 Single-copy control regions and their corresponding coordinates in NIeu1.0

Further, the copy number distribution in the control regions was used in order to define sample specific gain/loss cutoffs as the mean copy number plus/minus three units of standard deviation (calculated not considering those windows exceeding the 1% highest copy number value). Then, we merged 1 kbp windows with copy number larger than sample-specific gain cutoff (but lower than 100 copies) and we identified as duplications the regions that comprise at least five 1 kbp windows and >10 kbp. Finally, only duplications with >85% of their size not overlapping with repeats were retained.

We identified between 35.19 and 49.08 Mbp of duplicated sequence in the reference (once removed gaps from the assembly) in the eight samples (Table ST3.8). To check the consistency of our data we analyzed the Asteriks sample and we intersected the duplication calls from both methods. We only considered the calls greater than 20kb in both sets. In total we obtained 17.70 Mb of events >20 kb that overlap in both duplication sets, this corresponds to 78.35% of the calls from Illumina data (17.70/22.59 Mb, duplications >20 kb and removing gaps) and to the 67.17% of the duplication calls that come from the Sanger reads (17.70/26.35 Mbp, gaps are not included either).

Genus	Sample	Bps without Gaps	% Genome
Hoolock	Drew	40,791,641	1.389
	Maung	43,133,688	1.469
Siamang	Karenina	38,096,714	1.298
	Monty	35,187,318	1.198
Hylobates	Domino	38,634,497	1.316
	Madena	45,180,206	1.539
Nomascus	Vok	49,080,692	1.672
	Asteriks	42,305,215	1.441

Table ST3.8 Duplications detected in each sample from WSSD against Nleu1.0

Comparison of WSSD on two assemblies (GRCh37 and Nleu1.0)

We calculated and compared genus-specific duplication calls from Nleu1.0 and GRCh37 by intersecting calls for each pair of samples in a genus and subtracting any calls from each genus that were found in any of the other genera. We calculated ancestral duplications by intersecting calls from all samples. We lifted over these genus-specific and ancestral calls from Nleu1.0 coordinates to GRCh37 coordinates using a minMatch setting of 0.5. The set of shared calls between the two assemblies was determined by intersecting the GRCh37 calls with the lifted-over Nleu1.0 calls and calculating the total size of the lifted-over duplications that overlapped. We required at least 1 kbp per overlapping duplication. Almost

all of the original Nleu1.0 duplications survived the liftover and size filtering steps (>96.24%). Of these lifted over duplications, the fewest calls shared between Nleu1.0 and GRCh37 calls were 34.33% for *Siamang* specific calls while the most were 61.45% for *Hylobates*, while for the ancestral calls the intersection was 47.73% (Table ST3.9).

	# Bps (nomLeu1)	# Bps (in hg19, result of the lift over)	% Survival of the nomLeu1 calls	# Bps (hg19)	# Bps intersect	% Bps intersect (of the ones lifted to hg19)
Ancestral	16,611,021	21,425,136	98.72	24,199,737	10,225,507	47.73
Hoolock	3,299,164	3,583,923	99.55	3,315,692	1,741,290	48.59
Siamang	1,286,508	1,669,780	96.54	1,388,575	573,236	34.33
Hylobates	1,239,038	1,299,892	99.38	1,534,072	798,756	61.45
Nomascus	3,989,744	3,457,347	96.24	2,759,616	1,460,606	42.25

Table ST3.9 Intersection of duplications between calls in GRCh37 and Nleu1.0.

Validation of duplications by FISH of calls from mapping to Nleu1.0 and GRChr37

To validate predicted genus-specific or ancestral duplications found from both mapping to Nleu1.0 and GRChr37, we lifted over calls from Nleu1.0 to GRChr37 coordinates and intersected these calls with the GRChr37 calls. For validation, we selected duplications ≥10 kbp from this intersection (20 regions) as well as from lifted over Nleu1.0 calls that didn't intersect with GRChr37calls (15 regions) and calls from GRChr37 that didn't intersect with Nleu1.0 calls (15 regions). Human fosmids were used as probe to test these regions and we observed that predicted duplication status was confirmed for 13 out of 20 (65%) regions selected from the intersection of lifted over Nleu1.0 calls and GRChr37 calls, for 10 out of 15 regions (66,67%) selected from GRChr37 unique calls and for 0 out of 15 regions selected from Nleu1.0 unique calls. Not confirmed results for both GRChr37 unique calls and for the intersection of Nleu1.0 lifted over calls with GRChr37 calls were mostly due to the absence of genus specificity for the duplication (3/5 and 4/7 respectively). Instead, for Nleu1.0 unique calls wrong predictions were mostly

due to the observation of a single signal in regions described as duplicated (11/20). The results from the FISH suggest that the Nleu1.0 based duplications had a higher rate of false positives, so, the following comparison of the different gibbon genera was based on the GRChr37 mapping.

3.4 Identification of genus-specific duplications

We calculated total duplications by genus, duplications ancestral to all genera, and genus-specific duplications. *Hylobates* had the most total duplications (52.9 Mbp) followed by *Nomascus* (48.9 Mbp), *Siamang* (46.0 Mbp), and then *Hoolock* (43.3 Mbp). The four genera shared 22.8 Mbp of ancestral duplications. *Hylobates* also contained the most genus-specific duplications with 8.3 Mbp while *Hoolock* had 5.6 Mbp, *Nomascus* had 4.9 Mbp, and *Siamang* had 3.4 Mbp (Table ST3.10). The fraction of these genus-specific duplications that were fixed (appeared in both samples) ranged from 17% in *Hylobates* to 60% in *Hoolock*.

a)	**TOTAL DUPLICATIONS**	**All**	**Genus-fixed**	**Genus-polymorphic**
	Hoolock	43,285,836	35,157,990	8,127,846
	Siamang	46,048,327	39,382,069	6,666,258
	Hylobates	52,855,069	36,135,636	16,719,433
	Nomascus	48,895,084	40,175,433	8,719,651
b)	**GENUS-SPECIFIC DUPLICATIONS**	**All**	**Genus-specific fixed**	**Genus-specific polymorphic**
	Hoolock	5,557,300	3,315,692	2,241,608
	Siamang	3,445,078	1,388,712	2,056,366
	Hylobates	8,324,253	1,389,277	6,934,976
	Nomascus	4,861,046	2,723,699	2,137,347

Table ST3.10 Duplications by genera, showing fixed and polymorphic of a) the whole set of duplications; b) the genus-specific ones.

Using the genus-specific duplications, we searched for genus-specific gene duplications (Supplementary file 4) and found 20 (8 fixed) for *Hoolock* with a functional enrichment for sensory perception and cognition (enrichment score: 1.66), 51 (7 fixed) for *Hylobates* with a functional

enrichment for defensins and antibiotics (enrichment score: 5.55), 29 (10 fixed) for *Nomascus* with a functional enrichment for nuclease activity (enrichment score: 3.2), and 20 (10 fixed) for *Siamang* with no strong functional enrichment.

3.5 Cross-species cDNA array CGH

Cross-species cDNA array CGH was conducted using three hylobatidae species obtained from Coriell Cell Repositories[28], including *Hylobates gabriellae* (buff-cheeked or red-cheeked gibbon, Coriell PR00381), *Simphalangus syndactylus* (siamang, Coriell PR00721), and *Hylobates lar* (white-handed gibbon, Coriell PR00495) as the test samples. The reference sample was a sex-matched control human. Gene copy number gains and losses were called if the log2 ratio of the red (test genomic DNA signal) to green (reference genomic DNA signal) was based on a threshold of a log2 ratio greater than the absolute value of 0.5, and at least 2 of the three gibbons were required to share the copy number change in the same direction. Additionally, the absolute value of the average log2 ratio for the gibbon species had to be at least 2.5 fold greater than the average log2 ratios for human versus human comparisons. Detailed methods are reported in Dumas et al., 2007[28]. Using these criteria, we detected 336 lineage specific gains and 213 gibbon-specific losses. Because the arrays were designed according to the human cDNA clones, sequence divergence between the gibbon and human genomes could have overestimated the number of gibbon-specific losses. Therefore, in the following analyses, only the gibbon lineage-specific copy number increases were included.

For confirmation of gene duplications in the gibbon genome, the top 50 EST sequences that gave the strongest duplication signals from a consensus of all three gibbon arrays run were used as BLAT queries against the human hg19 and gibbon genome builds. BLAT hits were considered significant with a score of greater than or equal to 100. Of the 50 ESTs queried, 12/50 or 24% showed more copies in the gibbon genome build than in the human genome. The gibbon genome build and hg19 report that 29 of the cDNA array CGH-predicted gibbon lineage-specific increases have the same copy number in humans and gibbon, which likely indicates that the genome assembly has collapsed those sequences.

The genomic coordinates from the human genome BLAT queries were then compared to the coordinates of duplications found in the WSSD analysis. Only 7 genes (14%) of the top 50 were found to be common to both lists. Reasons for the weak overlap may include differences in criteria used to call duplications and the use of repeat masking. In order to be considered a duplication by WSSD, the sequence had to have >94% identity and length ≥20Kbp. This cutoff would have omitted detection of smaller size duplications such as could have been found via cDNA array CGH. The maximum cDNA probe size on the arrays was only 1.5kp, with the average size being near 500 bp. Additionally, WSSD calls were filtered to omit regions with >85% overlap with RepeatMasker calls or 75% overlap with Tandem Repeat Finder calls. All satellites and L1P repeats in the reference genome were masked to limit false positives seeded by those repeats. The array CGH experiments included Cot-1 DNA to block repetitive elements during hybridization. However, the bioinformatics analysis conducted on the gibbon lineage-specific signals predicted by array CGH did not mask for any repeats. Approximately 6% (3/50) of the array content was masked out of the WSSD analysis.

Of the 7 genes in common between the array data and the WSSD, 5 show multiple copies in the gibbon genome build. These genes are *EFHC2*, *DLG1*, *ZNF74*, *NPEPPS*, and *CEP112*. Two of these confirmed genes may have potentially interesting biological consequences, with DLG1 shown to be involved in T cell signaling and viral protection[29] and NPEPPS inhibits Tau induced neurodegeneration[30].

Supplemental Section S4 – Estimating timing of the gibbon / great ape split

Overview

In order to infer the population divergence time of gibbons from great apes using the Nleu1.0 assembly we applied a coalescent-based approach described by Rannala and Yang[31] that has been further developed to increase computational efficiency by Gronau et al.[32] in the software package G-PhosCS. We applied this method to ~15,000 independent 1kb loci from across the Nleu1.0 genome as well as aligned sequence from the GRCh37, ponAbe2 and rheMac2 genomes. In addition we performed simulations that take into account the observed sequence context (e.g. GC content) and a more complex mutational model in order to determine whether such features may bias our G-PhosCS estimates, which are based on more simplified assumptions. We report an estimate of the divergence time for small apes and great apes as a fraction of the divergence time between apes and old world monkeys.

Data

Alignments of human (Hs), orangutan (Pp), gibbon (Nl) and macaque (Mm) autosomal sequences were obtained from the 11-way vertebrate multiz alignments available at UCSC (http://hgdownload.soe.ucsc.edu/goldenPath/gorGor3/multiz11way/). These alignments were masked to remove bases with low quality (Sanger Phred quality score <50), repetitive sequence, and CpG sites (i.e. with at least one individual having the dinucleotide CG and at least one individual having either the dinucleotide TG or CA). The latter filter was included as the multiway alignments demonstrated a higher mutation rate for CpG sites (as has been demonstrated previously[33]) and G-PhosCS assumes that all mutation types are equally likely. In order to identify independent segments of sequence, we identified loci that were between 1,000 bp and 14,374 bp in length (the latter corresponds to the longest multiple sequence alignment). To perform an analysis that assumes a lack of intralocus recombination and independence among loci we applied two additional filters to this dataset. For the first filter, we required that the first 1000 bp of a locus that are not masked occur within the first 3000 bp and for the second

that all loci are at least 50kb from each other. This results in 14,962 alignments each of length 1kb. Divergence values between the various species pairs can be found in the ST4.1.

	Human	Orang	Gibbon	Macaque
Human		0.90	1.04	1.54
Orang	2.43		1.04	1.55
Gibbon	2.90	2.92		1.58
Macaque	4.71	4.73	4.87	

Table ST4.1 % mean pairwise divergence (lower diagonal) and standard deviation (upper diagonal) for 14,963 1kb alignments.

G-PhoCS analysis

We analyzed the mutilocus dataset described above using the program G-PhoCS[32] to estimate the divergence times of 1) Old World monkeys and apes, 2) small and great apes, and 3) humans and orangutans. For a model of sequential population splits, G-PhoCS implements a coalescent-based Markov-Chain Monte Carlo (MCMC) search in the space of demographic parameters θ, τ and *m* (the last of which we ignore for this scenario as in the method of Rannala and Yang[31]). The topology and all parameters we assume for this analysis are shown in Fig. SF4.1. The method assumes that all types of mutations are equivalent and allows for multiple mutations at a site.

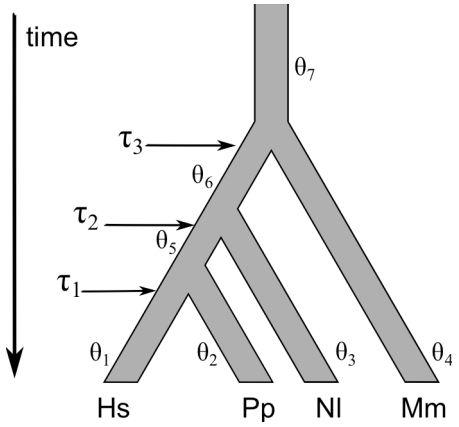


Figure SF4.1 Model and parameters assumed for the G-PhoCS analysis

Assuming the mutation rate per year depends only on the locus and is the same for all species, two factors affect the expected number of mutations per branch: (1) the mutation rate for the locus and (2) the branch lengths. The variation in the expected number of mutations could be due to variation in the mutation rates across loci, but can also be caused by the stochasticity of the coalescent process, which results in variable branch lengths. We evaluated the distribution of mutation rates across loci by considering the distribution on the number of differences between human and macaque sequences. For this purpose we used a model with species divergence of ~27 Mya (Million years ago) (1,080,000 human generations) from an ancestral effective population size of 50,000. We assumed an ancestral generation time of ~11 years[34], and modeled the variation in mutation rates using a β distribution. We compared the distribution of the number of mutations between macaque and human in simulated and observed data sets (Fig. SF4.2). A model of constant mutation rate provides a poor fit to the data (Fig. SF4.2-a and SF2.4-b). Likewise, a uniform distribution of mutation rates, which corresponds to a beta distribution with alpha and beta parameters equal to 1 (the default for G-PhoCS) results in too many regions with very low divergence (Fig. SF4.2-c). We adjusted the values of the parameters of the β distribution and mutation rate such that the empirical and simulated distributions have a good agreement on their mean, variance, skew and kurtosis. We determined that a β distribution with parameters 2.64 and 1 provided a good fit (Fig. SF4.2a,d).

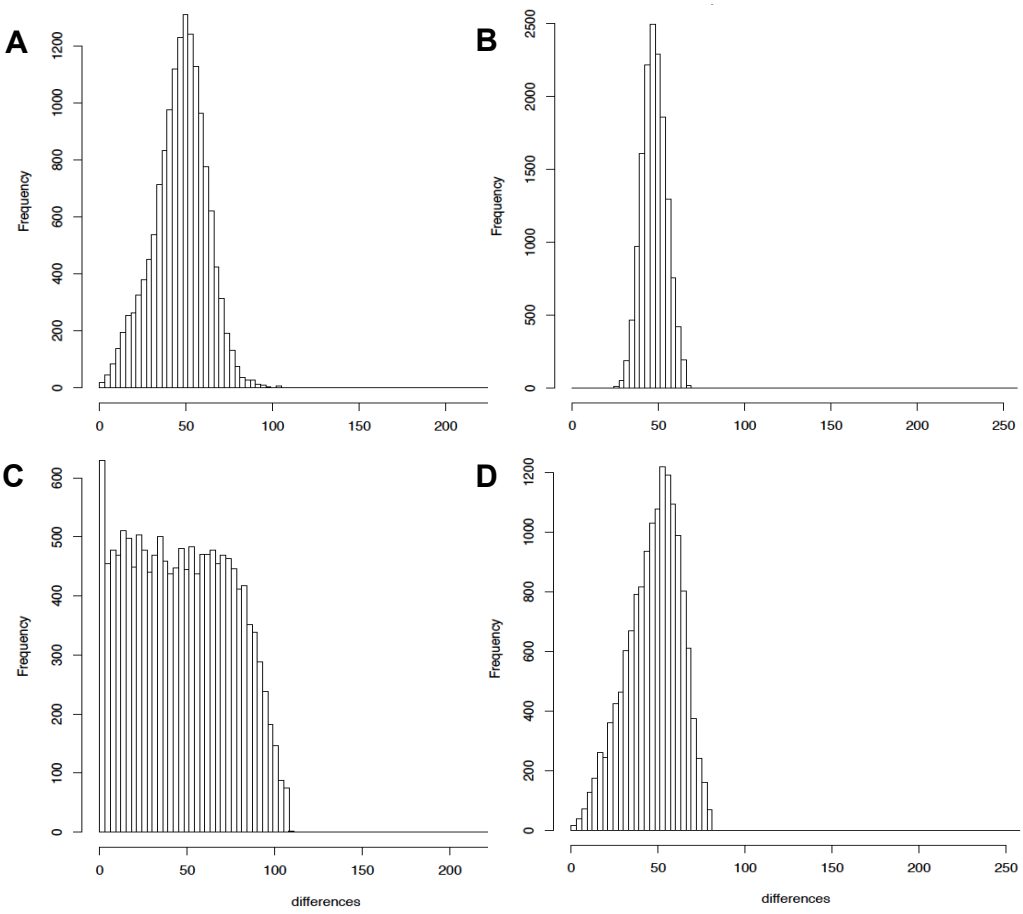


Figure SF4.2 a) Distribution of pairwise differences observed between humans and macaques at 14,962 1kb alignments; b) Distribution of simulated pairwise differences under a single mutation rate with N_e = 50,000. ; c) Distribution of simulated pairwise differences assuming a uniform distribution of mutation rates (the GPhoCS default); d) Distribution of simulated pairwise differences assuming a β distribution with parameters 2.64 and 1.

Possible Influence of a Hominoid slowdown

G-PhoCS assumes that all lineages share the same mutation rate. However, it has been proposed that there has been a slowdown in evolutionary rate since the divergence of apes and other primates, correlating somewhat with life history traits such as increased body size, increases in generation times and lower metabolic rate[138]. Molecular evidence for this slowdown comes from shortened branch lengths in larger primates, including humans, and a discrepancy of ~50% between mutation rates estimated using human-chimpanzee divergence and whole genome sequencing of pedigrees[139-141]. We

observe that levels of divergence from gibbon and from macaque are very similar for human and orangutan (Table ST4.1). Divergence from macaque is also similar for gibbons and the two great apes, although a small increase is noticeable for the former, a pattern expected if gibbons have a slightly higher mutation rate (though higher reference sequence error rate for gibbons could also contribute to this observation). If a slowdown in evolutionary rate is acting from macaques through to orangutans and humans, there may be a slight underestimation in the speciation time for gibbons, though if the slowdown only began after the split of great apes from gibbons, this effect is likely to be minimal. However, it is important to appreciate that our final estimate of divergence times using G-PhoCS does not depend on an assumed mutation rate, it depends only on the relative divergence times and the absolute divergence time for apes and Old World Monkeys. This last time is not well known and will almost certainly be the main source of error in our estimates, dominating that from the underestimation due to variation in evolutionary rate with a slow down in larger primates. While we assumed a divergence of 29 million years, a likely lower bound of 25 mya was recently described, with a plausible upper bound perhaps extending substantially past 30 mya[142].

Description of the priors

G-PhoCS uses a Bayesian approach for the estimation of the parameters, with priors specified by the user. The priors for the effective population sizes and divergence times are γ distributions, and that for inter-locus variation in mutation rates is a Dirichlet distribution. We chose exponential priors (an exponential distribution is a β distribution with shape parameter equal to 1) for the effective population size and divergence times (Table ST4.2), and β (2.64,1) for the mutation rates (a β distribution is a Dirichlet distribution with only two dimensions).

Parameter	Distribution
q_H	Gamma(1,10^{10})
q_O	Gamma(1,10^{10})
q_G	Gamma(1,10^{10})
q_M	Gamma(1,10^{10})
q_{HO}	Gamma(1.5,750)
q_{HOG}	Gamma(1.5,750)
q_{HOGM}	Gamma(1.5,750)
t_{HO}	Gamma(1,72)
t_{HOG}	Gamma(1,60)
t_{HOGM}	Gamma(1,40)
Locu-mut-rate	Beta(2.64,1)

Table ST4.2 Priors used in G-PhoCS

Because at each locus we are using a single haploid sequence per species, and therefore have no information on the heterozygosity of the population, we use an exponential distribution with a very large rate parameter for the priors of the current population size, which effectively fixes the posterior estimate at 0. The priors of current effective population sizes should not affect posterior estimates of ancestral effective population sizes or split times, because at the time of the split there should be a single ancestor per locus, given the assumption of no intra-locus recombination. We also note that a preliminary study demonstrated that the posterior distributions of split times are robust to the choice of the priors when the means were doubled or halved (results not shown).

Naïve G-PhoCS results

We ran G-PhoCS for 100,000 steps, discarding the first 50,000 as burn-in, and analyzing the distributions of the θ and τ parameters (and ratios of τ) from the remaining 50,000 steps. Assuming a

rate of 0.96×10^{-9} mutations per site per year and a generation time of 25 years, we estimated the effective size and split times as in Table ST4.3.

Generation times (years)	N_e			Divergence time (My)		
	Hs-Pp	Hs-Pp-NI	Hs-Pp-NI-Mm	Hs/Pp	Hs-Pp/NI	Hs-Pp-NI/Mm
25	40600	35300	54400	11.3	13.7	23.4
20,15,10	50800	58900	136000	11.3	13.7	23.4

Table ST4.3 Time estimates from G-PhosCS.

Note:Species names are abbreviated with the initial of their binomial names. - indicates population ancestral to those. / indicates separation.

The use of more realistic generation times of 20, 15 and 10 years for the ancestors of great apes, all apes, and the four lineages, respectively, result in larger estimates for the effective sizes (Table ST4.3), but does not affect the split time estimates. Alternatively, assuming the oldest split time as 29 Mya[35] we use the distribution of ratios to infer that apes and great apes split ~17My and ~14My, respectively.

Accounting for biases introduced by the mutation model and sequence context

Although G-PhoCS is a powerful method in ideal circumstances, the model used in G-PhoCS makes a number of simplifying assumptions about sequence evolution. In addition, we also applied a number of filters with regard to sequence content (CpG sites removal) and observed that this filtering strategy changed some properties of the simulated sequences, such as GC content and the transitions (Ti) to transversions (Tv) ratio. These factors may potentially bias our estimate of the population divergence times in complicated ways. Therefore, we conducted simulations that would allow us to evaluate the effect of these potential biases on the estimates of τ described above.

Our simulations take into consideration the fact that bases are not equivalent, with A and T being more common than C and G, and most mutations being transitions. We incorporated a mutation model that takes into account these asymmetries and also incorporates CpG sites. We simulated datasets of DNA

sequences that evolved following the established branching order that we have for Hs, Pp, Nl, and Mm. In the simulations, we specify all current effective population sizes to 10,000. Because all our data is haploid, this parameter is unimportant.

Our algorithm for generating an alignment of sequences consisted of the following steps:

1) Sample a value of GC content from a normal distribution
2) Simulate a non-recombining genealogy for the four sequences (coalescent tree) using the program ms[36].
3) Use the branch lengths in the coalescent tree to calculate the expected number of mutation in each branch.
4) Sample the number of mutations in each branch from a Poisson distribution with parameters calculated in 3).
5) Generate a random sequence of 1100 bp with bases sampled randomly according to their stationary frequencies.
6) Let the sequence evolve for 100 steps (mutations) according to the specified mutation model. The goal of this step is that CpG sites reach an equilibrium distribution. This is the sequence ancestral to all lineages considered.
7) Implement mutations sequentially in each branch. Positions are sampled randomly according to their probabilities given by the mutation model. Immediately following a mutation there is a small probability of a subsequent transition at CpG sites, with probabilities as specified in the section *Mutation Model* (see below).
8) The sequence produced after each iteration of this algorithm is filtered in a manner analogous to that with which we filtered our observed data.

In order to identify the set of sequence context parameters (mean and standard deviation of GC content, and Ti/Tv) that would consistently produce sequence alignments that matched those of the empirical data, we simulated sequences under different combinations of values for these parameters and performed a fit with the general linear model to infer the values of these parameters (data not

shown). All subsequent simulations used for determining biases of G-PhoCS were performed with the inferred values of 0.450 and 0.100 for the mean and standard deviation of GC content and 2.341for the ratio of Ti/Tv.

Mutation model

We used a time-reversible mutation model with the following characteristics:

1) All sites are independent (sites mutate in a context-independent fashion)
2) Mutation rates are the same in both complementary strands (resulting in a mutation scheme as in Fig. SF4.3). This also implies that if p_C, p_G, p_A, and p_T are the respective stationary frequencies of C, G, A, and T, $p_C=p_G$ and $p_A=p_T$.
3) At any site both transversions have the same rate.
4) The relative rate of transitions to transversions is the same for each base.

The last three conditions imply that for the rates shown in Fig. SF5.4 $\lambda_3=\lambda_5$, $\lambda_4=\lambda_6$,, $\lambda_1= 2*K*\lambda_3$, $\lambda_2= 2*K*\lambda_4$, and $\lambda_1/\lambda_2 = \Gamma/(1-\Gamma)$, where K is the rate of transitions to transversions and Γ is the GC content. We allow the GC content to vary across loci, resulting in λ_1/λ_2 also being variable. The ratio of the rates of transitions and transversions is constant. Though eventually we filter out CpG sites, we model their evolution to emulate the properties of real sequence. At these sites we assign mutation rates from C to T that are 10 times in excess of the average mutation rate for the sequence.

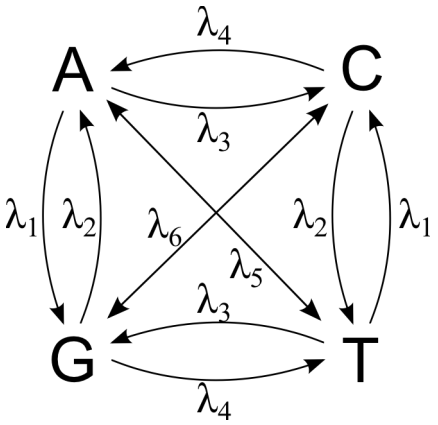


Figure SF4.3. Mutation Model for forward simulations of sequence context

Validation of simulations

Before attempting to infer potential G-PhoCS biases, in order to check that our simulations were implemented correctly, we simulated sequence evolution under the Jukes-Cantor model of mutation[37], where all mutations are equivalent, and compared the inferred and simulated parameter values for data sets of 1000 loci. We performed this experiment 500 times with simulation parameter values sampled from log-uniform distributions for the effective size, uniform for the divergence times (Table ST4.4) and a mutation rate of 2.4×10^{-9} mutations per site per generation. We used this data set of 500 simulations to regress the inferred value of an individual parameter estimate from G-PhoCS on the actual value used in the simulation. In each case we used the median value from the G-PhoCS output for the estimate.

Parameter	Range	Every
$\log_{10}(N_{eHO})$	4.301-5.176	0.097
$\log_{10}(N_{eHOG})$	4.301-5.176	0.097
$\log_{10}(N_{eHOGM})$	4.301-5.176	0.097
Td_{H-O} (My)	12My-16My	0.5My
Td_{HO-G} -Td_{H-O} (My)	1My-5My	0.5My
Td_{HOG-M} -Td_{HO-G}(My)	9My-15My	0.5My

Table ST4.4 Priors for parameters in the simulations for validation of G-PhoCS analysis

We observe good accuracy and a strong correlation between simulated and inferred split times (relative error<0.024 , R^2>0.85, P<3x10^{-16} with the assumption of normality of the error). Also, there is a moderate correlation between simulated and inferred effective population size for the population ancestral to great apes (R^2 ~0.58, P<3x10^{-16}, assuming normality of the error), with G-PhoCS underestimating the ancestral population size by about 30% (data not shown). However, the correlation between simulated and inferred ancestral effective size for the other two ancestral population size is very weak (R^2<0.01, P>0.7). These results suggest both that the simulations work correctly and that there is little power to estimate the ancestral population sizes of all apes and of all species using this method.

Properties of the sequences

Having established that we can reliably simulate realistic sequence evolution we attempted to infer whether G-PhoCS could reliably recover the true underlying demographic parameters that produced the simulated data and to quantify potential biases in the G-PhoCS inference. We simulated data under our mutational model (priors in Table ST4.5) and compared the inferred parameter estimates produced by G-PhoCS after filtering (as before we used the median) with the true values. We obtained the

regression line and observed an underestimation of the inferred split times between 7 and 12% (relative value of the difference between the fit and theoretical value for the range considered). G-PhoCS run on post-filtered sequences also underestimates the ancestral effective population size of great apes by about 6%.

Parameter	Range	Every
$\log_{10}(Ne_{HO})$	4.204-4.794	0.097
$\log_{10}(Ne_{HOG})$	4.107-4.982	0.097
$\log_{10}(Ne_{HOGM})$	4.107-4.982	0.097
Td_{H-O} (My)	11.5My-15My	0.5My
Td_{HO-G} -Td_{H-O} (My)	1My-4My	0.5My
Td_{HOG-M} -Td_{HO-G}(My)	9My-14My	0.5My

Table ST4.4 Priors for parameters in the simulations for inference with GPhoCS

As in the case of the simple mutation model there was very limited information on the other two ancestral effective population sizes. The fraction of the variance in the estimated absolute split times explained by the regression ranged between 0.89 and 0.98. Instead, when we analyzed the ratio of the split times for apes and for apes and monkeys, we found that the linear regression was unbiased (slope ~0.99, ordinate ~0.03, R^2 ~0.87). A Q-Q analysis of the residuals suggests that they are approximately normally distributed with a standard deviation of approximately 0.015. This demonstrates that we can use our estimate for the real data and apply the correction resulting from the linear regression to obtain an unbiased estimate of this ratio.

We then simulated ~5000 data sets of 1000 loci each and estimated confidence intervals for the ratio of the split times of apes and of apes with Old World monkeys. To find confidence intervals for this ratio, we first binned its possible values for each of the simulations, taking intervals of 0.005 in a range from 0.53 to 0.65. We then used the empirical estimates for the different data sets to generate an empirical

distribution of estimated ratios given the actual (binned) ratio. For each bin we sampled with replacement a value from the corresponding empirical distribution 1000 times. Each time we also sampled a value from the posterior distribution of the inferred ratio as estimated from G-PhoCS using the empirical data. We tabulated what fraction of the times the value sampled from the simulated data produces a value larger than that sampled from the empirical data. Of the more than 5000 data sets, approximately 4400 had simulated values ranging from 0.53 to 0.65. For each bin, the number of simulations ranged from 73 to 252.

To find our symmetric 95% confidence interval, we first assumed that the values obtained above come from the cumulative distribution function of a normal distribution and inferred the mean and variance using the method of least square error to fit the distribution. We then found the extremes of our confidence intervals as the 0.025 and 0.975 quantiles for this normal distribution. We provide as point estimate the mean of the fitted normal distribution. Our estimate of the ratio is 0.578 (CI:0.550-0.605). Assuming a split time with macaque of 29 My, the estimate of the split time for apes would be 16.8My (CI:15.9-17.6 My). However, again we emphasize the split time with macaque is not well known, and is likely to be a greater source of error. We note that although the ratio of split times inferred from G-PhoCS without the correction for the mutation model is biased, the error introduced is considerably smaller than the confidence interval, which takes into consideration the uncertainty introduced by the other demographic parameters. Analogously, we infer a the split time of great apes of 14.2My (CI:13.4-15.1 My).

Our approach has assumed that all alignments were based on orthologous sequences. If any of the genomes had a larger fraction of paralogous sequences in the alignment we may expect to overestimate the amount of divergence associated with those. It is not clear which lineage is likely to produce more spurious alignments. While macaque is more distantly related to the other lineages, the gibbon genome sequence is less complete. Analysis of paralogous sequences may also produce a heavier upper tail in the distribution of pairwise differences.

Supplemental Section S5 – Analysis of gibbon-human synteny breakpoints

5.1 Overlap with genomic features: repeats and genes

To analyze the enrichment of genomic features in the regions flanking evolutionary breakpoints, we used a permutation based approach. The number of overlaps between breakpoint flanks and each feature of interest in Nleu1.0 (the observed overlap count) was compared to a background distribution calculated by randomly permuting the locations of breakpoint regions 100,000 times. The Nleu1.0 version of the assembly was used for these analyses. For our breakpoint regions, we chose those breakpoints for which we had single nucleotide resolution, and added flanking regions to either side of the breakpoint; in breakpoints in which the breakpoint fell within a gibbon-specific repeat element, we chose the flanking regions of the repeat. In each permutation, the location of each breakpoint region was randomly changed, while keeping the length and scaffold assignment of the breakpoint region the same. We then counted the number of overlaps between the randomized breakpoint regions and the feature of interest (the permuted overlap count). Enrichment p-values were computed as the proportion of permuted overlap counts that were more extreme than the observed overlap count. We also visualized the spatial relationship of breakpoint regions to each type of feature by simultaneously shifting the locations of the breakpoint regions up to 1 Mbp in each direction, in increments of 20 kb, and counting the proportion of shifted breakpoint regions that overlapped a feature of interest (regions that were shifted beyond the beginning or end of a scaffold were discarded). Permutation testing and shift testing were carried out using custom Python scripts and the BEDtools[38], pybedtools[39], and BEDOPS[40] libraries; the code is publicly available at https://github.com/cwhelan/permuting-feature-enrichment-test

We tested for enrichment in the breakpoint regions of the following features: genes, segmental duplications, and several classes of repetitive elements: *Alu*, L1, LAVA, and LTR. In addition to testing the entire *Alu* family, we also tested the subfamilies *AluS, AluJ,* and *AluY* individually. Gene locations were taken from Ensembl build 70. For the segmental duplication analysis, we used the segmental

duplications identified by the WSSD method (Section S3). *Alu*, L1, and LTR locations were identified from the RepeatMasker output. In order to determine the distance from the breakpoints at which enrichments are strongest, we varied the size of the breakpoint flanking regions by adding differently sized intervals; we tested flanking regions with a size of 100 bp, 250 bp, 500 bp, and 1000 bp. We corrected for multiple testing using the FDR under dependency method of Bejamini and Yekutieli[41]. Breakpoint regions are depleted for genes, but enriched for *Alu* elements and segmental duplications (Fig. SF 5.1, Table ST5.1). However, some of the genes overlapping with breakpoints belong to interesting biological categories (Supplementary File 1). The enrichment for *Alu* is primarily due to a strong enrichment of the *AluS* subfamily.

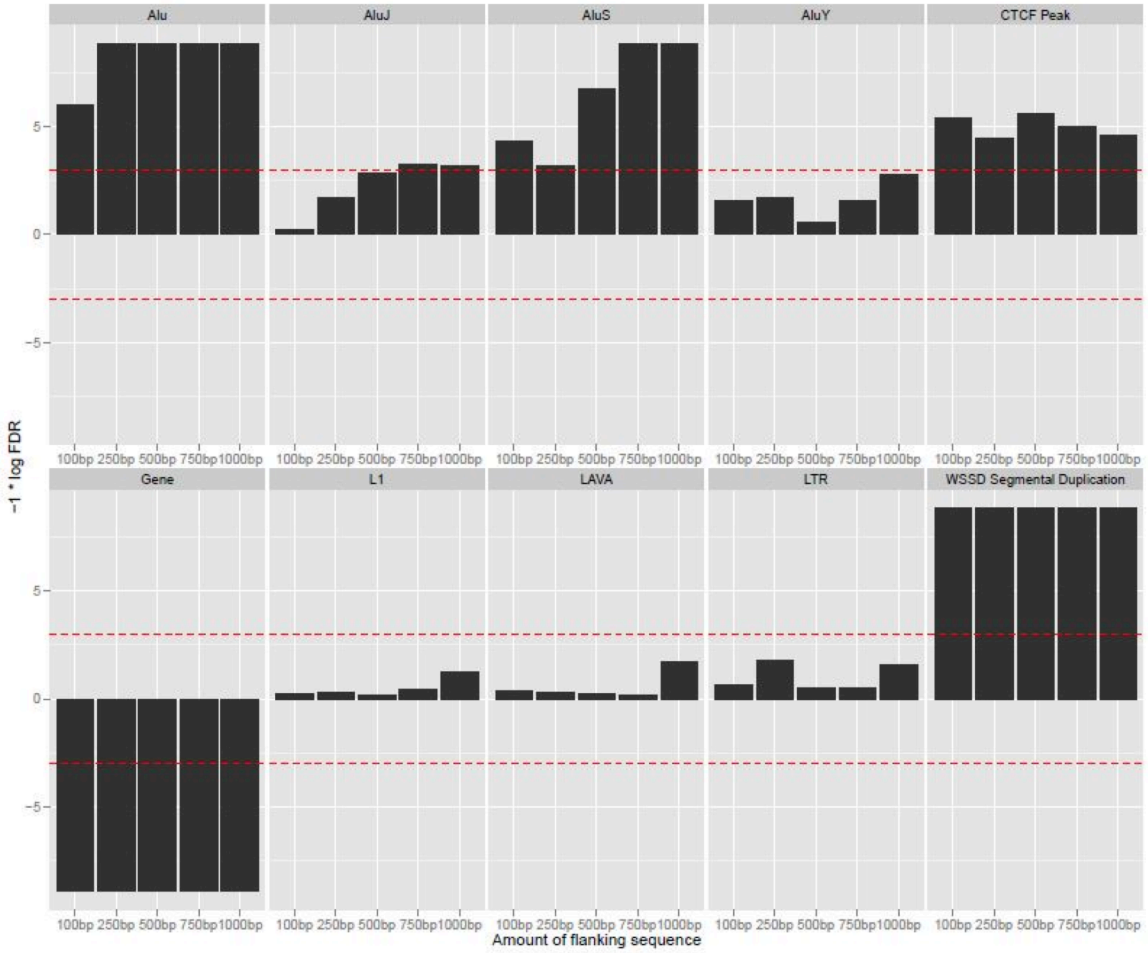


Figure SF5.1 Graphs show the results of permutation analyses to test the enrichment of breakpoints for different genomic features using different window sizes for the regions flanking the breakpoints (ranging from 100 bp to 1 kbp). Red dotted line indicates significance.

Feature	Number of overlapping breakpoints	Quantile	Adjusted -1 * log(P-value)
Alu	109	1.0000	8.8696
*Alu*J	35	0.9952	3.2191
*Alu*S	81	1.0000	8.8696
*Alu*Y	26	0.9918	2.7598
CTCF Peak	12	0.9990	4.5930
Gene	22	0.0000	-8.8696
L1	70	0.9516	1.2222
LAVA	1	0.9755	1.7284
LTR	48	0.9697	1.6054
Seg Dup	35	1.0000	8.8696

Table ST 5.1 Enrichment counts and scores of features in breakpoint flanking regions. For each feature type, we display the number of 1kb regions flanking breakpoints that overlap with a feature of that type, the quantile of that count in the empirical distribution obtained by permuting breakpoint flank locations 100,000 times, and the negative log FDR of that quantile treated as a p-value. Negative values indicate a depletion rather than an enrichment. Prior to FDR correction quantiles of zero were adjusted to p-values of 0.00001

Breakpoint regions were simultaneously shifted in increments of 25 kb, up to a maximum of 1 Mbp in each direction, and the proportion of breakpoint regions that overlap a feature of interest is reported. Shifts show that breakpoints are centered on regions that are depleted in genes but close to regions that contain genes, while the opposite is true for segmental duplications. *Alu* elements are more evenly spread across the shift regions (Fig. SF 5.2).

Finally, in addition to testing the count of overlaps between breakpoint flanking regions and repeats, we conducted a complementary test that examined the distance of each breakpoint to the nearest repeat of a given class. For this test we used only the 42 Class I breakpoints for which we had single nucleotide resolution. We compared these breakpoint locations to 10,000 randomly selected regions in the Nleu1.0 genome using the randomBed program from the BEDTools suite. The distance to the nearest repeat for either the breakpoints or random positions was determined using BEDTools closestBed. We

compared the distribution of distances to a repeat for the breakpoints to the distribution of distances to a repeat for the 10,000 randomly selected positions using the Kolmogorov-Smirnov (K-S) test (Table ST5.2). We examined the distance to any repeat, as well as those for *Alu*, LINE, and LTR elements, and finally the *AluJ*, *AluS*, *and AluY* subfamilies. After FDR correction for multiple hypotheses testing, the distance test showed similar results to the overlap test described above, with significant results for the *Alu* family as a whole, and for the *AluJ* and *AluS* subfamilies, indicating that the breakpoints tend to be closer to those repeats than random locations in the genome.

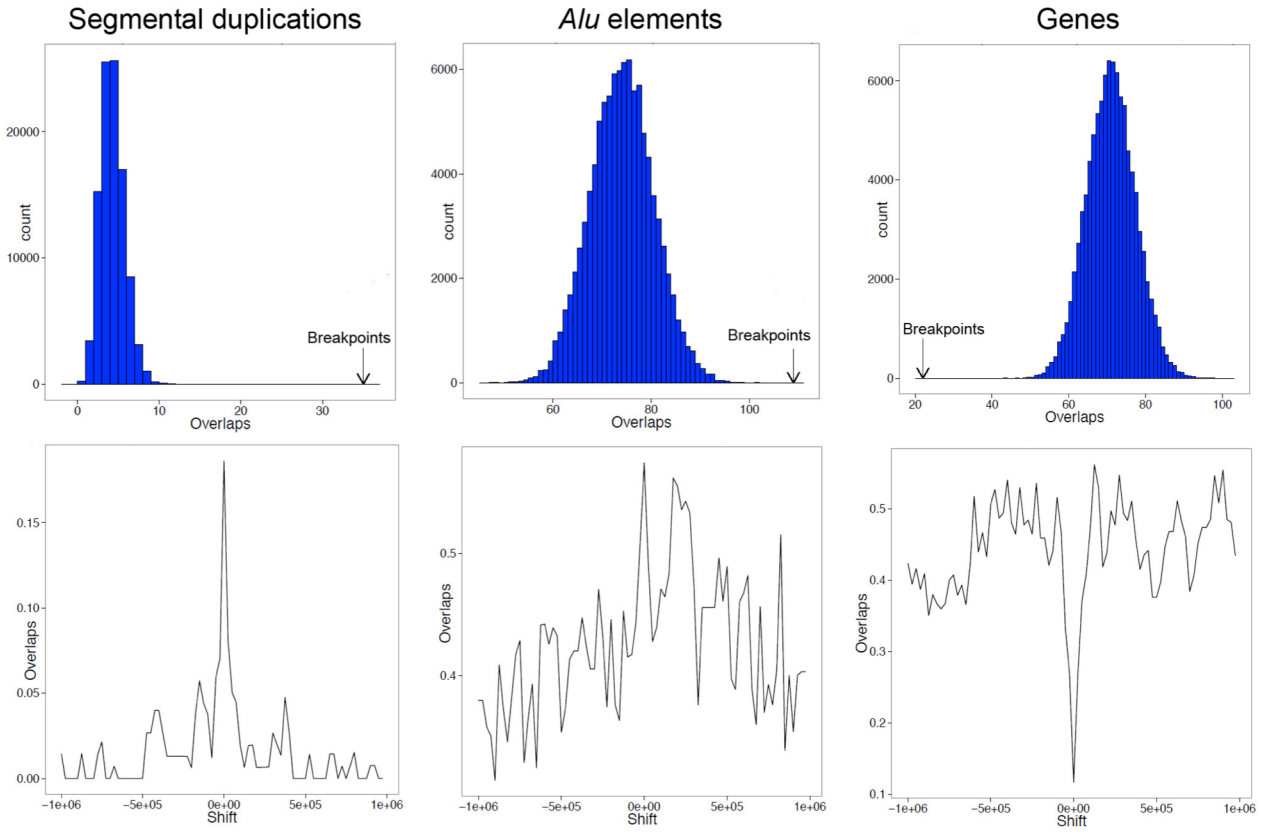


Figure SF5.2 The histograms display the results of permutation analyses used to test for association between breakpoints and genomic features. Breakpoints are enriched in segmental duplications and *Alu* elements and depleted of genes. This enrichment is lost when the breakpoints are shifted from their original position (i.e. 0) as shown by the graphs below the plots.

Repeat Type	D statistic	p-value
All repeats	0.1672	0.1727
Alu	0.3579	2.67e-05
AluJ	0.304	0.0006078
AluS	0.3875	3.86e-06
AluY	0.229	0.0257
LINE	0.1856	0.09792
LTR	0.104	0.731

Table ST5.2 K-S test for the equality of distributions between the distance to the nearest repeat for ClassI breakpoints and 10,000 randomly selected positions

5.2 Overlap with CTCF binding events

Chromatin Immunoprecipitation (ChIP) sequencing for CTCF

CTCF ChIP-seq assays were performed according to Schmidt et al.[42] on eight EBV-transformed lymphoblastoid cell lines established for the same individuals used for the diversity panel (Table ST2.1). In brief, CTCF-bound DNA was immunoprecipitated using an Anti-CTCF rabbit polyclonal antibody (07-729, Millipore, Billerica, MA, USA). End-repair was performed on both immunoprecipitated and input DNA prior to A-tailing and ligation to single-end Illumina sequencing adapters. DNA was amplified using Illumina primers 1.1 and 2.1 in an 18-cycle PCR reaction. Gel electrophoresis was used to select 200-300 bp DNA fragments. DNA libraries were sequenced using 36 bp reads on an Illumina Genome Analyser II according to the manufacturer's instructions.

Peak calling from CTCF ChIP-seq Data

We aligned reads to the Nleu1.0 reference using the BWA aligner (version 0.62)[16] with default parameters, and removed non-uniquely mapping reads. We then called peaks using CCAT[43], with parameters fragmentSize 100, slidingWindowSize 150, movingStep 10, isStrandSensitiveMode 1, minCount 10, minScore 4.0, and bootstrapPass 50. We combined the peaks called across the different individuals and chose the following set for further analysis: any peak called in an individual by CCAT with an FDR of less than 0.05, as well as any peak that was called in more than one individual with an FDR of less than 0.1 (data not shown).

Determination of gibbon-specific and shared CTCF binding events

Gibbon binding events were classified as shared or gibbon-specific based on whether the binding locations are conserved in three other primate species- human, orangutan and rhesus macaque[44]. First, orthologous locations of gibbon CTCF binding events in other primate species were determined using a local installation of the Ensembl Compara multi-species alignment database (http://ensembl.org/info/docs/api/compara/index.html). This database contains alignments of the reference genomes for human (GRCh37), chimpanzee (CHIMP2.1.4), gorilla (gorGor3.1), orangutan (PPYG2), rhesus macaque (MMUL_1), and gibbon (Nleu1.0). A multi-species alignment of each gibbon CTCF binding event region was generated in paml format. Gibbon CTCF binding events for which no multi-alignment was present, or where the nucleotide alignment identity to human and rhesus macaque was less than 70%, were excluded from analysis. We created a non-redundant list of non-gibbon CTCF binding events by converting the genomic coordinates of the human, orangutan and rhesus macaque binding events to gibbon coordinates using the Compara database, and then merging the results using the mergeBed tool from the BEDTools suite to remove redundant entries. Shared and gibbon-specific CTCF binding events were then identified as those gibbon CTCF binding events that did or did not intersect a peak in the non-redundant list of non-gibbon CTCF binding events.

Analysis of CTCF binding events in relation to gibbon-human synteny breakpoints

We identified 52,685 CTCF binding events across the eight gibbon individuals. Because of CTCF's function as an insulator and its known association with the boundaries of DNA topological domains[45], we tested the overlap of CTCF binding events and the gibbon breakpoints. We find 24/96 breakpoint regions with CTCF binding events (one example shown in Fig. SF5.3). Using the same permutation analysis described in section 3 for the SDs and transposable elements, this overlap has an enrichment p-value of 0.0028. This effect was even stronger when we took a ~20 kb window around the breakpoint regions as 84/96 expanded regions overlap CTCF peaks for an enrichment p-value <0.0001.

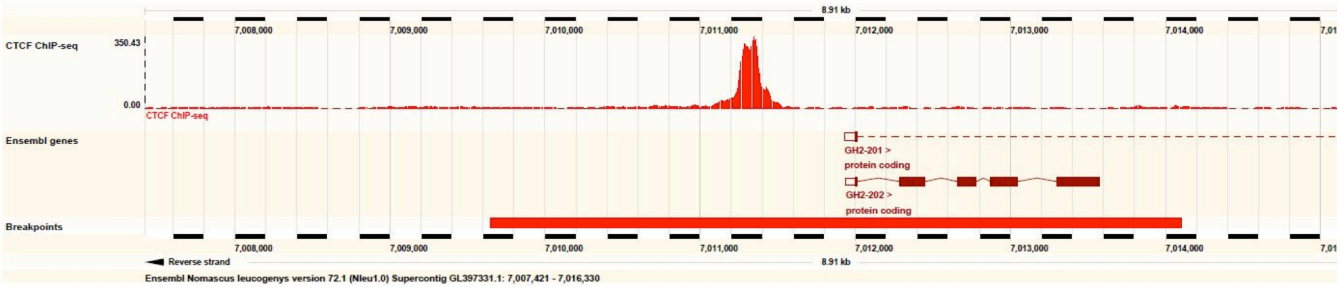


Figure SF 5.3 Screenshot from the Ensembl browser shows an example of a CTCF binding event overlapping gibbon-human synteny breakpoint (red bar).

Additionally, we tested whether the CTCF binding events causing the enrichment in breakpoint regions are specific to gibbons. Using CTCF ChIP-seq data from human, rhesus macaque, and orangutan individuals, we classified CTCF binding events as unique to gibbons (11,449 sites) or shared with a primate ancestor (41,236 sites). We found that the gibbon breakpoint regions are heavily enriched for CTCF binding events shared with a primate ancestor (enrichment p-value = 0.0006) but are not significantly enriched for gibbon-specific CTCF binding events (Fig. SF5.4). Again, the enrichment is stronger in the 20 kb expanded breakpoint regions (enrichment p-value <0.0001) for shared binding sites, but not for gibbon-specific binding events. This suggests that the formation or selection of gibbon genome rearrangements was associated with ancestral CTCF binding events.

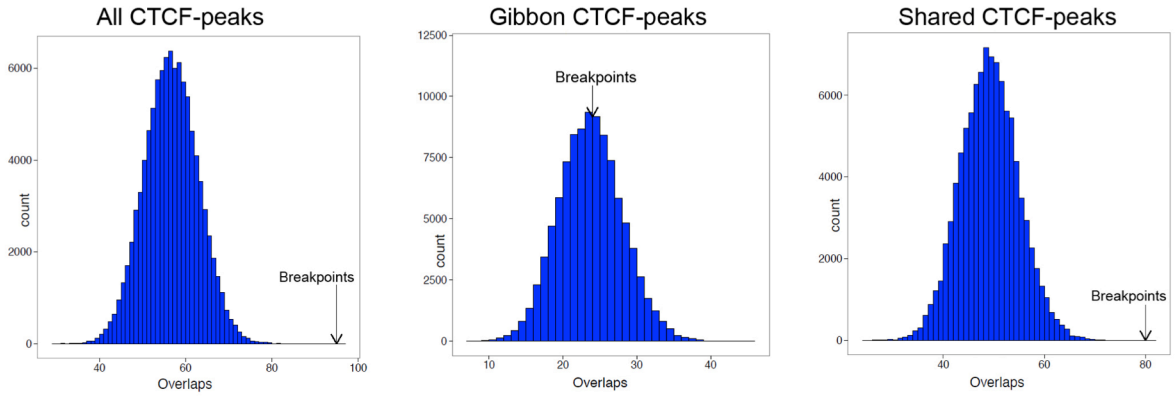


Figure SF5.4 - Gibbon breakpoints are associated with CTCF ChIP-seq peaks (All CTCF peaks). Such association is lost when only gibbon-specific peaks are considered (Gibbon-specific CTCF peaks) whereas it is still present for peaks shared with other primates (Shared CTCF peaks) indicating an evolutionary conserved chromatin structure at breakpoint regions.

Finally, since CTCF-binding events have previously been found to be associated with transposable elements in different species[44], we explored a possible association in gibbon. Moreover, since we are considering only uniquely mapping reads, we wanted to make sure that enrichment in repeated sequences was not preventing us to detect CTCF-binding events overlapping with breakpoints as these regions are enriched in transposable elements. We find that 5,722 out of 52,685 CTCF-binding events overlap an *Alu* and, of these 5,722 peaks, 3 also intersect a breakpoint. The percentage of *Alu*-containing peaks that overlap a breakpoint (3/5,722 = 0.05%) does not seem to be significantly different than the percentage of all peaks that overlap a breakpoint (35/52,685 = 0.07%). Therefore, we believe that we are able to discover peaks that overlap *Alu*s, and that our analysis is sensitive to *Alu*-associated peaks near the breakpoints.

Supplemental Section S6 – The gibbon Ensembl gene set

6.1 Initial Ensembl gene set

Raw Computes Stage: Searching for sequence patterns, aligning proteins and cDNAs to the genome.

The annotation process of the high-coverage Gibbon assembly was done with the Ensembl pipeline and began with the raw compute stage whereby the genomic sequence was screened for sequence patterns including repeats using RepeatMasker[46] (version 3.2.8, run twice, with parameters '-nolow -Gibbon "Nomascus leucogenys" –s' and '-nolow -mammal –s'), Dust[47] and TRF[25] (http://tandem.bu.edu/trf/trf.html). RepeatMasker and Dust combined masked 54% of the Gibbon genome (Fig. SF6.1).

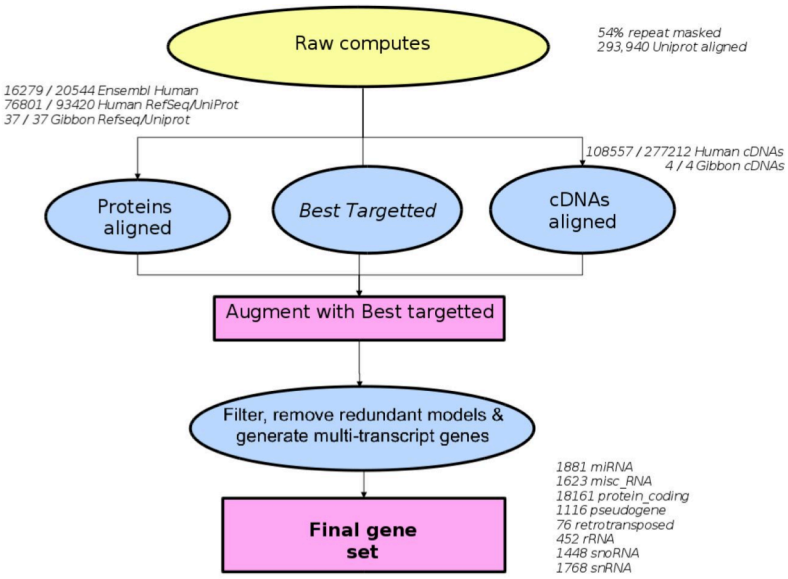


Figure SF6.1 Summary of gibbon genome annotation pipeline

Transcription start sites were predicted using Eponine–scan[48] and FirstEF[49]. CpG islands and tRNAs[50] were also predicted. Genscan[51] was run across RepeatMasked sequence and the results were used as input for UniProt[52], UniGene[53] and Vertebrate RNA (http://www.ebi.ac.uk/ena/) alignments by WU-BLAST[54]. This resulted in 293,940 UniProt, 343,641 UniGene and 336,483 Vertebrate RNA sequences aligning to the genome.

Generating coding models from Human Ensembl Translations

First, Human Ensembl data from e!61 was taken and aligned to the genome using Exonerate[55]. This resulted in 16279 models after cut offs were set at 85% coverage and 80% identity. Additionally, 'mid-ranged' models as low as 50% coverage and identity were taken where they matched a best targetted entry by intronic regions and the best targetted model had a translation of ≥50 amino acids.

Generating a supportive evidence coding model set from Human and Gibbon proteins

Gibbon and Human protein sequences were downloaded from public databases (UniProt SwissProt/TrEMBL and RefSeq[53]). The Gibbon and Human protein sequences were mapped to the genome using Pmatch as indicated in Fig. SF6.2. Models of the coding sequence (CDS) were produced from the proteins using Genewise[56] and Exonerate[55]. Where one protein sequence had generated more than one coding model at a locus, the BestTargetted module was used to select the coding model that most closely matched the source protein to take through to the next stage of the gene annotation process. The generation of transcript models using Gibbon-specific (in this case Gibbon and Human) data is referred to as the “Targetted stage”. This stage resulted in 99069 coding models and was used as evidence for alignments of “midranged” matches from the Human Ensembl exonerate alignments mentioned in the previous stage.

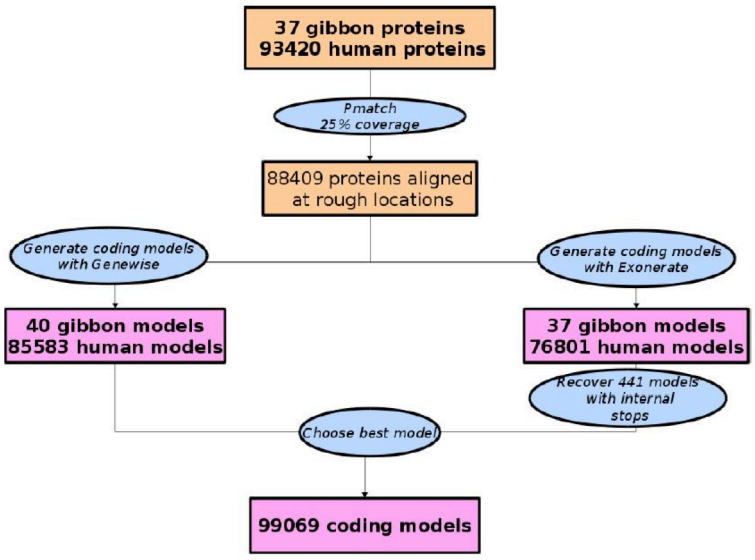


Figure SF6.2 Targetted stage using Gibbon protein sequences

Recovery of internal stop entries

The exonerate alignments can produce transcript models with stop codons, which cannot be used in the final gene set because the GeneBuilder module removes models which include internal stops. For models with only one stop we attempt to replace the stops with small introns where they lie in the middle of the exon. For models with more than one stop attempts to get a better alignment are then made on the region using exonerate in 'exhaustive' mode.

cDNA Alignments

Gibbon and Human cDNAs were downloaded from ENA/Genbank/DDBJ, clipped to remove polyA tails, and aligned to the genome using Exonerate (Fig. SF6.3). Of these, 108,557 (of 277,212) Human cDNAs aligned and 4 (of 4) Gibbon cDNAs aligned. Human alignments were at a cut-off of 90% coverage and 90% identity.

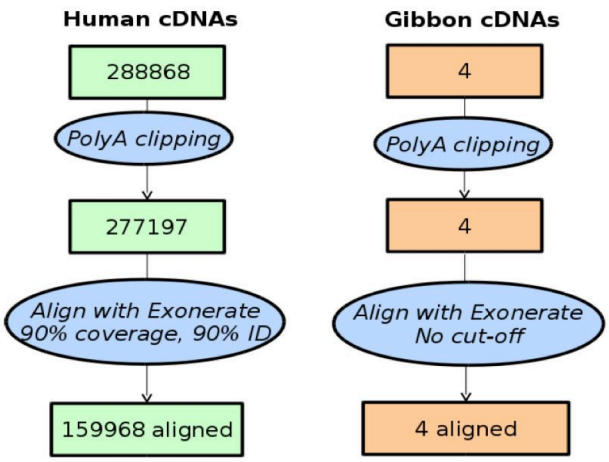


Figure SF6.3 - Alignment of Gibbon and Human cDNAs

Addition of UTR to coding models

The set of coding models was extended into the untranslated regions (UTRs) using Human and Gibbon cDNA sequences. This resulted in 19 (of 26) Gibbon coding models with UTR and 21427 (of 77888) Human coding models with UTR.

Generating multi-transcript genes

The above steps generated a large set of potential transcript models, many of which overlapped one another. Redundant transcript models were removed and the remaining unique set of transcript models were clustered into multi-transcript genes where each transcript in a gene has at least one coding exon that overlaps a coding exon from another transcript within the same gene. The final set of 19,461 coding genes included 13 genes with at least one transcript supported by Gibbon proteins with the remaining having at least one transcript supported by Human evidence (Fig. SF6.4)

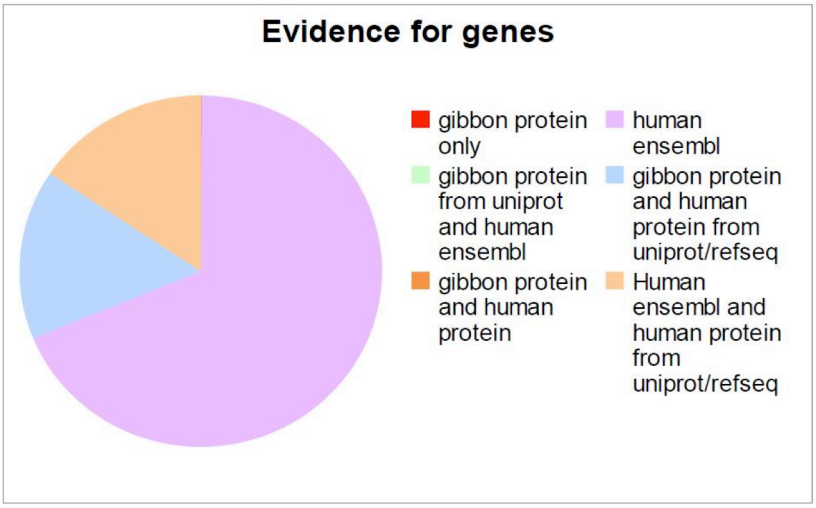


Figure SF6.4 - Supporting evidence for Gibbon final gene set

The final transcript set of 24554 transcripts included 18 transcripts with support from gibbon proteins, 13190 transcripts with support from Human Ensembl proteins and 19,974 transcripts with support from UniProt/RefSeq.

Pseudogenes, Protein annotation, Cross-referencing, Stable Identifiers

The gene set was screened for potential pseudogenes. Before public release the transcripts and translations were given external references (crossreferences to external databases), while translations were searched for domains/signatures of interest and labeled where appropriate. Stable identifiers were assigned to each gene, transcript, exon and translation. (When annotating a species for the first time,

these identifiers are auto-generated. In all subsequent annotations for Gibbon, the stable identifiers will be propagated based on comparison of the new gene set to the previous gene set).

6.2 Ensembl gene set update

Making the RNAseq-based gene models from transcriptome data

RNAseq-based gene models for gibbon were produced by running the Ensembl RNAseq pipeline with gibbon lymphoblastoid transcriptome data from Asia provided by the Gibbon Consortium.

The RNAseq-based gene models were produced in a two-step alignment process. First, raw reads were aligned to the gibbon genome using BWA[16]. These alignments were collapsed to create alignment blocks roughly corresponding to transcribed exons. Read pairing information was used to group exons into approximate transcript structures called proto-transcripts. In the second alignment step, any reads that were partially aligned by BWA were extracted and aligned to the proto-transcripts using Exonerate[55]. Exonerate is splice-aware and the spliced alignments from this step allowed us to identify exon-intron junctions and therefore the location of introns. The result of the Exonerate alignment step was a set of spliced alignments representing canonical and non-canonical introns. Transcript models were created by combining the proto-transcripts with the spliced alignments to create all possible transcript isoforms. Our pipeline was configured to keep only the isoform with the most read support.

The RNAseq pipeline described above produced transcript models that could be protein coding or non-coding protein. We therefore ran BLAST of UniProt PE1 and PE2 proteins against the set of RNAseq models, in an attempt to identify those transcript models that are protein coding. Those models that aligned to UniProt sequences were considered to be protein coding models and they were used as input for the RNAseq update pipeline.

Updating the Ensembl gibbon gene set using transcriptome data

The RNAseq update pipeline took as input the existing Ensembl gene set on gibbon Nleu1.0/nomLeu1 and the RNAseq-based models produced by our RNAseq pipeline (as described above). The two sets of input models were compared and merged to produce an updated gene set. The update pipeline

allowed truncated genes to be lengthened, adjacent gene fragments to be merged, and artificially merged genes to be split.

In addition to the updated Ensembl gene set, we have released the RNAseq-based gene models that were used as input in the update pipeline. These RNAseq gene models include only the best supported transcript model for each gene. For some genes, additional splice junctions (introns) may have been identified but not represented in the best supported transcript model. We therefore also provide users with the full set of introns identified by our RNAseq pipeline.

6.3 Coding exon assessment

Correct gene models are needed in order to be able to investigate presence of positive selection and identify possible biases due to gene model errors. We assessed the complete set of coding exons in the gibbon genome by checking how they align with coding exons in the human genome. One would expect a large degree of agreement between the human and the gibbon gene models if both genes have been correctly predicted. Miss-annotations in the gibbon genome can happen because of assembly gaps and can be detected when comparing them to the human model (Fig. SF6.5). In addition, this approach also allowed us to indirectly assess the quality of the predicted gene models.

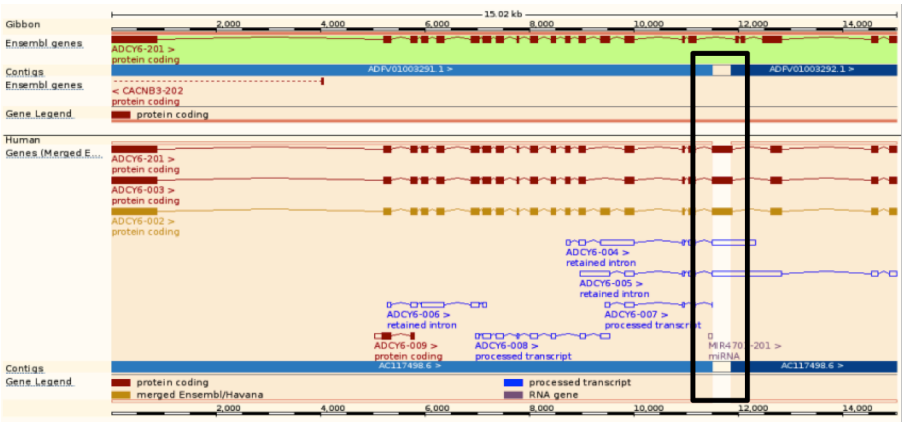


Figure SF6.5 The top panel shows the gibbon locus for the ADCY6 gene. The bottom panel shows how the human genome aligns to the gibbon sequence. Note how the 17th exon of the human gene model (red box) maps on an assembly gap (blue track) of the gibbon genome. As a result, the modeling software has created two artificial exons downstream of the assembly gap to represent the whole protein sequence.

For each gene in the gibbon genome we focused on the transcript used to build the GeneTrees (Beal et al, in prep.) in Ensembl 70 (http://e70.ensembl.org). We tested each gibbon gene against every transcript of every human orthologous gene. In each case, we look at how well each exon of the gibbon gene model aligns with the human orthologous sequence. The alignments are extracted from the whole-genome LASTz or BLASTz alignments available in Ensembl 70 (Beal et al, in prep.). It is important to note that the gene models were not used at any stage for building the pairwise alignments.

Methods

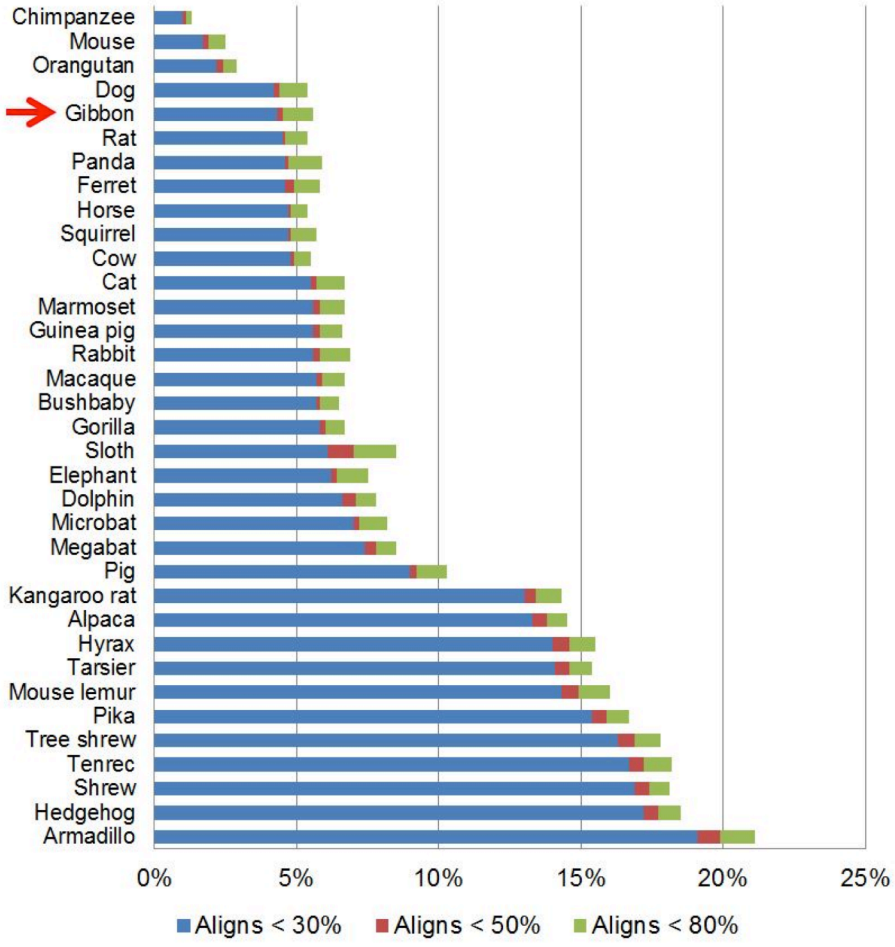
Gibbon genes for which there is no human orthologous or there is no pairwise alignment were ignored. All other gibbon exons were assigned a score that represents the ratio between the length of the aligned portion of the human exon and the length of the gibbon exon itself. Based on the average exon score, we can define the best matching human transcript. Using the best matching transcripts only, we can assess the confidence for each coding exon.

In addition, we looked at the percentage of exons for which less than 30%, less than 50% or less than 80% of the sequence aligns with a human orthologous exon. Fig. SF6.6 shows the results for the gibbon genome and all other placental mammalian species in Ensembl 70 (http://e70.ensembl.org). The gibbon genome appears in the 5th place with about 5% of the exons largely unaligned to human coding sequence.

Results

Genomic alignments can be used to build a list of suspicious gibbon exons defined as the ones that do not align to the coding sequence of their human counterpart. About 8.4% of the transcripts were not analyzed because they either did not have an orthologous gene in human (6.9%) or no genomic alignment could be used for the test (1.5%). We were able to filter about 5% of the remaining 177,384 exons. It is worth noting that we expect a bias in favor of the primate genomes. One of the reasons is

that all comparisons are made against the human genome. The primate genes are evolutionary closer to the human ones, hence they should align much better. In addition, human proteins are normally used to annotate primate genomes, increasing this effect. These considerations should be taken into account if one wants to use this method to rank the quality mammalian genome annotations.



SF

Figure SF6.6 The percentage of coding exons in mammalian genomes for which less than 30%, 50% or 80% of their sequence aligns to human coding exons.

Supplemental Section S7 – The LAVA element

7.1 Characterization of LAVA elements in Nleu1.0 / nomLeu1

Sequence retrieval and analysis

LAVA elements in Nleu1.0 were identified and retrieved by BLAST search using the 3' part (U1-*Alu*Sz-U2-L1ME5) of the published LAVA consensus sequence[57] as query. The set of potential LAVA elements retrieved was manually curated resulting in a final set of 1,797 non-redundant LAVA sequences. For 1,583 out of the 1,797 elements the 5' end could be established (no assembly gaps). Out of these 760 (48%) were found to be full-length, i.e. they contain 5' CT hexameric repeats, the *Alu*-like domain, VNTR region and the 3' U1-AluSz-U2-L1ME5.

The set of 1,797 LAVA elements was validated using RepeatMasker analysis with the human repeat library. For this analysis, we added 5 kb of flanking sequence up- and downstream of each LAVA position. Next, we categorized each locus using a series of "in-house" Perl scripts that parsed the resulting repeat annotations for each locus. A 3'-complete LAVA locus in this analysis appears as a characteristic arrangement of three components: an SVA followed by an *Alu*Sz element followed by an L1ME5 fragment in the opposite orientation. 1,256 of the elements in our data set met this requirement and were used in the subsequent analyses of LAVA genomic distribution (see below) (Supplementary File 6).

LAVA subfamily and phylogenetic analysis

Initial manual subfamily determinations were conducted on an early subset of elements that were aligned using BioEdit. Six initial subfamily groups (A through F) were suggested, and in some cases several subfamilies within these were identified. We then used Coseg (http://www.repeatmasker.org/COSEGDownload.html) to perform a subfamily analysis of 1,097 validated complete LAVA elements that also had no internal non-LAVA repeat insertions and no Ns. Due to the highly repetitive VNTR section in the 5' half of LAVA elements, we confined our analysis to the 3' half of the elements. A consensus sequence consisting of the U1 through L1 portion of LAVA was produced

using a majority rule approach, and we used cross_match (http://www.phrap.org/phredphrapconsed.html) to align this consensus sequence to the LAVA elements prior to the Coseg analysis. We allowed Coseg to call a subfamily if it could find significant co-segregating mutations (2-3 bp) shared between ≥10 elements. This produced evidence of 16 subfamilies, whose relationships were in general agreement with the six subfamily groups suggested by our initial manual inspection. Further manual annotation of these 16 groups using alignments in BioEdit allowed us to support a further 6 small subfamilies discernible within some of the larger subfamilies but missed by Coseg. In total, our analysis identified 22 LAVA subfamilies (LAVA_A1, LAVA_A2, LAVA_B1B, LAVA_B1D, LAVA_B1F1, LAVA_B1F2, LAVA_B1G, LAVA_B1R1, LAVA_B1R2, LAVA_B2A, LAVA_B2C, LAVA_B2R1, LAVA_B2R2, LAVA_C2, LAVA_C4A, LAVA_C4B, LAVA_D1, LAVA_D2, LAVA_E, LAVA_F0, LAVA_F1, and LAVA_F2). This subfamily taxonomy was included in the latest RepeatMasker library (http://www.girinst.org/server/RepBase/index.php). A network analysis showing the relationships between the consensus sequences of these 22 subfamilies, as well as the number of elements in each subfamily, is shown in Fig. 3-C. This network was generated using the application, Network version 4.611[58]. An alignment of the sequences was first performed using the alignment algorithm within the Geneious version 5.6.5 (http://www.geneious.com/) and exported as a multi-fasta file. This was then converted into the Rohl Data File (.rdf) format using DNAsp version 5.10.1[59]. The .rdf file was then imported into Network and a median-joining analysis was performed.

7.2 LAVA element PCR validation and phylogenetic analysis

Based on the subfamily structure of LAVA elements we designed oligonucleotide primers for locus specific PCR of 200 LAVA insertions representing all known LAVA subfamilies, with an emphasis focus on the youngest appearing elements from each subfamily based on low divergence from their respective subfamily consensus sequences (Extended Data Fig. 4). Each locus required two PCR reactions for genotyping, one using locus-specific PCR primers which flanked the predicted insertion coordinates, and a second using an internal forward primer (designed inside the LAVA element

consensus sequence) and the flanking reverse primer. A PCR product in the first reaction amplified the empty site, while a product in the second reaction confirmed the presence of a LAVA element insertion. In some cases, if the LAVA element was truncated, the flanking primers amplified a complete filled site sequence. Primer3 (http://bioinfo.ut.ee/primer3/) was used for primer design. Each primer was checked against the multi-alignment to ensure a high likelihood for amplification in human, chimpanzee, gorilla and orangutan as well as all gibbon species. In addition, each primer was aligned by BLAT[60] against the human (GRChr37) and gibbon (Nleu1.0) reference genomes to confirm the uniqueness of the primers. If necessary, alternative primers were selected and tested. An *in-silico* PCR was performed for each primer combination a) to confirm that only one amplicon was predicted, and b) to determine the size of the predicted filled (LAVA present) and empty (insertion absent) PCR products. Oligonucleotide primers were obtained from Sigma Aldrich (Woodlands, TX) and all primer sequences are available on the Batzer Lab website (http://batzerlab.biology.lsu.edu/) and as Supplementary File 6. PCR reactions were performed using standard procedures and amplification products were checked using gel electrophoresis. The DNA panel used for these PCR experiments included 17 individual gibbons representing 13 species and all four gibbon genera. Out-group species included human, chimpanzee, gorilla, orangutan, and African green monkey (Table ST7.1). Sanger sequencing was used to confirm PCR results if necessary.

Of the 200 LAVA candidate loci selected for PCR validations, 176 were successfully genotyped with over half (52%; N=92) being Hylobatidae specific; i.e. being shared across all four gibbon genera (*Nomascus*, *Hoolock*, *Hylobates* and *Symphalangus*). Another 27% (N=48) were specific to the *Nomascus* lineage. The remaining 36 LAVA insertions used for our phylogenetic analyses showed varying degrees of incomplete lineage sorting, or specificity to either *N. leucogenys* or *N. leucogenys* and *N. siki*. A summary of the LAVA phylogenetic results is shown in Table ST7.2. All inconclusive or contradictory genotype results were further analyzed by Sanger sequencing to confirm either the presence of a shared LAVA insertion or to confirm a precise pre-insertion site (absence of the insertion

as opposed to a deletion). A total of 160 PCR products were Sanger sequenced in both directions for a total of 320 sequencing reactions. All LAVA insertions were confirmed as shared insertions and no evidence was found for deletion events.

Gel Well	ID	Species Name	Common Name	Source
1	100bp ladder			
2	empty			
3	TLE		Negative control	
4	CCL2	*Homo sapiens*	Human (HeLa)	ATCC
5	NS06006	*Pan troglodytes*	Chimpanzee	IPBIR
6	AG05251	*Gorilla gorilla*	Gorilla	Coriell
7	GM06213A	*Pongo abelii*	Sumatran orangutan	Coriell
8	NLL605	*Nomascus leucogenys*	Northern white-cheeked gibbon	GCC
9	1232	*Nomascus siki*	Southern white-cheeked gibbon	Christian Roos
10	PR00652	*Nomascus gabriellae*	Southern yellow-cheeked gibbon	Coriell
11	990838	*Nomascus gabriellae*	Southern yellow-cheeked gibbon	Los Angeles Zoo
12	P109	*Nomascus annamensis*	Northern yellow-cheeked gibbon	Christian Roos
13	9087	*Hylobates lar*	White-handed gibbon	Gladys Porter Zoo
14	PR00715	*Hylobates lar*	White-handed gibbon	Coriell
15	98274	*Hylobates moloch*	Javan gibbon	Fort Wayne Children's Zoo
16	15353	*Hylobates agilis*	Agile gibbon	Henry Doorly Zoo
17	212067	*Hylobates albibarbis*	Bornean white-bearded gibbon	Louisiana Purchase Zoo
18	8136	*Hylobates muelleri*	Mueller's Borneo gibbon	Gladys Porter Zoo
19	1225	*Hylobates muelleri*	Mueller's Borneo gibbon	Christian Roos
20	8097	*Hylobates pileatus*	Pileated gibbon	GCC
21	1230	*Hylobates klossii*	Kloss's gibbon	Christian Roos
22	HH305	*Hoolock leuconedys*	Eastern hoolock gibbon	GCC
23	SS901	*Symphalangus syndactylus*	Siamang	GCC
24	KB 11539	*Symphalangus syndactylus*	Siamang	SDFZ
25	CCL70	*Chlorocebus aethiops*	African Green Monkey	ATCC

Table ST7.1 Panel of DNAs used for investigating LAVA element Phylogeny.

Furthermore, we did not encounter any PCR products that were the same size as the targeted LAVA amplicon that were caused by another mobile element insertion (including another LAVA element) across all loci analyzed. Thus, we did not find evidence for (near) parallel insertions that can rarely occur in phylogenetic analyses using mobile element markers[61,62]. We did, however, find evidence for a few mobile element insertions that resulted in amplicons with unexpected sizes. In all cases, sequencing revealed that the size disparity was caused by an *Alu* insertion that occurred within the flanking unique sequence of the selected candidate loci (within the amplicon). One explanation for the lack of (near) parallel insertion events with amplicon sizes similar to the filled LAVA amplicon could be

Description of PCR Results for LAVA Phylogeny	Number of Loci
Selected loci (young elements of SFS* represented)	**200**
Able to genotype	176
Hylobatidae specific (shared among all 4 genera)	92
Nomascus specific (shared among *Nomascus* species)	48
Nomascus & *Symphalangus* specific	3
Nomascus & *Hoolock* specific	2
Nomascus, *Hoolock* & *Symphalangus* specific	4
Nomascus, *Hylobates* & *Symphalangus* specific	3
General incomplete lineage sorting within gibbon genera	6
Shared by *Nomascus leucogenys* and *N. siki* only	1
Polymorphic within *Nomascus* species	10
Specific to *Nomascus leucogenys* & polymorphic (young)	7

Table ST7.2 PCR Results for LAVA Element Phylogeny

the relatively low copy number of LAVA elements in gibbons in conjunction with a relatively slow retrotransposition rate.

Altogether, 12 loci provided evidence for incomplete lineage sorting of ancestral polymorphic loci among the four gibbon genera (Table ST7.2 highlighted in orange). A closer relationship of *Nomascus* and *Symphalangus* compared to *Hoolock* and *Hylobates* was indicated based on presence/absence PCR data confirmed by sequencing of three LAVA insertions (i.e. *Nomascus* and *Symphalangus* shared a LAVA insertion that was absent from *Hoolock* and *Hylobates*). Two LAVA insertions suggest that *Hoolock* is most closely related to *Nomascus*. In addition our analyses revealed four insertions that group *Nomascus*, *Hoolock*, and *Symphalangus* most closely together while three insertions provide support for a closer relationship of *Nomascus*, *Hylobates*, and *Symphalangus*. An additional six loci provided evidence for lineage sorting between as well as within gibbon genera, meaning that the LAVA insertion was detected in all tested individuals of at least one genus and also in some (but not all individuals of another genus. The identification of this number of lineage sorting candidate loci suggests rapid speciation and/or extended hybridization events across gibbon genera. Within the *Nomascus* genera, we observed one locus shared by *N. leucogenys* and *N. siki* while being absent from the other *Nomascus* species on our panel. However, we observed another 10 LAVA loci that were polymorphic

among the *Nomascus* individuals analyzed and thus not phylogenetically informative. This provides further evidence for possible hybridization events across gibbons.

The phylogenetic analyses also revealed evidence for the ongoing retrotransposition of LAVA elements. All LAVA insertions specific to *N. leucogenys* were tested on a population panel that included 6 *N. leucogenys* individuals. We detected 7 LAVA loci specific to *N. leucogenys*, 4 being polymorphic among *N. leucogenys* individuals and 3 that were unique to only one *N. leucogenys* individual based on our analyses, suggesting relatively recent insertion events.

7.3 LAVA subfamily age estimates

We used the maximum likelihood method for estimating the age of retrotransposon subfamilies described by Marchani et al.[63] to provide age estimates for the 22 LAVA subfamilies. The dataset consisting of 1,256 3'-intact LAVA elements used in our other analyses was divided into 22 subfamily-specific datasets based on RepeatMasker annotations. All sequences of the member elements for each subfamily were then aligned to their respective subfamily consensus sequences, which had first been trimmed to include only the 3' end of the LAVA element consisting of the U1-to-L1 portion of the LAVA. The SVA-derived portion of the LAVA element contains a stretch of VNTR sequences that are quite difficult to align properly, and, as they are highly variable, would skew an age estimate; we therefore chose to exclude this portion from our age analysis and focus on the 3' end of the element.

The alignments were generated with the ClustalW implementation in MEGA 5.22 using default parameters. First, we manually trimmed of any sequence aligning to the 5' SVA-derived portion of the member elements as well as any sequence from the 3' poly-A tails. Next, we cleaned the alignments by deleting any insertions in member elements relative to their consensus sequences as well as any clustered substitutions, inversions, and other mutations not resulting from a single base misincorporation. Then, we manually aligned sections of sequence that were sometimes poorly placed at the margins of gaps in member elements by the alignment algorithm. Finally, in three cases elements were removed entirely from the alignment because the internal rearrangements were deemed too

extensive to leave sufficient aligned sequence after cleaning. The resulting alignments consisted of ungapped consensus sequences and trimmed, cleaned member elements. These were separated into consensus sequence and member element files which were then used to calculate the maximum likelihood estimate (T) and sample standard deviation (σ) for each subfamily. By dividing T for each subfamily by an appropriate DNA mutation rate, age estimates in years were calculated. The σ values and number of member elements (n) in each subfamily were used to calculate 95% confidence intervals around these age estimates.

The question of mutation rate (μ) deserves special attention, as small changes in μ can result in large differences in the final age estimate. We therefore show age estimates based on several values of μ in Table ST7.3. A common mutation rate used for apes that was estimated from human-chimp sequence divergence at pseudogenes is 1.05×10^{-8} substitutions per site per generation, or 0.105% per million years[63,64]. However, some studies of retrotransposon sequences have shown rates greater than the genome average. Since no studies specifically detailing the LAVA-specific rates have yet been performed, and because the portion of the LAVA element used to obtain the estimates consists largely of Alu- and L1-derived sequences, we argue that it is reasonable to use the higher mutation rates that have been associated with these elements in previous analyses. These rates are 0.23% and 0.25%[64] [65,66]. The maximum likelihood age estimates with 95% confidence intervals under these three values of μ can be found in Table ST7.3.

LAVA Subfamily	Members (n)	Estimate (T)	Std. Dev. (σ)	μ=0.105%			μ=0.230%			μ=0.250%		
				Lower 95%	Mean	Upper 95%	Lower 95%	Mean	Upper 95%	Lower 95%	Mean	Upper 95%
A1	26	0.0429	1.0911E-05	40.853	40.857	40.861	18.650	18.652	18.654	17.158	17.160	17.162
A2	28	0.0420	1.0411E-05	39.996	40.000	40.004	18.259	18.261	18.263	16.798	16.800	16.802
B1B	26	0.0410	9.3264E-06	39.044	39.048	39.051	17.825	17.826	17.828	16.399	16.400	16.401
B1D	37	0.0443	7.5806E-06	42.188	42.190	42.193	19.260	19.261	19.262	17.719	17.720	17.721
B1F1	10	0.0334	1.9065E-05	31.798	31.810	31.821	14.517	14.522	14.527	13.355	13.360	13.365
B1F2	77	0.0479	4.6454E-06	45.618	45.619	45.620	20.826	20.826	20.827	19.160	19.160	19.160
B1G	80	0.0457	4.4711E-06	43.523	43.524	43.525	19.869	19.870	19.870	18.280	18.280	18.280
B1R1	31	0.0388	7.3999E-06	36.950	36.952	36.955	16.868	16.870	16.871	15.519	15.520	15.521
B1R2	101	0.0446	2.9800E-06	42.476	42.476	42.477	19.391	19.391	19.392	17.840	17.840	17.840
B2A	55	0.0381	4.7487E-06	36.285	36.286	36.287	16.565	16.565	16.566	15.239	15.240	15.241
B2C	29	0.0486	1.8055E-05	46.279	46.286	46.292	21.128	21.130	21.133	19.437	19.440	19.443
B2R1	41	0.0316	4.0059E-06	30.094	30.095	30.096	13.739	13.739	13.740	12.640	12.640	12.640
B2R2	107	0.0389	2.3255E-06	37.047	37.048	37.048	16.913	16.913	16.913	15.560	15.560	15.560
C2	38	0.0419	9.9353E-06	39.902	39.905	39.908	18.216	18.217	18.219	16.759	16.760	16.761
C4A	52	0.0469	7.8931E-06	44.665	44.667	44.669	20.390	20.391	20.392	18.759	18.760	18.761
C4B	86	?	?	?	?	?	?	?	?	?	?	?
D1	48	0.0352	5.0153E-06	33.522	33.524	33.525	15.304	15.304	15.305	14.079	14.080	14.081
D2	11	0.0345	2.0568E-05	32.846	32.857	32.869	14.995	15.000	15.005	13.795	13.800	13.805
E	168	0.0269	9.3022E-07	25.619	25.619	25.619	11.696	11.696	11.696	10.760	10.760	10.760
F0	33	0.0205	2.6014E-06	19.523	19.524	19.525	8.913	8.913	8.913	8.200	8.200	8.200
F1	99	0.0169	6.5810E-07	16.095	16.095	16.095	7.348	7.348	7.348	6.760	6.760	6.760
F2	69	0.0148	7.4227E-07	14.095	14.095	14.095	6.435	6.435	6.435	5.920	5.920	5.920

Table ST7.3 Different age estimates for LAVA subfamilies using three different values for the mutation rate (μ). [Subfamily C4B failed the analysis and could not be run]

Given that LAVA elements are gibbon specific and the divergence between gibbons and other hominoids is ~16.8 mya (as described in Section 4), we would expect this divergence time to set an upper bound on the age the basal LAVA subfamilies, LAVA_A1 and LAVA_A2. Age estimates for these two subfamilies under μ = 0.105%, however, are much too old, with means of 42.9 and 42.0 million years, respectively. Alternatively, if the *Alu*-specific rate of μ = 0.23% is used those estimates drop to 18.652 and 18.261 million years, while under the L1-specific rate of μ = 0.25% the estimates further drop to 17.160 and 16.800 million years. Further, we estimate the age of the youngest subfamilies LAVA_F1 to be 7.348 and 6.76 million years and the age of LAVA_F2 to be 6.435 and 5.920 million years, respectively. These retrotransposon-specific mutation rates provide estimates that much more closely match the diversification of gibbons from other apes as well as the radiation within gibbons of the four extant gibbon genera. We therefore argue that the age estimates obtained using these rates are likely to be more accurate than those obtained under the background mutation rate of the genome.

7.4 Analysis of LAVA insertions into genes and GO term analysis

Gibbon repeats identified in the Nleu1.0 assembly were intersected with gene predictions based on Ensembl release e70. The set of 1,256 3'-intact LAVA elements were intersected along with *Alu* subfamilies, L1 and L2 identified by RepeatMasker[67]. In addition, human (GRCh37) Ensembl release e70 gene predictions were intersected with *Alu* subfamilies, L1, L2 and SVA identified by RepeatMasker. Intersections were performed using both the gene body and the gene body +/- 5 kbp upstream and downstream. Intersections with the gene body were classified as either involving an exon or intron only (Supplementary File 6).

For LAVA elements, permutation testing was performed to examine the significance of intersections with gene bodies, exons and introns. Genomic loci matched to the lengths of the 1,256 LAVA elements were randomly selected in 1,000 iterations. Genomic loci intersecting with assembly gaps were removed from consideration. Pearson's chi-squared tests were performed on the observed intersections of LAVA with gene elements compared to the expected intersections based on the average number of intersections across the 1000 iterations (Extended Data Fig. 3).

Genes that intersected with LAVA, *Alu* elements, L1, L2, or SVA were analyzed for enrichment of Gene Ontology (GO) terms and tissues with expression using the Database for Annotation, Visualization and Integrated Discovery (DAVID)[20]. GO enrichment analyses used GOTERM_BP_FAT, GOTERM_CC_FAT, and GOTERM_MF_FAT annotations and tissues with expression enrichment used UNIGENE_EST_QUARTILE annotations. Analyses were run using human genes as the background. In order to further examine the GO enrichment of LAVA insertions, the 1,000 permutation iterations were ran through DAVID and the number of times a GO term was enriched at FDR <0.05 was calculated (Supplementary File 6).

Using enrichment and simulation analyses similar to the ones used for LAVA insertions, we discovered that genes from the 'microtubule cytoskeleton' categories are also enriched for *Alu*S and *Alu*Y, but not for L1, LTR and SVA elements in both gibbon and human (Supplementary File 6). *Alu* elements, however, lack a termination signal and therefore cannot cause early termination of

transcription. On the other hand, this observation suggests a tendency from this group of genes to be targeted by retrotransposons.

7.5 Analysis of LAVA elements inserted into genes-of-interest

In order to better understand the timing and potential impact of LAVA insertions into genes-of-interest, we analyzed 24 LAVA insertions in genes from the 'microtubule cytoskeleton' GO category. Subfamily attribution and percent divergence from subfamily consensus sequences were obtained from RepeatMasker for each LAVA element; we also annotated presence of an antisense polyadenylation signal near the 3' end of each LAVA element and found that 50% (12/24) of the elements have a perfect TTTATT antisense polyadenylation signal (Supplementary File 6 and Extended Data Table 1).

We assessed the orthology of these insertions among gibbon species using both *in silico* and experimental methods. The *in silico* approach involved the mapping and visualization of the Illumina reads from nine gibbon individuals (8 from the diversity panel and the reference) with the Integrative Genomics Viewer (IGV)[68]. All datasets for each *Nomascus* individual were merged into a single bam file for each individual using samtools[69]. Datasets for the *Hoolock*, *Hylobates*, and *Symphalangus* individuals were merged into genus-specific bam files. These bam files were filtered based on quality score at a variety of cutoffs (q1, q20, q35, and q50) and were further filtered to include only read pairs with at least one mate mapping within one of the 24 genes-of-interest +/- 750 bp of flanking. The final orthology calls were made using the q50 filtered dataset. We RepeatMasked the 24 regions to annotate their repeat content and IGV batch files were then used to generate screenshots for each region. These images show the coverage in each *Nomascus* individual, the coverage in the other genera, RepeatMasker annotations of the 750 bp flanking regions, and the LAVA element itself (Extended Data Fig. 4). Each image was visually inspected and a locus was putatively annotated as "shared" between genera (and therefore considered ancestral) if read pairs mapping was consistent between the four genera (i.e. one mate inside the LAVA and the other inside one of the flanking regions). On the other hand, loci were putatively annotated as "*Nomascus*-specific" if *the Nomascus* individuals showed reads

as above, but non-*Nomascus* datasets showed substantial coverage in the flanking regions but no read pairs spanning the margins of the LAVA element. Instead, read pairs were found to map in the flanking regions with a larger-than-expected insert size.

Independently from the *in silico* orthology calls, we designed PCR primers for each of the 24 LAVA insertions using Nleu1.0 sequences as template. The software "Primer3" was used to design all PCR primers. For each LAVA insertion, one internal primer was designed near the 3' end of the element and one primer was designed in one of the flanking regions. In some case, two flanking primers were designed to amplify the empty allele and confirm results from the internal primers. PCR was performed by pairing the internal primer with the 3' flanking primer; the *in silico* PCR tool from the UCSC Genome Browser was used to verify each primer pair and estimate amplicon sizes. In the case of internal PCRs, amplicons are expected only for species in which the LAVA element is present, while no amplification is expected in species lacking the LAVA insertion. All PCRs were performed using standard methods on a panel of four genomic DNAs from gibbon individuals representing the four genera (Asia for *Nomascus*, Domino for *Hylobates*, Karenina for *Symphalangus*, and Drew for *Hoolock*, see Table ST2.1). Amplicon sizes were visualized by gel electrophoresis. PCR orthology was called based on these results and compared with the orthology calls from the IGV analysis. In some cases one of the assessment methods (IGV visualization, internal PCR, or flanking PCR) failed. Final orthology calls were made for each locus based on agreement between at least two of the assessments and in no cases did any two orthology assessments disagree at a locus. The orthology calls, PCR primers, and estimated sizes are summarized in Supplementary File 6.

7.6 Analysis of distance from the nearest exon

Of the 481 full-length LAVA elements that have inserted into introns, 124 (25.78%) are in the same orientation as the gene and 357 (74.22%) are antisense. This suggests selection pressure against sense insertions which has been documented for other transposable elements in the human and mouse genomes[70]. We tested the possibility of the presence of selective pressure to weed out LAVA

elements that insert too close to exons. Following the procedure outlined in Zhang et al.[71] we calculated the distance to the nearest exon for each intronic LAVA and compared this to what would be expected for random insertions. We found fewer insertions than would be expected by chance within 1 Kbp of the nearest exon (Extended Data Fig. 3). Specifically we considered distance bins of 1-50 bp, 50-100 bp, 100-200 bp, 200-500 bp, 500-1 kbp, 1-2 kbp, 2-5 kbp, 5-10 kbp, 10-20 kbp, 20-50 kbp and 50-100 kbp base pairs. We then simulated 1 million random positions in the gibbon genome and filtered out those that did not land in introns. For each bin we calculated the percent of intronic LAVA elements observed in each bin, the percent that would be expected based on the simulated insertions and an enrichment score c with

$$c = \log_{10}\left(\frac{\%\ obs}{\%\ exp}\right).$$

Standard errors (before log transformation) were calculated similar to an odds ratio using the observed and expected counts

$$se = \sqrt{\frac{1}{obs_{bin}} + \frac{1}{obs_{tot}} + \frac{1}{exp_{bin}} + \frac{1}{exp_{tot}}}.$$

7.7 Network building and pathway functional enrichment of LAVA gene sets

We used the meta-pathway database tools Genemania[72] and Consensus Pathway Database (CPDB)[73-75] to observe the network structure of the LAVA gene set and analyze its enrichment over many publicly available pathway databases. Genemania is a meta-database system for protein function prediction, which integrates multiple genomics and proteomics data sources to infer the function of unknown proteins. In Genemania, reactions are collected from Reactome and BioCyc, via PathwayCommons. Network building allows extension of the initial gene set by the most likely associated genes and the enrichment over Gene Ontology (GO) labels is computed. CPDB includes interaction data for network building from 32 databases (see Kamburov et al, 2013 for a complete list)[73], and computes pathway

functional enrichment over practically all the pathways in publicly available databases. In particular, ConsensusPathDB-human integrates interaction networks including binary and complex protein-protein, genetic, metabolic, signaling, gene regulatory and drug-target interactions, and biochemical pathways. The interaction data are complementary, i.e. free of redundancies, thus the functional enrichment can be computed over interaction networks that represent pathways containing many different types of interactions. This analysis results in estimation of functional enrichment over complete pathways, rather than GO labels.

We have performed Genemania network construction, and GO enrichment as well as CPDB pathway enrichment. Using CPDB, we performed pathway enrichment analysis for the LAVA gene set over the 32 included interaction databases. Limiting the FDR to 0.2 resulted in 32 pathways. All enriched pathways are shown in Table ST7.4. We find significant enrichment for chromatin related pathways, including establishment of sister chromatid cohesion, cohesion loading into chromatin, and separation of sister chromatids.

p-value	q-value	pathway	source	members_input_overlap
8.70E-05	**0.067**	**Establishment of Sister Chromatid Cohesion**	**Reactome**	**SMC3; PDS5A; ESCO1; PDS5B**
0.001	0.12	Epithelial cell signaling in Helicobacter pylori infection - Homo sapiens (human)	KEGG	ADAM10; CASP3; PTPRZ1; IKBKB; ATP6V0A1; MAP2K4; ATP6V1C1
0.0012	**0.12**	**Mitotic Anaphase**	**Reactome**	**CLASP2; MAD1L1; PDS5A; SMC3; PPP2R2A; PDS5B; ANAPC1; ANAPC5; TAOK1; PPP1CC**
0.0013	**0.12**	**Mitotic Metaphase and Anaphase**	**Reactome**	**CLASP2; MAD1L1; PDS5A; SMC3; PPP2R2A; PDS5B; ANAPC1; ANAPC5; TAOK1; PPP1CC**
0.0014	**0.12**	**Cohesin Loading onto Chromatin**	**Reactome**	**SMC3; PDS5A; PDS5B**
0.0016	**0.12**	**M Phase**	**Reactome**	**CLASP2; NUP54; MAD1L1; PDS5A; SMC3; PPP2R2A; PDS5B; ANAPC1; NUP214; ANAPC5; TAOK1; PPP1CC**
0.0016	0.13	Binding of RNA by Insulin-like Growth Factor-2 mRNA Binding Proteins (IGF2BPs-IMPs-VICKZs)	Wikipathways Reactome PID	IGF2BP3; IGF2BP2
0.0016	0.13	Amyotrophic lateral sclerosis (ALS)	Wikipathways	BCL2L1; PPP3CC; RAB5A; CASP3; APAF1
0.0019	**0.13**	**Cell Cycle, Mitotic**	**Reactome**	**CEP164; DNA2; PPP1CB; CENPJ; NUP54; MAD1L1; PDS5A; ESCO1; PPP2R2A; TAOK1; ANAPC1; NUP214; PDS5B; CLASP2; SMC3; ANAPC5; NINL; PPP1CC**
0.0025	**0.14**	**Separation of Sister Chromatids**	**Reactome**	**CLASP2; MAD1L1; SMC3; PDS5A; PDS5B; ANAPC1; ANAPC5; TAOK1; PPP1CC**
0.0025	0.14	Alzheimers Disease	Wikipathways	ATF6; GNAQ; ADAM10; ITPR2; PPP3CC; APAF1; CASP3

p-value	q-value	pathway	source	members_input_overlap
0.0026	0.14	Pyruvate metabolism - Homo sapiens (human)	KEGG	ACAT1; ME1; ACACA; ACYP2; PC
0.0027	0.14	Metabolism of lipids and lipoproteins	Reactome	PI4KA; GBA; SCAP; CHKA; NCOR1;
0.0033	0.15	Apoptosis	Wikipathways	TAZ; ACAT1; SPTLC2; LIPC; ACACA; SLC44A5; IDH1; PDSS1; PPP1CB; PHYH; INPP5D; PPP1CC; ACSL3; PIP5K1A; OSBP; CDS2; COQ3 BCL2L1; TP63; CASP3; HELLS; IKBKB; MAP2K4; APAF1
0.0038	0.15	apoptotic signaling in response to dna damage	PID BioCarta	
0.0039	0.15	Mitotic Telophase/Cytokinesis	Reactome	SMC3; PDS5A; PDS5B
0.0039	0.15	TGF-Ncore	Signalink	NUP214; DAB2; SIRT1; PPP1CC; ZFYVE9
0.0044	0.15	TNF alpha Signaling Pathway	Wikipathways	BCL2L1; CASP3; TRAP1; IKBKB; MAP2K4; KSR2; APAF1
0.0048	0.15	TNF	INOH	CASP3; PSMD4; IKBKB; PSMA6; MAP2K4; APAF1
0.0053	0.15	Activation of caspases through apoptosome-mediated cleavage	Reactome PID	CASP3; APAF1
0.0053	0.15	Cytochrome c-mediated apoptotic response	Reactome	CASP3; APAF1
0.0053	0.15	Alanine Metabolism	SMPDB	PC; AARS
0.0057	0.16	Mitotic M-M/G1 phases	Reactome	CLASP2; NUP54; MAD1L1; PDS5A; SMC3; PPP2R2A; PDS5B; ANAPC1; NUP214; ANAPC5; TAOK1; PPP1CC
0.0078	0.19	N-linked glycosylation	PID	STT3A; RPN2
0.008	0.19	Amyotrophic lateral sclerosis (ALS) - Homo sapiens (human)	KEGG	BCL2L1; PPP3CC; RAB5A; CASP3; APAF1
0.008	0.19	Pyruvate metabolism	INOH	ACACA; ACAT1; ME1; ACYP2
0.0086	0.19	miR-targeted genes in lymphocytes - TarBase	Wikipathways	CHD1; NAA15; MATR3; DHX57; IDH1; P4HA2; RCOR1; NPR3; ATP6V0A1; WDR82; SFXN1; ANAPC1; ATP6V1C1; TRPV6; PLAG1; ARID4B; CDKAL1; GNL3L
0.0087	0.19	TGF_beta_Receptor	NetPath	SNX1; XPO4; SNX4; TRAP1; PPP2R2A; ZFYVE9; ANAPC1; NUP214; DAB2; ANAPC5
0.0089	0.19	Integrated Cancer pathway	Wikipathways	MSH2; MRE11A; CASP3; MSH6
0.0089	0.19	Signaling mediated by p38-alpha and p38-beta	PID	ATF1; MAPKAPK5; ATF6; RAB5A
0.0089	0.19	Cell Cycle	Reactome	CEP164; DNA2; PPP1CB; CENPJ; NUP54; MAD1L1; PDS5A; ESCO1; PPP2R2A; TAOK1; ANAPC1; NUP214; PDS5B; CLASP2; SMC3; ANAPC5; NINL; HIST1H2BJ; PPP1CC
0.01	0.2	Intrinsic Pathway for Apoptosis	Reactome	BCL2L1; PPP3CC; CASP3; APAF1

Table ST7.4 Full Pathway Enrichment of the LAVA geneset for FDR <0.2

7.8 Identification of LAVA-encoded major antisense polyadenylation sites (MAPS)

Luciferase reporter assay

To generate the firefly luciferase reporter construct pmiRGlo_ΔAATAAA (Fig. 3-B), the SV40 late polyadenylation signal terminating transcription of the luciferase gene in pmRIGlo (promega) was deleted by using the Stratagene QuikChange Site-Directed Mutagenesis Kit and the oligonucleotides pAmut_FW 5' TTGTAACCATTATAAGCTGCCAAGTTAACAACAACAATTG 3' and pAmut_REV 5' CAATTGTTGTTGTTAACTTGGCAGCTTATAATGGTTACAA 3'. The 3' ends (including polyA tail and 3' TSD) of the LAVA_E and LAVA_F1 elements (Ianc et al. submitted) were amplified using primer pairs LA_E_Sal (5' ATGTCGACCTACCACCGAGGCCAGAAGCAATG 3')/ LA_E_Sac (5' ACGAGCTCGGTCTTCACAATTACAGGCTAAGCAC 3' – 537 bp fragment) and LA_F_Sal (5' ATGTCGACCTACCATGGAGGCCAGAAGCAATG 3')/ LA_F_Sac (5' ACGAGCTCTTCTGAAAGTCAAAACGTTACGTCGG 3' - 559bp fragment). PCR products were subcloned, sequenced and subsequently inserted into pmiRGlo_ΔAATAAA via SacI/SalI in opposite orientation relative to the firefly luciferase reporter gene (luc2), resulting in luciferase reporter plasmids pmiRGlo_LA-E and pmiRGlo_LA-F (Fig. 3-B).

Luciferase assays were performed in the human teratocarcinoma cell line GH[76] using the Dual Luciferase Assay System (Promega) according to the manufacturer's instructions. 2.5 x 10^5 GH cells were seeded per well in 12-well plates and grown for 24h at 37ºC. Cells were co-transfected with 0.9µg of the plasmid pmiRGloLA_E, pmiRGloLA_F or pmiRGlo_ΔAATAAA, and 0.1µg of the renilla luciferase expression plasmid phRL-TK (Addgene) using TransIT®-LT1 Transfection Reagent (Mirus) according to the mansufacturer's instructions. Light units were quantified in a microplate luminometer (Tecan, Infinite F200). For normalization, the number of firefly luciferase light units was divided by the number of renilla luciferase light units. P-values for all pairwise comparisons LA_F vs. LA_E, delta_PA vs. LA_F, and delta_PA vs. LA_E respectively (with 95% CI) were adjusted for multiple comparisons according to Bonferroni. The statistical analysis was performed with SAS®/STAT software (PROC GLM), version 9.3, SAS System for Windows.

To identify LAVA-encoded transcription termination site(s) and 3' ends of luc2-LAVA_F fusion transcripts, we applied the 3'RACE system for rapid amplification of cDNA ends. For that purpose, poly(A) selected RNA was isolated from pmiRGlo.LA_F transfected GH cells using the Dynabeads mRNA purification kit (Ambion). We generated cDNAs applying the GeneRacerTM Kit (Invitrogen, Cat-No: L1502-01) and the GeneRacer™ Oligo dT Primer according to the manufacturer's instructions. Subsequently, PCR with the GeneRacer™ 3′Primer and the luc_FW_3 primer (5'-CTCCTTCTCGGTCATGGTTTTACC-3') was performed using the generated cDNAs as templates. For a nested reaction, GeneRacer™ 3′Primer and the primer luc_FW_1 (5'-GTCCTTTAGGCACCTCGTCC-3') were used. PCR was performed with Taq polymerase (Applied Biosystems) using the following cycle conditions: one cycle 94°C for 5min; 36 cycles 95°C for 30s, 68°C for 30s, 72°C for 60s; 1 cycle 72°C for 10min. Reaction products were analyzed by agarose gel electrophoresis, purified with the MinElute gel extraction kit (Qiagen), cloned into the pGEM-T Easy vector (Promega) and sequenced. Multi-alignment of the sequences is represented in Extended Data Fig. 4.

7.9 Analysis of RNA-seq data to identify premature transcription termination

Paired end 76 bp Illumina reads (Table ST2.4) were trimmed by 20 bp at the 5' end of each read and then aligned as fasta sequences to nomLeu1 using Bowtie 2 verson 2.1.0, using default parameters. Alignments to the plus strand of genes, in which the second (3') mate fell into an intronic antisense LAVA element, were kept for further processing. For these alignments, the original full length sequences were retrieved and realigned to a database of the genes recovered in the first step. Any pair that no longer aligned was then a candidate for a prematurely polyadenylated transcript, possibly containing an untemplated poly(A) sequence. For these pairs, the read aligning to the antisense LAVA transcript was examined further, using bl2seq (v2.2.26) to create local alignments, using a decreased mismatch penalty (-1) and filtering at hash (-F F). For any read that still did not align full length, if the 5' end of the read contained more than 8 T residues in the last 10 residues, it was reported as a polyadenylated read. Fig. SF7.1 shows one of the results from this analysis and genes identified

through this approach are listed in Table ST7.5. While the human L1 element has a discrete premature polyadenylation signal, the transcripts we identified were truncated and polyadenylated at several different sites in the 3' end of the LAVA element, suggesting that a more diffuse, regional signal may play a role in transcript termination in this retrotransposon (data not shown).

Ensembl ID	Gene name	Intron	LAVA Position nomLeu3	Subfamily	Spanning reads
ENSNLET00000000032	MAPKAPK5	7	chr10:85,664,861-85,666,365	B1F2	6168
ENSNLET00000001094	TRPM3	25 (last)	chr1a:64,679,978-64,682,392	E	1
ENSNLET00000001284	CHIC1	3	chrX:66,332,533-66,333,941	B1D	177
ENSNLET00000001422	COG6	8	chr5:67,309,429-67,312,234	C2	9677
ENSNLET00000001463	WDR25	3(4)	chr22a:5,496,920-5,498,302	E	126
ENSNLET00000001972	ZCCHC17	7 (last)	chr12:674,836-679,776	F2	124
ENSNLET00000002251	CDKAL1	4	chr8:71,456,753-71,458,184	B2R2	78
ENSNLET00000002650	ADCK2	7 (last)	chr13:92,113,638-92,114,784	B1D	1606
ENSNLET00000003966	GPR114	2	chr2:65,892,631-65,894,343	B2R2	1742
ENSNLET00000006409	SLC22A15	2	chr12:61,680,456-61,682,257	F2	1865
ENSNLET00000007727	APAF1	12	chr10:6,816,512-6,819,443	F1	2156
ENSNLET00000008982	NBEAL1	32	chr22a:93,479,290-93,480,906	B2C	891
ENSNLET00000012580	PRPF3	8	chr12:51,828,997-51,830,437	C4B	5476
ENSNLET00000014274	ZSCAN25	4 (last)	chr17:34,699,184-34,700,709	B2R2	15050
ENSNLET00000014734	ATP6V0A1	2	chrUn_GL397460_1:1,370,203-1,371,757	C4B	4052
ENSNLET00000016108	ZNF346	4	chr2:3,603,291-3,604,481	D1	645
ENSNLET00000016411	C9orf171	5(6) last	chr8:111,835,916-111,839,320	C4B	977
ENSNLET00000016572	HELLS	14	chr3:39,689,696-39,691,784	C2	12792
ENSNLET00000017191	CD97	1	chr10:60,085,209-60,086,757	B2C	2890
ENSNLET00000017984	TMEM53	2	chr12:14,007,502-14,009,434	F1	50
ENSNLET00000017987	TTC26	16	chr13:90,548,838-90,550,315	B1R2	6288
ENSNLET00000018217	C11orf49	1	chr15:109,075,715-109,077,284	B1F2	726
ENSNLET00000018883	NCOR1	2	chr19:65,551,527-65,552,888	B1F2	500
ENSNLET00000019227	TEX2	1	chr19:63,023,120-63,025,045	C4B	299
ENSNLET00000019260	PGM1	1	chr5:44,108,713-44,109,812	B1R2	3455
ENSNLET00000019553	PPP1CB	5	chr19:61,544,023-61,544,986	A1	2275
ENSNLET00000019605	SORCS3	11	chr3:29,096,318-29,097,416	E or B1F2*	22184
ENSNLET00000019886	ABL2	1	chr12:23,140,831-23,143,438	F2	67
ENSNLET00000020116	TDRD5	5	chr12:22,687,467-22,689,246	B1G	1947

Table ST7.5 List of genes for which evidence of premature polyadenylation was found. All LAVA elements are inserted in antisense direction with respect to the gene. Subfamilies are indicated in column 5. The

spanning read column indicates the number of sequences for which both pairs aligned on either side of the LAVA-containing intron and the distance between the mapping sites indicates splicing out of the intron, therefore suggesting presence of the full transcript.

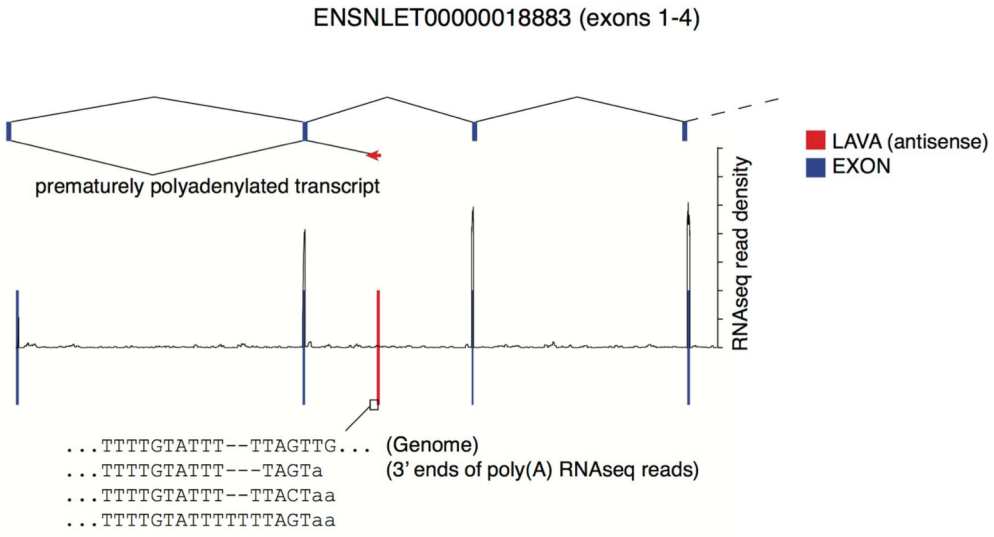


Figure SF7.1 The fragment read shown does not fully align to the genomic sequence due to untemplated A residues at its 3' end. After trimming of its 3' end it aligns to the antisense intronic LAVA element in the gene and terminates near an antisense polyadenylation site. While this read can be aligned by itself, paired end reads were used throughout to unambiguously place the polyadenylated sequences in intronic LAVA elements shown in table ST7.5.

Possible effects of LAVA on transcriptional output

There are different ways in which LAVA could affect transcriptional output, besides causing lower levels of full-length transcripts. First, looking at genes orthologous to the gibbon genes listed in the EDT1 we find human isoforms containing alternatively spliced exons upstream of the site where LAVA integrated in gibbons for 7 of the 15 genes (Supplementary File 6). As alternative splicing and RNA polymerase II transcript termination/ polyadenylation are tightly coupled processes, LAVA-mediated premature termination could (i) affect different isoforms to a different extent and (ii) influence the ratio between isoforms. Second, to investigate possible effects of LAVA insertions at the protein level we performed *in silico* translation of the exons using the human reference to avoid problems resulting from gaps etc. in the gibbon assembly. This analysis revealed that in 6 of the 15 cases complete functional domains

should be retained even after LAVA-mediated premature termination. However, the complete structure of the transcripts is unknown. For instance, depending on which part of the LAVA containing intron is included, these domains might be combined with sites of post-translational modification that are different in the LAVA terminated isoform compared to the wild-type. Alternatively, the truncated proteins could act in a dominant negative manner.

Finally, we analyzed all the alternatively spliced forms of genes orthologous to the gibbon genes listed in the EDT1 (i.e. microtubule cytoskeleton category with LAVA insertions) in human and compared them with the transcripts that would result from early transcription termination by LAVA (Supplementary File 6). We found isoforms resulting from premature termination for 9 out of 15 of the human orthologs. However, except for one case (a *MAP4* isoform terminating in intron 3), all of these prematurely terminated isoforms extend more 3' than the ones predicted to result from antisense LAVA-mediated premature termination indicating that LAVA-mediated premature termination has the potential to generate isoforms that do not have any counterparts in humans. Moreover, the potentially LAVA-truncated genes all contain fewer exons than the shortest human isoform starting with the same exon.

7.10 LAVA elements can function as exon traps

VNTR composite retrotransposons (such as LAVA and SVA) that are localized in introns can influence gene expression by functioning as exon traps in sense. In the case of SVA, it has been demonstrated that splicing of upstream exons of cellular genes to the retroelement RNA can occur in two different ways: i) exons are spliced directly to the element using the splice acceptors present in the SVA *Alu*-like region, or ii) SVA elements can activate cryptic splice acceptors leading to exonization of the part of the intron directly adjacent to the element[77]. Analysis of LAVA 5' transductions (Fig. SF7.2) provides evidence that both of these mechanisms are operative for LAVA as well. A LAVA element in Nleu 1.0 was considered to contain a 5' transduction if it displayed unambiguous TSDs (>6bp) and a >20-bp sequence between the TSD and the 5'-end of the LAVA sequence. This way, 12 LAVA elements containing 5' transductions could be identified. The source loci of the transductions were identified

using the transduced sequences as queries in BLAT searches at http://genome.ucsc.edu. In 9 out of the 12 cases (75%) the transduced sequence was found to map to loci annotated in the "Non-gibbon RefSeq Genes" track of the UCSC genome browser – indicating that the transduced sequence originated from a spliced cellular RNA. In all cases this finding could be corroborated by aligning the transduced sequence to the respective human reference RNA sequence. The fraction of LAVA-associated 5' transductions consisting of spliced RNAs is considerably higher than that found among SVA 5' transductions (excluding the independently amplifying SVA_F1 subfamily[77]). There are up to six exons of cellular RNAs included in 5' transduced sequences of LAVA elements (Table ST 7.6). Source elements can still be found in Nleu1.0 for only two of the spliced transductions, suggesting selective pressure against LAVA elements capable of exon trapping. Interestingly, one of the LAVA elements was found to have used the bona fide splice acceptor of an upstream exon (GTPBP8 exon 6). Differences to SVAs are observed in cases of splicing to the *Alu*-like region: whereas in SVAs the four splice acceptors used in transduction are distributed over the entire *Alu*-like region, fusion to LAVA occurs only to an acceptor immediately downstream of the CT hexameric repeats.

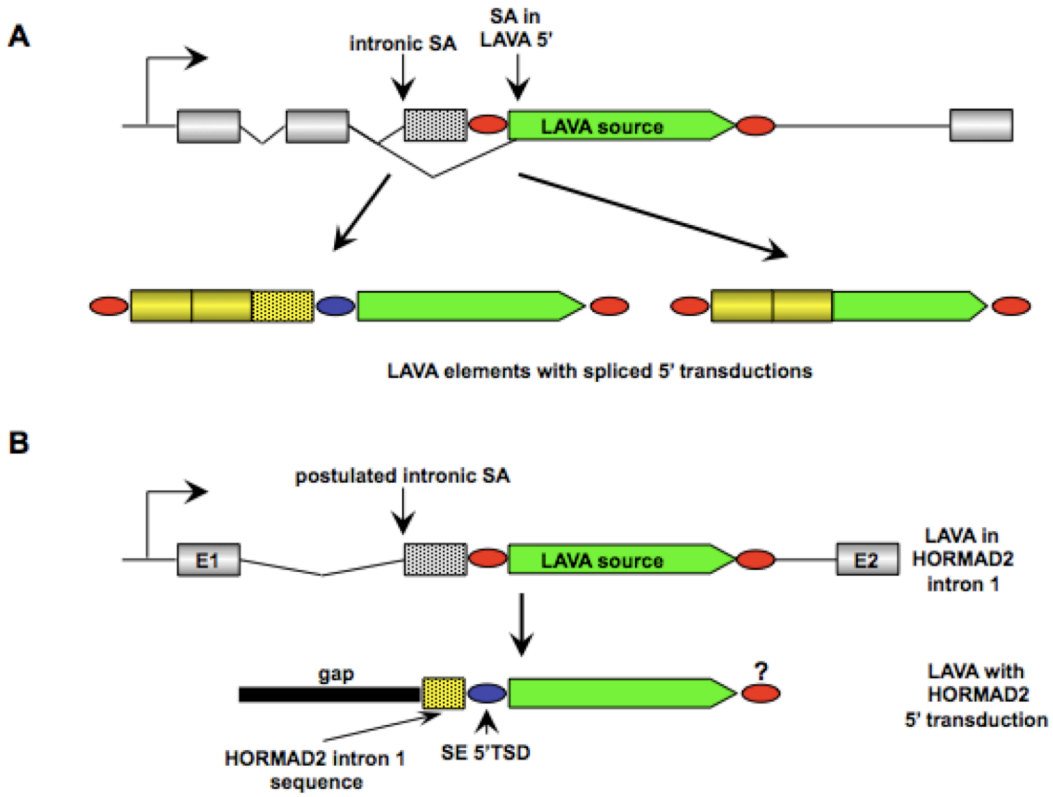


Figure SF7.2 LAVA elements can function as exon traps. (A) Intronic LAVA elements in sense orientation can transduce upstream exons at their 5' ends through either activation of intronic (cryptic) splice acceptors (SA) or splicing to a splice acceptor at the 5' end of the element. Exons at the source locus are in grey. Transduced intron sequence is hatched. Transduced exons are in yellow. Target site duplications (TSDs) are shown as red ovals. The 5' TSD of the source element found in the offspring carrying a spliced 5' transduction is shown in blue. (B) In case of the 5' transduction carrying LAVA that originates from a source element in HORMAD 2 intron 1 the source element 5'TSD and upstream intron sequence are present in the current genome build. Exon 1 (E1) is likely to be found in the sequence covered by a gap in the current assembly.

A. Activation of intronic cryptic splice acceptors					
position (subfamily)	**Splice Acceptor position**	**Splice acceptor**	**inferred position of SE or SE**	**Exons included**	**observations**
GL397360:5,231,615-5,233,405 (D1)	CLIP1 intron 1 (GL397506: 659259)	TTTTTTTTTTTGCAG	intron 1 of CLIP1 GL397506: 659122	CLIP1 exon 1	
GL397303:18,213,910-18,218,331 (B1R2)	TESK2 intron 3 (GL397442 2243550)	AGCTTTCTTTTATAG	intron 3 of TESK2	TESK2 exon 1-3 (alt. exon from intron 1)	
GL397583:92,542-94,118 (B1R2)	ARMC6 intron 3	TTTTTTGTTTTTCAG (human seq)	intron 3 of ARMC6 NM_033415 in HSA	ARMC6 exon 1-3	SA could not be determined in NLE
GL397356:5,833,295-5,836,765 (F2)	MTMR12 intron 8 (GL397369 3304527)	TTTTATATTCCTAAG	intron 8 of MTMR12 (3' part)	MTMR12 exon 1-6; 8	SE must have been 5' truncated VNTR
GL397266:14,505,919-14,508,676 (A2)	GTPBP8 intron 5/ exon 6	CTTTTCTTTTCACAG	downstream of GTPBP8	GTPBP8 exon 1-6 (exon 6 longer than in HSA)	SE must have been 5' truncated in Alu-like
GL397386:gap-4,975,086 (F2)	HORMAD2 intron 1	nd – assembly gap	intron 1 of HORMAD2 (GL397272:2,459,766-2,462,107)	HORMAD2 exon 1	3' truncated element
B. Splicing to a splice acceptor in the LAVA 5' part					
position (subfamily)	**SA position (in SF cons)**	**Splice acceptor (in SF consensus)**	**inferred position of SE or SE**	**Exons included**	**observations**
GL397301:6,288,467-6,290,304 (B1R2)	11	CCCTCTGTGGCCCAG (SE sequence)	intron 6 BRWD3 (GL397321:5,544,606-5,546,072)	BRWD3 exon 1-6	acceptor prediction on SE
GL397327:6,787,931-6,789,367 (B1R1)	11	CCCTCTGTGGCCCAG (SF consensus)	intron 2 of MND1	MND1 exon 1-2	
GL397373:1,469,169-1,471,541 (E)	11	CCCTCTGTGGCCCAG (SF consensus)	intron 2 of ZNF714	ZNF714 exon 1-2	

Table ST7.6 LAVA elements containing exons of cellular genes as 5' transduced sequences. The 100% conserved AG dinucleotide at the intron 3' end is highlighted in green. (SE = source element; SA = splice acceptor; SF = subfamily; HSA = Homo sapiens; NLE = *Nomascus leucogenys*

Supplemental Section S8 – Phylogenetic analysis using autosomal DNA

8.1 Next-generation sequencing of the four gibbon genera

Samples

Previous work examining genetic diversity amongst the four known gibbon genera have been limited to studies examining uniparentally inherited markers[34,70,78-81], very short stretches of autosomal sequence[82], or *Alu* repeats[83]. In order to examine diversity at the whole genome level we performed next-generation sequencing on two individuals from each of the four genera (Section S2; Table ST2.1). For the genus *Nomascus* we examined two individuals different from the reference Asia; for the genus *Hylobates* (the most diverse genus with ~9 species) we examined one individual each from *H. moloch* and *H. pileatus*. It is important to bear in mind in all subsequent discussion of our results based on these samples that only the two *Hoolock* samples represent wild born individuals. Hybridization resulting in viable (but not nessearily fertile) offspring is known to occur between gibbon genera and species in captivity[84] and may affect our analysis in unexpected ways. In particular we would have little power to identify hybridization amongst species within the same genus.

Read Mapping and Variant Calling

Sequences in FASTQ format were trimmed with cutadapt[85] to remove Illumina TruSeq adapter sequences. Reads with less than 25 nucleotides left after trimming were dropped, along with their mates. The remaining reads were aligned to Nelu1.0 with stampy (v. 1.0.17) [86]. For the two *N. leucogenys* (NLE) samples, stampy was used in its "hybrid mode" where alignment with BWA (v. 0.5.9)[16] is attempted first. A substitution rate of 0.001 was specified, along with BWA minimum seed length of 2, fraction of missing alignments 0.0001, and quality threshold 10. For the non-NLE samples, stampy was used with a substitution rate of 0.015[82]. Local realignment at indel sites was performed with the Genome Analysis Toolkit (GATK, v. 1.4-37)[87,88]. PCR duplicates were then removed with samtools. Picard (v. 1.70) (http://sourceforge.net/projects/picard/) CleanSam was run on the output. The two samples from each genus were then merged using Picard MergeSamFiles, and the resulting files were

split using samtools[69] into 100 files containing ~180 contigs each to facilitate further parallel processing. The GATK UnifiedGenotyper was run and Single Nucleotide Variants (SNVs) and indels with a quality score of at least 50 were retained to create a mask of variant sites to be excluded from base quality score recalibration (BQSR). The BQSR steps were run with the standard set of covariates, and the resulting files were merged across all samples. The GATK indel realignment tools were then run again to standardize alignment of indels across the samples. Default settings were used except that "BadCigar" reads were excluded and BAQ calculation was added. The UnifiedGenotyper from GATK version 2.1-11 was then used to call SNVs and indels in each genomic part using the "EMIT_ALL_SITES" mode (with the BAQ calculation included) to produce VCF files with data for all genomic positions. (Version 2.1 was used for this step to allow multiallelic calling). VCFs for all genomic parts were then merged using a custom perl script. Annotations were added to specify the consensus quality score of the Nleu1.0 reference sequence at each position (see Fig. SF8.1 for an overview).

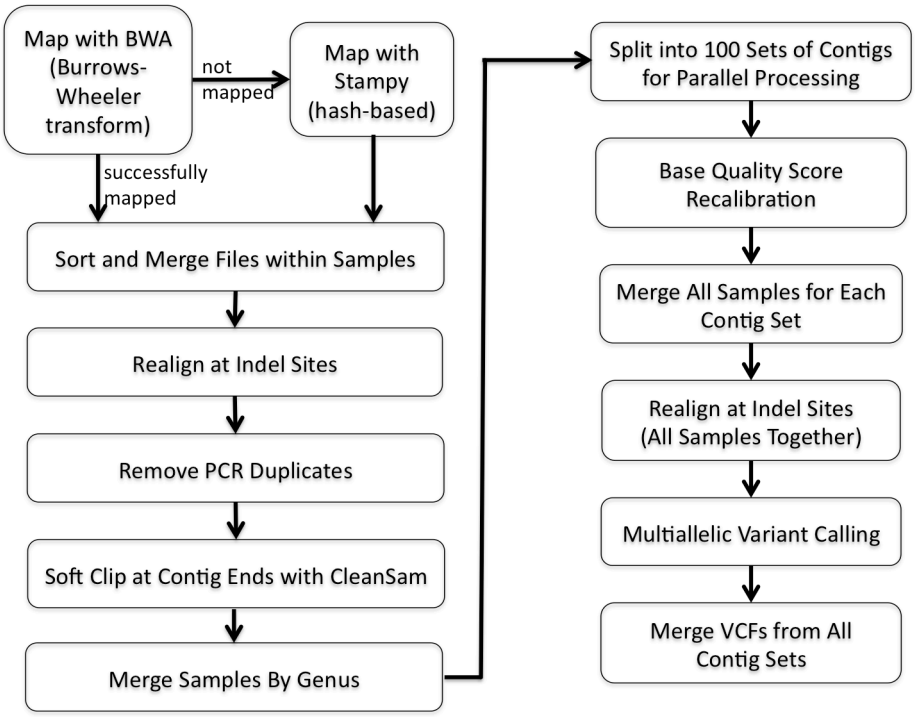


Figure SF8.1 Overview of our the pipeline for the analysis of next-generation sequencing data

Masks

The Nleu1.0 genome is composed of 17,968 contigs, ranging in size from 2,496 bases to ~74 Mbp. As small loci may be compressed, and represent duplications in the gibbon genome that have not been properly separated during the assembly process, we masked out all scaffolds less than 1 Mbp in length, yielding 273 scaffolds that span ~2.73 Gbp. UCSC's gibbon-human pairwise alignments where used to identify non-autosomal sequence. Specifically, gibbon loci that aligned to human X, Y or M in UCSC's "net" alignments[3] were masked, along with locations in the gibbon genome that were not primary alignments to locations in the human genome. Further, locations where the gibbon reference quality was below a phred-quality of 50, repeats (identified by Tandem Repeat Finder[25] by RepeatMasker[46], and LAVA elements identified in this project), CNVs with an estimated ploidy >2.5 in any sample, infinite sites violations, positions where any sample has less than 7x coverage, or more than their 95th percentile read depth, and bases within 3bp of any called indel called were ignored, unless otherwise specified, from downstream analysis.

High Coverage Exome Validation

Exome capture using the TruSeq Exome Enrichment Kit (Illumina) was performed on one NLE sample (Vok, 116x coverage) and one SSY sample (Monty, 64x coverage), and the resulting data were run through the pipeline described in Fig.SF8.1. The exome targeted regions were lifted over to the Nleu1.0 genome using the UCSC liftOver utility using the default parameters, and the emit-all VCF of the exome capture data were restricted to these loci.

Based on exome calls with 30x ≤ coverage ≤ 200x and controlling for coverage in the whole genome data, homozygous reference, homozygous alternate and heterozygous sites in the exome data were concordant with the whole genome data 99.8%, 99.0% and 99.5% of the time for the NLE sample and 97.1%, 98.5% and 99.9% of the time for the SSY sample (Table ST8.1). The generally marginally greater concordance for the NLE sample suggested reference biases were still present in our data despite the use of the aligner Stampy with a greater allowance for substitutions for non-NLE samples

(see *Veeramah et al. submitted* for an in depth description of the exome validation results as well as a method to use this information to correct whole genome data in coalescent-based analysis).

NLE-VOK	HOMO REF	HET	HOMO ALT
HOMO REF	74431	95	0
HET	184	11788	9
HOMO ALT	1	30	1823

SSY-MONTY	HOMO REF	HET	HOMO ALT
HOMO REF	10088	44	0
HET	286	9011	53
HOMO ALT	9	96	68774

NLE-VOK	HOMO REF	HET	HOMO ALT
HOMO REF	99.75%	0.80%	0.00%
HET	0.25%	98.95%	0.49%
HOMO ALT	0.00%	0.25%	99.51%

SSY-MONTY	HOMO REF	HET	HOMO ALT
HOMO REF	97.16%	0.48%	0.00%
HET	2.75%	98.47%	0.08%
HOMO ALT	0.09%	1.05%	99.92%

Table ST8.1 Concordance between genotype calls from whole genome (rows) and whole exome sequencing (columns) in the NLE and non-NLE individual. First two tables represent absolute number and second two tables represents cells as percentage of exome genotypes

8.2 Genetic diversity within and among gibbon genera

To assess the overall level of nucleotide diversity in gibbons we calculated the average number of pairwise differences for each sample, π. Only unmasked sites called in all eight individuals that were mono- or di-allelic were considered, resulting in ~461 million callable sites (**Table ST8.2a**). The two NLE samples demonstrated the highest level of nucleotide diversity (π ~$2.2x10^{-3}$), while values as low as ~$7.3x10^{-4}$ were observed in the HPI sample (**Fig ST8.2**). The HMO sample was also relatively high at ~$1.7x10^{-3}$, followed by SSY (~$1.4x10^{-3}$) and then the two wild born HLE (~$8x10^{-3}$). Our results are largely concordant with the previous autosomal analysis of Wall et al.2013[91]. Within both SSY and HLE, the sample with the higher coverage had lower levels of nucleotide diversity. Our high coverage WES validation data suggest that this pattern may arise due to a bias resulting from a higher probability of calling true homozygous reference genotypes as heterozygous at lower coverage sites in non-reference individuals (this cell of **Table ST8.1** showed the greatest discordance rate at ~3%). Relaxing the requirement that all individuals be called at a site and allowing tri-allelic (which occurred at 386,766 callable sites) and tetra-allelic (which occurred at 2,156 sites) sites resulted in a slight (mean 5%) increase in nucleotide diversity in all cases; however, the relative values remained similar (**Table ST8.2b**).

Sample	Nb Hets	Nb Homo Ref	Nb Homo Alt
NLE_Vok	1,022,918	459,486,365	510,354
NLE_Asteriks	1,007,445	459,517,937	494,255
SSY_Monty	664,404	454,548,156	5,807,077
SSY_Karenina	576,059	454,581,738	5,861,840
HLE_Drew	471,078	454,726,713	5,821,846
HLE_Maung	376,071	454,766,836	5,876,730
HPI_Domino	340,249	454,567,388	6,112,000
HMO_Madena	764,013	454,418,909	5,836,715

Table ST8.2a: Breakdown of callable sites across all 8 gibbon samples using a strict filtering criteria

Sample	Nb Hets	Nb Homo Ref	Nb Homo Alt
NLE_Vok	1,973,254	860,876,984	981,445
NLE_Asteriks	1,608,099	707,449,852	784,791
SSY_Monty	1,249,515	824,063,223	10,658,674
SSY_Karenina	1,112,561	841,905,625	11,011,561
HLE_Drew	935,954	851,013,901	11,038,489
HLE_Maung	755,859	859,517,573	11,257,623
HPI_Domino	626,804	784,432,197	10,737,761
HMO_Madena	1,433,090	825,615,897	10,737,751

Table ST8.2b: Breakdown of callable sites across all 8 gibbon samples using relaxed filtering criteria

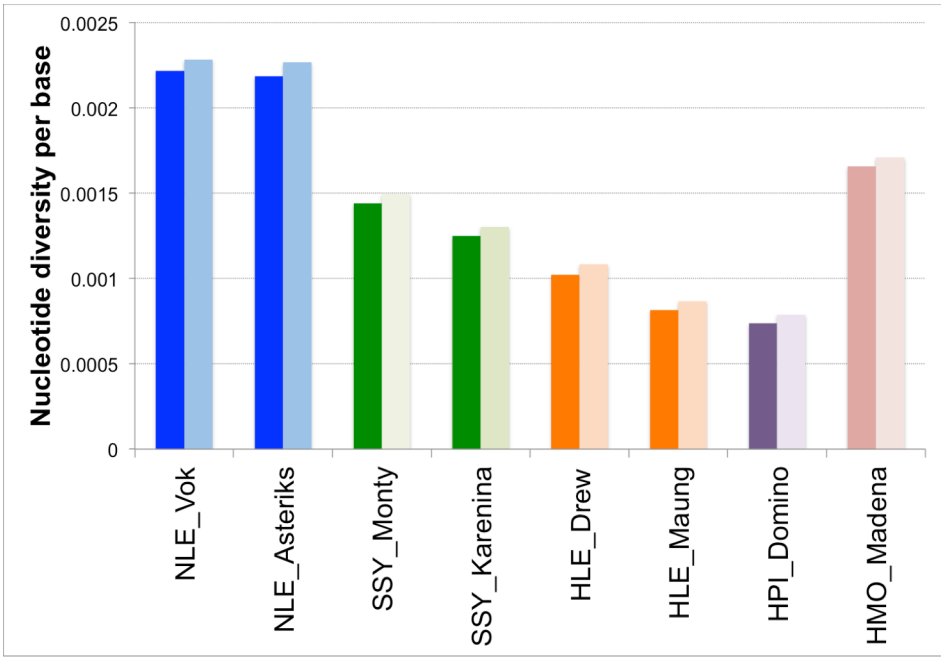


Fig SF8.2. Bar chart showing nucleotide diversity across gibbon genera. Lighter shading based on more relaxed filtering criteria

8.3 Establishing the phylogenetic relationships between gibbon genera

Whole Genome Sequence Divergence

A set of 12,413 loci at least 50 kbp away from the nearest exon with 1 kbp of sequence fully called in all 8 individuals across a contiguous stretch of no more than 3kb (non-genic loci) and a set of 11,323 loci overlapping with exons with 200 bp of callable sequence cross a contiguous stretch of no more than 4kb (genic loci), filtered for CpG sites and conserved phastCons elements, were identified in order to perform a phylogenetic analysis of the four gibbon genera (more details on these loci can be found in *Veeramah et al. submitted*).

Pairwise Sequence Divergence and NJ analysis

First we performed a simple pairwise sequence-divergence based phylogenetic analysis by concatenating all alignments (though non-genic and genic loci were examined separately). We calculated the sequence divergence, *k*, amongst all eight gibbons across the whole set of loci as well as from the aligned human reference genome (GRCh37) using the multiz 11-way alignments from UCSC. As all gibbon data were unphased and thus diploid, we counted the number of different alleles between individuals. Sites where two individuals had the same genotype had 0 different alleles, sites where one individual was homozygote and one individual was heterozygote had 1 different allele and sites where each individual was homozygote for a different allele had 2 different alleles. *k* was then estimated by the total number of different alleles/(2 X the total number of sites considered). When calculating *k* from the haploid GRCh37 sequence we assumed that it was homozygous at all sites. This will result in a slightly underestimate of the true *k* due to missing heterozygosity within humans on the order of ~1 difference per 1000bp, but as the purpose is to act as an outgroup for the within gibbon relationships this does not impact our results greatly. We also performed an analysis where a site that was heterozygous in both individuals was set as having 0.5 different alleles (rather than 0 because of an identical genotype), but such instances were so rare that it had a negligible effect on our inference of *k* and any downstream analysis, essentially resulting only in a very slight increase in *k* for pairs of

samples from the same species (where double heterozygotes are most likely to occur). A NJ tree was constructed and visualized from the pairwise *k* matrices using MEGA 5[89]. Bootstrapping was performed by resampling loci with replacement for 10,000 iterations using PHYLIP[90].

For the non-genic regions pairs of individuals from different genera all had similar *k* values, with a mean of 0.011 (range = 1.08-1.12%) (Table ST8.3), similar to previous analysis of small sections of autosomal loci[82,91]. Within species the sequence divergence was approximately a tenth of that between genera, while the two *Hylobates* species had an intermediate *k* of 0.0056.

	GRCh37	NLE Vok	NLE Asterik	SSY Monty	SSY Karenina	HLE Drew	HLE Maung	HPL Domino	HMO Madena
GRCh37		3.01E-02	3.00E-02	3.03E-02	3.03E-02	3.00E-02	3.01E-02	3.03E-02	3.02E-02
NLE Vok	3.50E-02		**1.01E-03**	7.81E-03	7.83E-03	7.55E-03	7.59E-03	8.06E-03	7.91E-03
NLE Asterik	3.50E-02	**1.53E-03**		7.81E-03	7.83E-03	7.54E-03	7.58E-03	8.05E-03	7.90E-03
SSY Monty	3.53E-02	1.09E-02	1.09E-02		**7.07E-04**	7.54E-03	7.57E-03	8.06E-03	7.90E-03
SSY Karenina	3.53E-02	1.09E-02	1.09E-02	**1.09E-03**		7.55E-03	7.59E-03	8.08E-03	7.92E-03
HLE Drew	3.51E-02	1.08E-02	1.07E-02	1.07E-02	1.07E-02		**5.37E-04**	7.87E-03	7.73E-03
HLE Maung	3.51E-02	1.08E-02	1.08E-02	1.07E-02	1.07E-02	**8.40E-04**		7.90E-03	7.77E-03
HPL Domino	3.52E-02	1.11E-02	1.10E-02	1.11E-02	1.11E-02	1.10E-02	1.10E-02		**3.85E-03**
HMO Madena	3.52E-02	1.10E-02	1.09E-02	1.10E-02	1.10E-02	1.09E-02	1.09E-02	**5.59E-03**	

Table ST8.3 Sequence divergence among gibbon samples for non-genic (lower diagonal) and genic (upper diagonal) loci. Bold type indicates sequence divergence with genera

Unsurprisingly a neighbor-joining (NJ) tree based on these pairwise *k* values and rooted by a human outgroup showed the same pattern as previous analyses, with very short internal branches and long external branches separating different genera. Consistent with Wall et al.[91] SSY and HLE appeared as sister taxa, while bootstrapping of loci with replacement for 10,000 iterations demonstrated an unstable phylogeny with regard to the relative placement of the more external NLE and HMO/HPI branches.

Values of *k* at the analogous set of ~11,000 independent 'genic' loci that span exons showed a similar pattern with an r^2 value of 1.0 (non-genic versus genic pairwise *k* values). However, *k* was ~29% lower than at non-genic loci (mean 0.78%), suggesting stronger sequence conservation at these regions, consistent with the effect of long term purifying or positive selection.

Sliding Window Phylogenetic Analysis

While popular, approaches that use consensus sequence trees to infer the underlying species tree can fail, particularly when the speciation time is relatively recent and internal branches are short, something that may apply to gibbons given the similarity in divergence values amongst the four genera. In order to further investigate the phylogenetic relationships among genera based on local gene trees across the genome (i.e. to examine the extent of ILS), we calculated *k* for each species as above within 100 kbp non-overlapping windows along the whole genome. From a total of 25,779 possible windows, 23,364 had at least 10,000 variable sites for which (Unweighted Pair Group Method with Arithmetic Mean) UPGMA trees were then calculated (UPGMA was chosen because of its significant speed over alternative methods for such a large dataset and the purpose here is simply to examine the extent of variation in tree topologies rather than to accurately infer any particular topology). Only eight windows did not include the two *Hylobates* species as a monophyletic group. Of the remaining windows, all 15 possible rooted topologies for the four genera were observed at considerable frequencies (2-15%, mean 6.7%) (Table ST8.4) (Extended Data Fig. 6) consistent with high levels of ILS

The most common tree was the same as that observed for our genic and non-genic loci (((SSY,HLE),NLE),(HPI,HMO)) but was only observed 15% of the time. Highly similar results were observed using 10kb non-overlapping windows with at least 1,000 variable sites, with the two *Hylobates* species being monophyletic in 92% of the 179,753 windows and an r^2 of 0.82 when comparing the relative counts of these trees in the 100 kbp and 10 kbp window datasets, though on this occasion (((SSY,HLE),NLE),(HPI,HMO)) was the second most common tree (n=16,363) after (((SSY,HLE)(HPI,HMO)),NLE) (n=16,367).

Topology	Count
(((SSY,HLE),NLE),(HPI,HMO))	3599
(((SSY,HLE),(HPI,HMO)),NLE)	3086
(((NLE,HLE),SSY),(HPI,HMO))	2537
(((NLE,SSY),HLE),(HPI,HMO))	1854
(((NLE,HLE),(HPI,HMO)),SSY)	1678
((((HPI,HMO),HLE),SSY),NLE)	1563
(((HPI,HMO),NLE),(SSY,HLE))	1320
((((HPI,HMO),HLE),NLE),SSY)	1219
((((HPI,HMO),SSY),HLE),NLE)	1201
(((NLE,SSY),(HPI,HMO)),HLE)	1109
((((HPI,HMO),NLE),HLE),SSY)	957
((((HPI,HMO),SSY),NLE),HLE)	899
((((HPI,HMO),NLE),SSY),HLE)	860
(((HPI,HMO),SSY),(NLE,HLE))	814
(((HPI,HMO),HLE),(NLE,SSY))	659
((((NLE,HMO),HPI),SSY),HLE)	1
((((SSY,HLE),NLE),HMO),HPI)	1
((((SSY,HLE),HPI),HMO),NLE)	1
((((SSY,HPI),HLE),NLE),HMO)	1
((((SSY,HMO),HLE),HPI),NLE)	1
((((SSY,HMO),HPI),NLE),HLE)	1
((((HLE,HMO),NLE),SSY),HPI)	1
((((HLE,HMO),HPI),NLE),SSY)	1

Table ST 8.4 UPGMA trees for 100kb non-overlapping sliding windows moving along the Gibbon genome

Coalescent-based analysis of population divergence

Sequence divergence essentially reflects an upper bound for when populations split and can give a false signal of the phylogeny if the time of coalescence for sequences can fall within the ancestral population of the extant populations of interest[98]. Therefore in order to investigate the gibbon phylogeny at the population divergence level we developed a coalescent-based Approximate Bayesian Computation (ABC) method that explicitly takes into account sequence and population divergence simultaneously, can cope with large amounts of sequence data, is not dependent on haplotype phase and can incorporate information derived from our exome capture validation. This method is also able to estimate divergence times and effective population sizes of gibbon taxa. This methodology was then applied to the gibbon shotgun sequencing data using the ~12,000 non-genic loci and ~11,000 genic loci described above. The methodology and detailed description of the result can be found in *Veeramah et al. submitted*. Briefly our results were that no particular topology (out of 12 asymmetric and 3 symmetric topologies for 4 genera) was particular well supported, though the most frequently observed topology in the sequence divergence-based analysis, (((SSY,HLE),NLE),(HPI,HMO)), did have consistently higher posterior probabilities than most other topologies. Parameter estimates for both an instantaneous speciation model and a bifurcating speciation model for gibbon divergence support the hypothesis that all four gibbon genera diverged at approximately the same time. Under the instantaneous speciation model and an overall autosomal mutation rate of 1×10^{-9}/site/year (accounting for the removal of CpG sites by multiplying by 3/4) we placed this speciation process at 5.5 Mya (95% CI 2.5-10.3Mya), with an ancestral population size of 132,000 (95% CI 107,000-162,000) individuals (assuming 10 years per generation) (Fig. 4-B). Under the best bifurcating speciation model we placed this speciation process at 5.1 Mya (95% CI 2.5-7.7Mya), with an ancestral population size of 113,000 (95% CI 87,000-147,000) individuals. However, we note that a model with a large ancestral population size cannot be distinguished from a model of ancestral population structure. The parameter estimates from this method were also in line with a related coalescent-based analysis applied to the same non-genic loci

implemented in the software G-PhosCS, though CIs from the latter method were narrower because of the ability to directly calculate likelihoods (*Veeramah et al. submitted*, for details).

8.4 Allele sharing

We examined inconsistencies between our most likely bifurcating species topology and the genome-wide data that might suggest more complicated demographic scenarios involving post divergence geneflow and/or ancestral population structure. Allele sharing statistics under various filters demonstrated that while HLE and SSY were sister taxa, both NLE and the two Hylobates species shared more alleles with the SSY (Table ST8.5). A D-statistic analysis demonstrated that this difference was highly significant with Z-scores >14 (assessed using an m-delete jackknife with 5MB windows as in Green et al. 2010[143]. This could be explained by two separate geneflow events involving SSY (between both NLE and *Hylobates*) or ancestral population structure between ancestors of HLE and other ancestral gibbon groups. More details of allele sharing statistics and the D-statistic analysis are provided in *Veeramah et al. submitted*.

NLE	SSY	HLE	HMO/HPI	Count
0	1	1	0	264,675
0	1	0	1	212,083
1	1	0	0	198,146
1	0	1	0	190,535
1	0	0	1	186,668
0	0	1	1	185,938

Table ST8.5: Summary of pairwise allele sharing counts between genera for derived alleles (orientated by comparison to hg19). These counts were obtained by randomly sampling one allele from the two genotypes from a genus that met certain quality criteria. A more complete analysis using alternative criteria is given in *Veeramah et al. submitted* (though the patterns remained the same regardless of the criteria utilized)

8.5 Pairwise sequentially Markovian coalescent (PSMC) analysis

In order to examine how the effective population size of gibbons has changed over the past ~5 million years we analyzed each of the eight genomes using the Pairwise Sequentially Markovian Coalescent (PSMC) model described by Li and Durbin[69]. This model uses a Hidden Markov Model (HMM) that examines patterns of heterozygosity along individual genomes to infer the changes in the time to the most recent common ancestor (TMRCA) for sections of the genome, which in turn can be related to the distribution of N_e over time in the population. Sequence data were converted into 100 bp windows that described whether at least one heterozygous site existed in the window using a binary indicator. The sequence data was also filtered/masked as described above with all individuals needing to be called at a site. 100 bp windows with less than 90 called bases were set as missing. We applied the default parameters of the PSMC software that were previously described in Li and Durbin[69] with 64 atomic time intervals with the pattern 1*4 + 25*2 +1*4 +1.6 and maximum coalescent time of 15. The 20th iteration of the expectation-maximization algorithm was used as the final estimate of the change in population size over time. Bootstrapping was performed by dividing the genome into smaller segments using the 'splitfa' tool and conducting the PSMC analysis by randomly sampling these segments with replacement.

Estimated θ values from the PSMC model are shown in Table ST8.6, and Fig. SF8.4 shows the change in N_e over time assuming of $\mu = 1x10^{-8}$ per site per generation and a generation time of 10 years. Fig. SF8.5 shows the results of the analysis for all samples including 100 bootstrap replicates, Fig. 4C (main paper) shows the results of the analysis only for the sample with the highest coverage from each species and Figures SF8.6-10 show the bootstrap results for each individual species. In general individuals from the same species follow the same trajectory of N_e through time, though it is noticeable that the lower coverage samples from the 3 species pairs (NLE, SSY, and HLE) all show some flare at the most recent time scales, which is likely due to increased sequencing errors (see below).

Both the NLE and HMO show an increase in N_e beginning 500 kya before declining again 100 kya, while the N_e of the other species has stayed relatively low during this period (reflecting the level of

overall nucleotide diversity we see in these species as well as the estimated N_e from the ABC and G-PhoCS analysis which assume constant effective population sizes), though the actual dates and the magnitude of the size change is dependent on the phylogenetic mutation rate and generation time we assumed. Relaxing our filtering so that usable sites in an individual did not depend on all other individuals also being called at the site, results in a much higher N_e toward the older age range of ~4-5 million years (Fig. SF8.10). Nevertheless, unmasked segmental duplications presenting as long tracts of heterozygosity likely contribute to this observation as the effect is less pronounced in the NLE who will generally show less mapping biases for such features in our CNV and segmental duplication masks.

Similarly it is tempting to attribute the left hand side of the plot to a recent expansion amongst the SSY. However this result should be treated with caution as power is known to be low for the inference of population sizes at recent and ancient time frames (<20 kya in humans), while there a number of potential sources of error in our data that could affect our results, especially for the non-NLE samples. Coverage lower than <20X is known to bias PSMC inference[99] by underestimating true diversity (missing heterozygotes), while reference bias against non-NLE samples in combination with low coverage may result in both missing heterozygotes and even the introduction of false heterozygosity. Our high coverage exome validation results suggested that Monty, the lower coverage SSY sample at ~13X, may produce more than double the number of false heterozygotes than Vok, the higher coverage NLE samples at ~14X. In addition the estimate of θ from the PSMC analysis for the lower coverage SSY and HLE samples had higher heterozygosity than their higher coverage counterparts, but this pattern was not seen in the two NLE samples, again suggesting an increased false heterozygote error rate due to non-NLE reference bias. While missing heterozygotes will only affect the overall scale of the PSMC inference by reducing θ (our conditioning on all sites being called in all individuals also may have this effect), false heterozygotes have been shown by Li and Durbin[69] to artificially increase recent population sizes by fragmenting long segments of DNA, which may explain the patterns seen in our low

coverage samples compared to the high coverage counterparts with regard to recent time frames. Thus in the main manuscript we show only the high coverage sample from NLE, SSY and HLE.

Sample		θ
Vok	NLE	1.17E-03
Asteriks	NLE	1.14E-03
Monty	SSY	0.65E-03
Karenina	SSY	0.52E-03
Maung	HLE	0.25E-03
Drew	HLE	0.37E-03
Domino	HPI	0.25E-03
Madena	HMO	0.80E-03

Table ST8.6 θ for each sample as estimated by the PSMC model

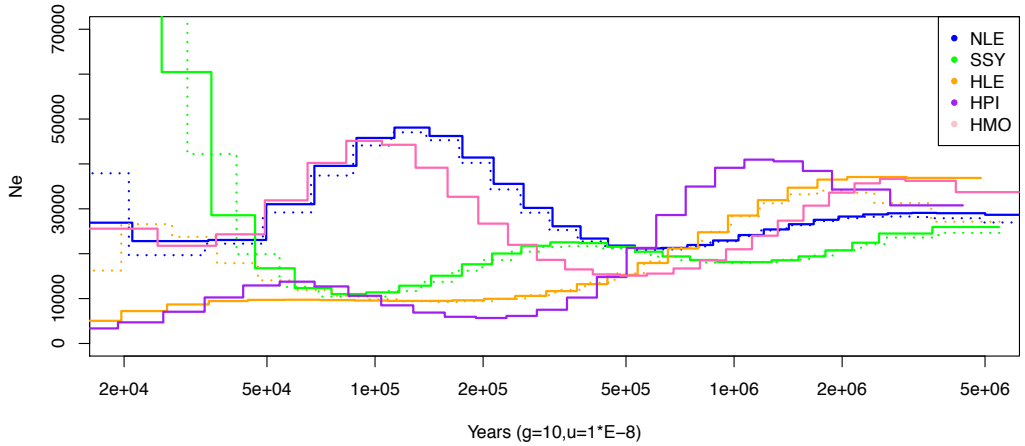


Fig. SF8.4 Change in N_e for gibbons as assessed by the PSMC model. Lower coverage samples from each species shown with dashed lines (NLE-Asteriks, SSY-Monty, HLE-Drew)

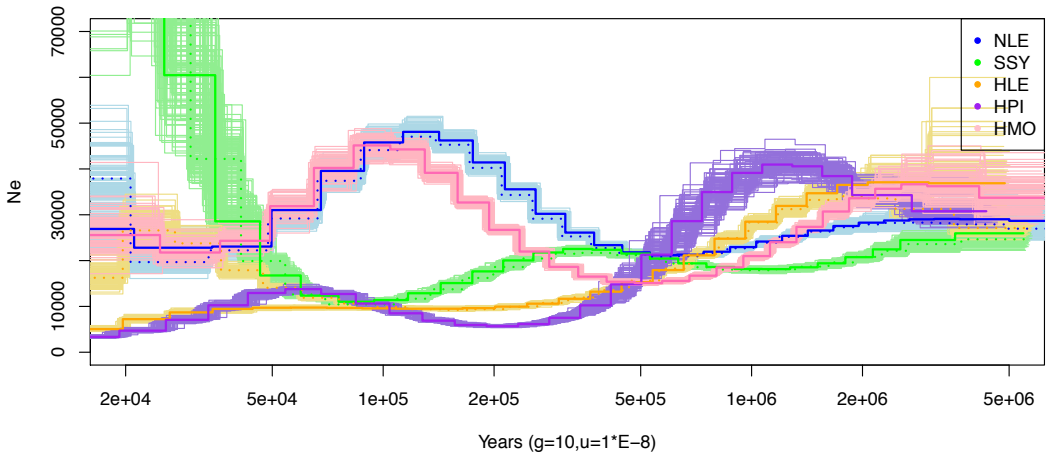


Fig. SF8.5 Change in N_e for gibbons as assessed by the PSMC model with 100 bootstrap replicates

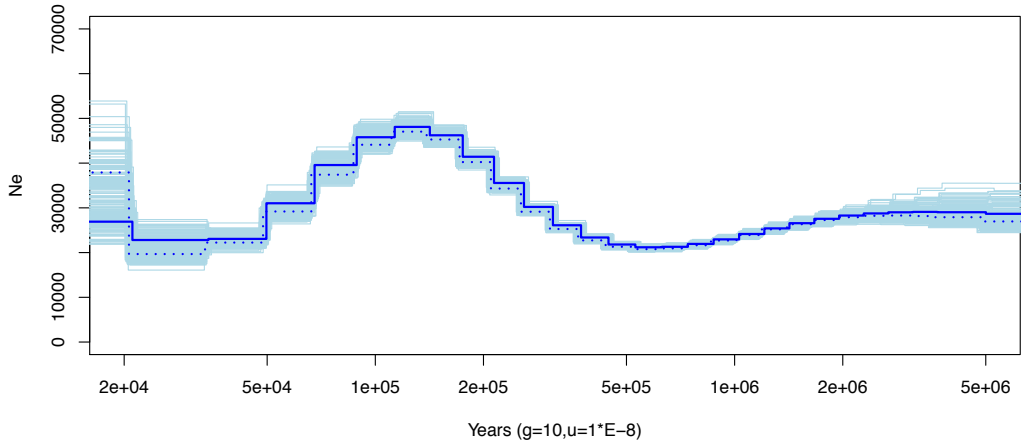


Figure SF8.6 Change in N_e for NLE as assessed by the PSMC model with 100 bootstrap replicate. (Asteriks = dashed line)

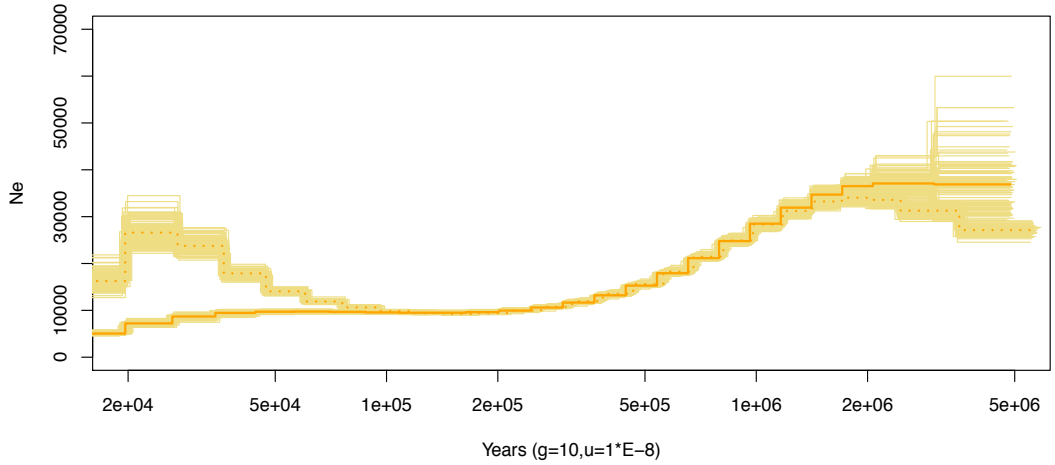


Fig. SF8.7 Change in N_e for HLE as assessed by the PSMC model with 100 bootstrap replicates. (Drew = dashed line)

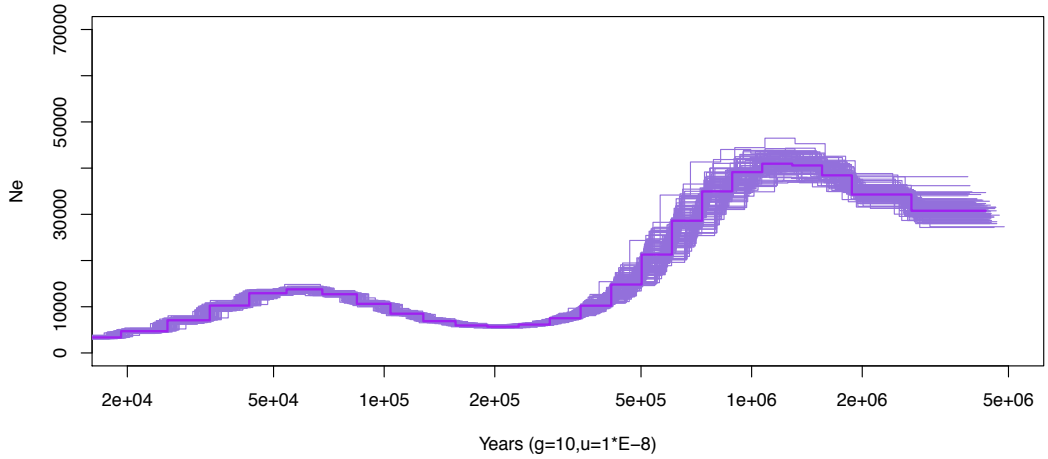
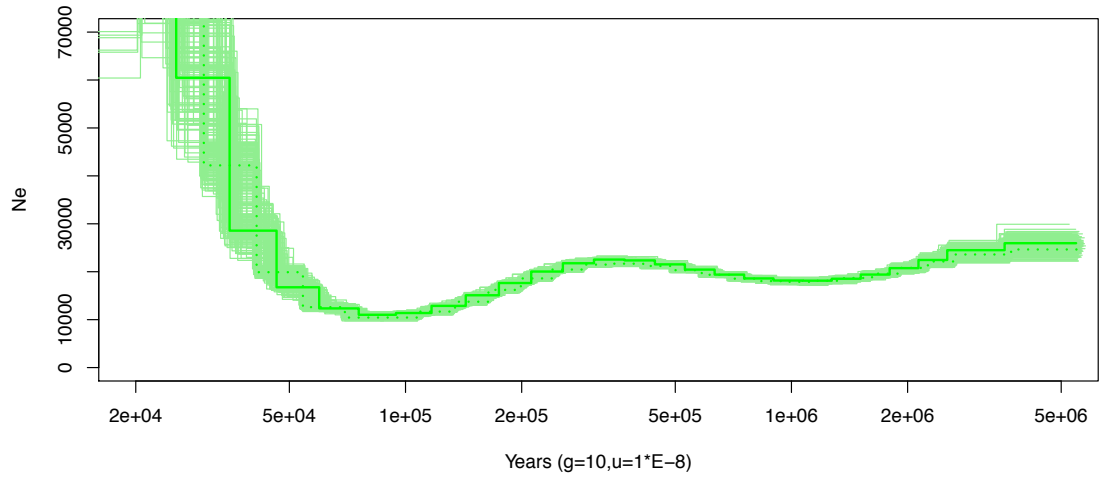


Figure SF8.8 Change in N_e for HPI as assessed by the PSMC model with 100 bootstrap replicates

a



b

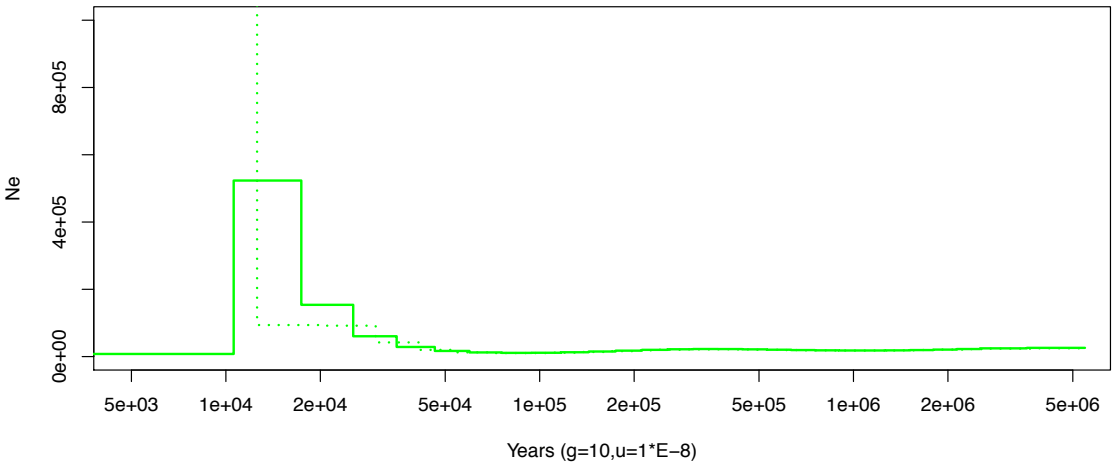


Figure SF8.9 Change in N_e for SSY as assessed by the PSMC model (a) using same scale as other plots and with 100 bootstrap replicates, and (b) using a scale to show the extension of the PSMC line to a unfeasibly high N_e during recent time frames for this particular species.

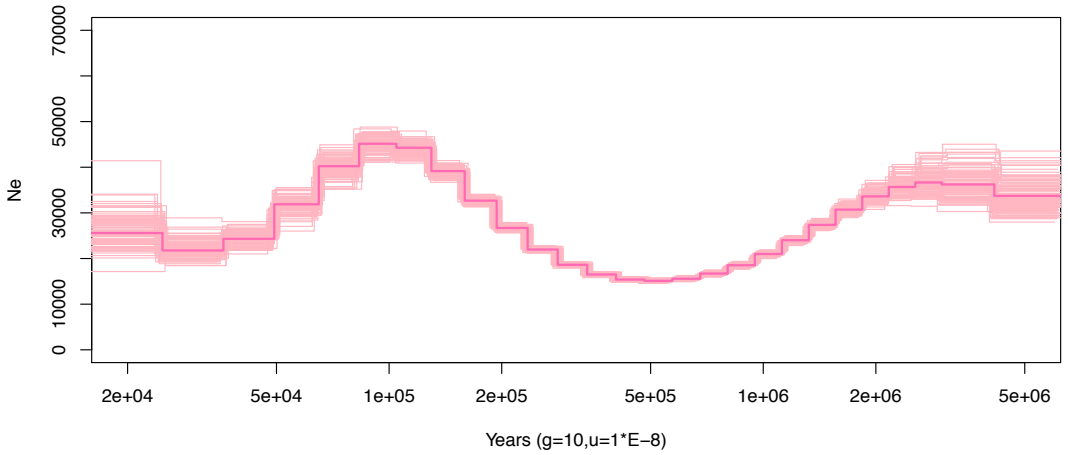


Fig. SF8.10 Change in N_e for HMO as assessed by the PSMC model with 100 bootstrap replicates

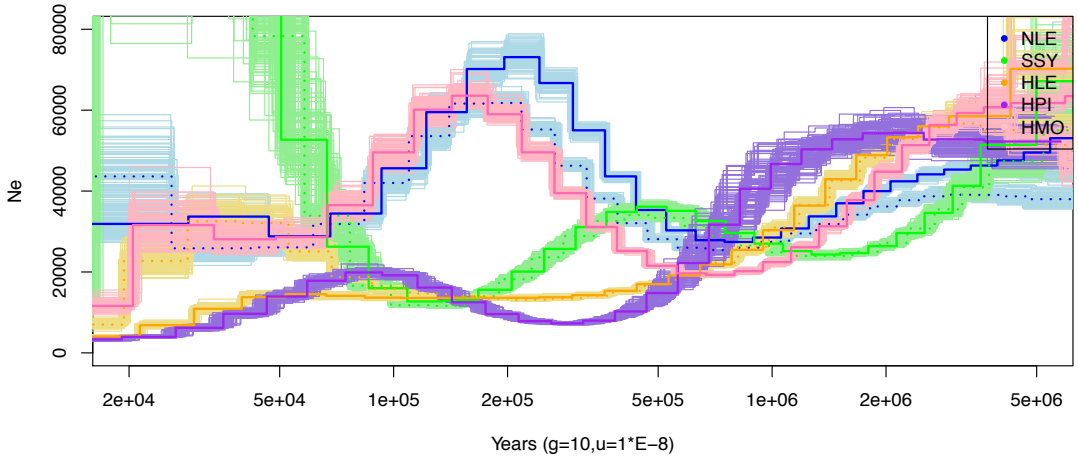


Fig. 8.11 Change in N_e for gibbons as assessed by the PSMC model using relaxed filtering criteria per individual with 100 boostrap replicates.

8.6 What is the gibbon mutation rate?

There is an ongoing debate regarding the appropriate mutation rate for anatomically modern humans[139-144] that has potential knock-on effects for the likely mutation rate in gibbons, which will then affect downstream estimates of divergence times and N_e. We do not have the data available to solve this debate, but we present the various issues such that the reader is aware of the caveats that are inherent in our estimates of these parameters.

As mentioned above in S4, a hominid slow down has been invoked to explain the discrepancy between phylogenetic estimates of the human mutation rate based on chimpanzee or old world monkey divergence (~1 x 10^{-9} per site per year, which can be converted to ~2.5 x 10^{-8} per site per generation assuming 25 years per generation[145-146] and whole genomes sequencing of trios (~1.1 x 10^{-8} per generation, which translates to ~0.5 x 10^{-9} per site per year[141]). Under this scenario, the long-term phylogenetic mutation rate may better represent (somewhat crudely) primates of lower body size or shorter generation times (for example Macaques), while the pedigree estimate is clearly applicable at least to modern humans and perhaps other great apes with larger bodies and longer generation times. Gibbons lie intermediate of macaques and great apes with regard to body size and generation time. We have assumed in this study that the phylogenetic mutation rate is more appropriate for gibbons. Assuming a generation time of 10 years[147]) as in a recent analysis of RADseq data by Kim et al. 2012[82](though 15 years is also plausible[144], this translates to a mutation rate of ~1 x 10^{-8} per site per generation (it is simply coincidence that this matches the recent human-based pedigree estimates). However, it would not be unreasonable to suggest that the mutation rate is in fact slower than this if gibbons were affected by the hominid slowdown at least to some extent. The effect of this would be to increase our divergence times and N_e estimates from our ABC and PSMC analysis (perhaps even doubling them depending on the extent of the proposed slowdown). Unfortunately there is currently no data on mutation rates from gibbon pedigrees to determine the correct approach and thus this adds substantial uncertainty to our parameter estimates that readers should be aware of.

Supplemental Section S9 – Phylogenetic analysis using mitochondrial DNA

9.1 Obtaining the mitochondrial sequences

We have reconstructed the mitochondrial genome (mtDNA) of the 8 gibbons of this project from the whole-genome shotgun (WGS) paired-end sequencing data of these individuals. For each sample, we retrieved a subset of the sequenced reads by mapping the whole set of paired-end reads (Table ST2.2) against a mitochondrial reference of the same species. We performed the alignments using the BWA aligner with parameters –n 6 –q 15[16]. We retained only high-quality paired-end reads by imposing that both pairs should be properly paired mapped (with samtools –f 2[69]) and by requiring that both pairs have a median Phred quality score greater than 32.

With the intention of increasing the number of reads that map to the extremes of the assemblies we applied a second round of read capturing in which we took advantage of the mtDNA circularity. This second time we aligned reads to a modified sequence assembly, in which the origin of the reference assembly was changed at the middle of the mtDNA (8 Kbp from the start).We constructed contigs from the captured reads with Hapsembler v.1.1. (-p Illumina -t 4 -d no --PHRED_OFFSET 33 --MIN_CONTIG_SIZE 1000 –EPSILON 0.05)[100], a haplotype-specific genome assembly toolkit. The efficiency of this assembler decreases when dealing with high coverage input data such as the ones obtained for the mitochondria sequences (Table ST9.1).

Genus	Species	Sample	Total Reads	Coverage*
Nomascus	*Nomascus leucogenys*	Asteriks	287,182,500	9.77x
		Vok	483,613,856	16.54x
Symphalangus	*Symphalangus syndactylus*	Karenina	782,617,756	26.76x
		Monty	434,653,704	14.90x
Hoolock	*Hoolock leuconedys*	Drew	707,839,272	24.20x
		Maung	783,804,634	26.66x
Hylobates	*Hylobates moloch*	Madena	376,780,080	12.91x
Hylobates	*Hylobates pileatus*	Domino	485,754,996	16.53x

Table ST9.1 Initial WGS reads used to reconstruct the mtDNA.

We, therefore, reduced the number of reads per sample to around 350X of mitochondrial coverage (for Karenina and Monty, that already have lower coverage, no reduction was done, and for Asteriks we re-sampled reads to have around 250x). However, this random representation of reads entails potential related problems such as the assembling of existing *numts* into the mitochondrial sequence. To compensate the randomness of the resampled data, we applied the reduction of coverage and posterior construction of contigs 20 times per reference assembly (the standard and the one with the origin changed).

Sample	Standard Assembly		Modified origin Assembly		Length mtDNA	Length mtDNA wo D-loop
	Total Reads	Coverage*	Total Reads	Coverage*		
Asteriks	64,594	372.13	64,322	370.56	16,481	15,446
Vok	572,376	3,297.50	570,562	3,287.05	16,477	15,446
Karenina	53,156	305.75	52,886	304.20	16,514	15,451
Monty	25,540	146.91	25,452	146.40	16,517	15,451
Drew	301,276	1,732.94	298,726	1,718.27	16,500	15,453
Maung	352,514	2,027.66	350,462	2,015.86	16,494	15,447
Madena	151,950	874.76	151,554	872.48	16,501	15,453
Domino	180,476	1,040.24	179,870	1,036.75	16,501	15,454

Table ST9.2 Number of captured high-quality reads and mitochondrial coverage per sample.
**Coverage relative to the length of the corresponding gibbon mitochondrial reference.*

For each of the 40 iterations of read reduction and subsequent assembling, we oriented the resultant contigs via local alignments to the corresponding reference genome (with BLAST[54]) and joined them using MAFFT[101] including N's in the existing gaps. We incorporated N's, also, in those positions where contigs overlap and differ. Thus, we reconstructed 40 mitochondrial assemblies per sample. The final

mitochondrial sequence per individual is the consensus sequence. We were able to reconstruct the complete mitochondrial genomes without gaps of the eight individuals. The sequence lengths are shown in Table ST9.2.

Validation of the mitochondrial sequences

Four out of the 8 individuals (Asteriks, Modena, Drew and Madena) were independently sequenced using traditional Sanger sequencing. Therefore, two overlapping LongRange-PCR products of the mtDNA were generated using primers and methods outlined in[102]. Subsequently, overlapping nested PCRs with product lengths of 1.0-1.2 kb were amplified with various primer sets and sequenced in both directions with respective primers. The resultant sequences are identical from both next-generation and Sanger sequencing.

9.2 Phylogenetic analysis

The 8 gibbon mtDNA sequences were aligned together with human (X93334), chimpanzee (D38113), bonobo (D38116), gorilla (NC_011120), Borneo orang utan (NC_001646) , Sumatra orangutan (NC_002083) and rhesus monkey (NC_005943) with Muscle 3.7[103] and manually checked. The complete alignment had a length of 16,787 bp. Two datasets were generated from the original alignment: in the first (dataset 1) poorly aligned positions and indels were eliminated with Gblocks 0.91b[104], while in the second (dataset 2) also the D-loop region was removed. Accordingly, dataset 1 and dataset 2 had alignment lengths of 15,957 bp and 15,164 bp, respectively.

The best-fit model for both datasets was the TPM2u+I+G model as calculated by the Bayesian Information Criterion (BIC) in jModeltest 2.1[105]. Phylogenetic trees were constructed with maximum-likelihood (ML) and Bayesian algorithms in GARLI 0.951[106] and MrBayes[107], respectively. In GARLI, only the model specification settings were adjusted, while all other settings were left at their default values. In total 500 bootstrap analyses were performed and a 50% ML consensus tree was calculated in PAUP 4.0b10[108]. For Bayesian analyses, four Markov Chain Monte Carlo (MCMC) runs with the default temperature of 0.2 were applied. Repetitions were run for 1 million generations with tree and

parameter sampling occurring every 100 generations and 25% of samples were discarded as burnin. The adequacy of the burnin and convergence of all parameters was assessed with the software TRACER 1.5 (http://tree.bio.ed.ac.uk/sofware/tracer/) and via the uncorrected potential scale reduction factor (PSRF)[109] as calculated by MrBayes. AWTY[110] was used to check whether posterior clade probabilities converged properly. Posterior probabilities and a phylogram with mean branch lengths were calculated from the posterior density of trees.

9.3 Divergence age estimation with BEAST

Divergence ages were calculated with a Bayesian MCMC method in BEAST 1.6.1[111,112]. We applied a relaxed lognormal model of lineage variation and a Birth-Death Process prior for branching rates. Four runs each with 10 million generations with tree and parameter sampling occurring every 100 generations was performed. The adequacy of a 10% burnin and convergence of all parameters were assessed by visual inspection of the trace of the parameters across generations in TRACER. Subsequently the sampling distributions were combined (25% burnin) using LogCombiner 1.6.1 and a consensus chronogram with node height distribution was generated and visualized with TreeAnnotator 1.6.1 and FigTree1.4 (http://tree.bio.ed.ac.uk/software/figtree/). As calibrations, we applied three fossil-based divergences: (1) the split between human and chimpanzee/bonobo 6-7 Mya[113-115], (2) the split between *Pongo* and African great apes and humans ca. 14 Mya[116], and (3) the divergence of Old World monkeys (rhesus monkey) from apes (great apes, small apes, humans) 24-29 Mya[117,118]. Instead of hardbounded calibration points, respective dates were used as normal distribution priors. Accordingly, calculations were calibrated with (1) 6.5 ± 0.5 (SD 0.3) Mya, (2) 14.0 ± 1.0 (SD 0.6) Mya, and (3) 26.5 ± 2.5 (SD 1.5) Mya.

Results

Tree reconstructions from both datasets and using ML and Bayesian algorithms revealed a strongly supported monophyly of the Hylobatidae family and each of the four gibbon genera as well consistently

strongly supported branching patterns among great apes and humans (Fig. SF9.1a-b). However, the branching pattern among the four gibbon genera remained unresolved and statistical support is mostly low. Moreover, the phylogenetic relationships among genera differ between datasets. In the phylogeny derived from dataset 1, *Nomascus* is indicated as basal, followed by *Hylobates*, while *Symphalangus* and *Hoolock* were the last to diverge. In dataset 2, *Hylobates* and *Symphalangus* are suggested as sister genera, *Nomascus* is indicated again as basal genus.

Estimated divergence ages from both datasets are similar and in agreement with various published dates derived from fossils or other molecular studies (Fig. SF9.1a-b, Table ST9.3)[34,35,102,119-125]. Accordingly, gibbons separated from great apes/humans ca. 19 Mya, which is in line with the estimates derived from nuclear data, while the four gibbon genera diverged from each other in a short time period 5-7 Mya. *H. pileatus* and *H. moloch* separated ca. 3 Mya.

Split	**Dataset 1**	**Dataset 2**
rhesus monkey – hominoids	27.90 (25.08-30.73)	27.65 (24.82-30.52)
hylobatids – hominids	18.66 (16.62-20.90)	19.10 (16.62-21.93)
Pongo – Gorilla+Pan+Homo	13.86 (12.80-14.87)	13.92 (12.89-15.02)
Sumatra – Borneo orang-utan	3.66 (2.61-4.72)	3.63 (2.28-4.95)
Gorilla – Pan+*Homo*	8.17 (7.20-9.23)	8.22 (7.16-9.42)
Pan – Homo	6.25 (5.69-6.83)	6.28 (5.70-6.86)
chimpanzee – bonobo	2.12 (1.50-2.77)	2.07 (1.33-2.90)
Nomascus – other hylobatids	6.86 (5.65-8.26)	6.59 (5.12-8.16)
Hylobates – Hoolock+Symphalangus	5.89 (4.80-7.03)	-
Hoolock – Symphalangus	5.21 (4.12-6.33)	-
Hoolock – Hylobates+Symphalangus	-	6.06 (4.65-7.52)
Hylobates – Symphalangus	-	5.56 (4.25-7.03)
MRCA *Nomascus*	0.07 (0.04-0.11)	0.06 (0.02-0.09)
MRCA *Hoolock*	0.27 (0.18-0.38)	0.24 (0.15-0.35)
MRCA *Symphalangus*	0.26 (0.18-0.36)	0.22 (0.13-0.31)
Hylobates moloch – H. pileatus	3.03 (2.23-3.79)	2.89 (1.94-3.87)

Table ST9.3 Estimated divergence ages in million years ago (95% highest posterior density) as obtained from the two mitochondrial datasets.

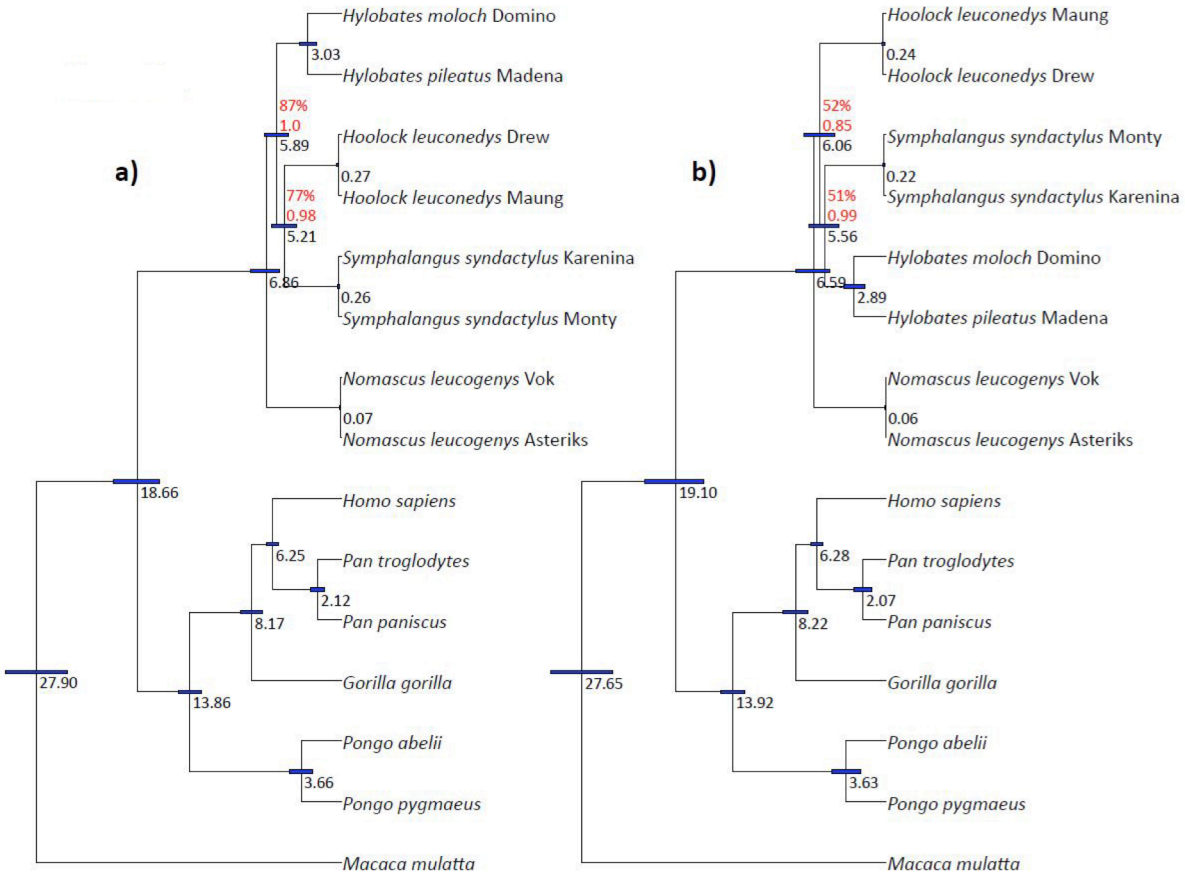


Figure SF 9.1 Ultrametric tree showing phylogenetic relationships and estimated divergence ages among gibbons based on mtDNA (a. dataset 1, b. dataset 2). Node support <100% and <1.0 is given at respective branches as red numbers (ML-Bayesian). Divergence ages are given in million years ago; the blue bars indicate 95% highest posterior densities of estimates (see also Table ST9.3)

The divergence of the four gibbon genera between 5 and 7 Mya is fairly easy to interpret with respect to the evolution of regional topography and environments in southwestern China at the time. This was a period of exceptionally rapid and extreme change in elevation. Existing topographical relief (the Heng Duan Mountains) was exaggerated by the Himalayan uplift and the increased rate of river incision, leading to increased environmental complexity and heterogeneity, and increased species richness and frequency of vicariants. Steep elevational and climatic gradients existed, and isolation of gibbon habitats was promoted by deeply incised rivers, which formed trenchant boundaries between regions that affected most non-vagile terrestrial species. From the Late Miocene onward, regional uplift, tilting,

and subsidence led to the creation and elimination of basins in southwestern China, creating a perfect "crucible" for speciation by vicariance. As for the divergence of *H. pileatus* and *H. moloch* ~3.0 Mya, there is no precise geological or biogeographical event that one could point to. However, we know that during the Late Pliocene, climatic deterioration resulting from intensification of the winter monsoon is occurring, and is leading to increasingly cold and dry conditions and the growth of glaciers in northern latitudes and on mountains. The historical details of the climate, environment and land bridges of Sundland are complex and still poorly understood. It is possible that climatic deterioration caused retraction of proto-*pileatus* and proto-*moloch* populations to respective core areas in Southeast Asia and Java. Subsequent genetic exchange was precluded by inundation of the land bridge.

Supplemental Section S10 – Genic positive selection

10.1 Ensembl gene trees and orthologs

We have performed a genome-wide phylogenetic analysis to infer the orthology relationships for the gibbon genes. The species used for this analysis include 10 primate genomes (*Homo sapiens*, *Pan troglodytes, Gorilla gorilla gorilla, Pongo abelii, Nomascus leucogenys, Macaca mulatta, Callithrix jacchus, Tarsier syrichta, Microcebus murinus, Otolemur garnettii*) as well as most of the high-quality vertebrate genomes. We have also added a few outspecies, namely like *Ciona intestinalis*, *Ciona savignyi*, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae*. We have used the Ensembl GeneTree pipeline, a fully automated approach (Beal et al, in prep.). In short, we cluster the proteins using hcluster_sg (http://treesoft.svn.sourceforge.net/viewvc/treesoft/branches/lh3/hcluster/) based on their BLAST e-value[54], we align the proteins with M-Coffee[126] and we infer the phylogenetic trees with TreeBeST (http://treesoft.sourceforge.net/treebest.shtml). From the final set of trees, we infer orthology and paralogy relationships. We use Mafft[101] instead of M-Coffee when the alignment is too difficult complex and these exceptions are built into the pipeline. Also, to cope with large trees, we recursively split them with QuickTree[127] until they reach a reasonable size (less than 400 genes). They then follow the normal procedure (alignment, tree building, homology inference). The final set of trees include 20,300 trees, 1,127 of them representing the sub-trees in the 440 super-trees Table TS10.1 shows a comparison of the number of orthologous genes between the human genome and the orangutan, gibbon and macaque genomes. In evolutionary terms, compared to gibbon orangutan is closer to human while macaque is further away. The number of one-to-one orthologous genes is quite similar in all 3 cases. These are orthologous such that one (and only one) human gene is related to one (and only one) gene in the other species. We also look at the total number of orthologous genes including more complex relationships and observed a slight reduction of these for the gibbon genome. This may be the result of differences in either the assembly or annotation quality. All the trees and homologies are available in

Ensembl 70 (http://e70.ensembl.org/) and can be downloaded from the corresponding FTP site in several formats (ftp://ftp.ensembl.org/pub/release-70/xml/ensembl-compara/homologies/).

10.2 Clusters of primate orthologs

For the PAML analysis, we have defined clusters of primate orthologs (human, chimpanzee, gorilla, orangutan, gibbon, macaque, marmoset). Because we want to allow only lineage-specific duplications in these sets, we used a single linkage approach. In a cluster, we do not require that all pairs of genes are orthologous to one another, but make it sufficient if both are orthologous to a third gene. This means that a cluster can contain, for instance, two human genes if they arose from a recent duplication, as they may have the same gibbon orthologs. However, we would generate two different clusters for two separate primate sub-families if they originated before the speciation of primates.

Using this method we defined 20,867 clusters (spanning 136,892 genes in total from only 10,555 trees), of which 17,479 contain at least a gibbon gene, and 14,644 at least a gibbon gene and a non-gibbon gene. For each cluster of orthologous genes, we extracted the sub-alignment from the original gene tree and removed the columns that contained only gaps. While it is possible to build a new alignment with the sequences in each cluster only, we opt for using the existing alignment. This is because the full alignment is built with the sequences from the other species and they provide additional signal that is useful to resolve the alignment.

10.3 Detecting genes under positive selection

The set of 20,867 orthologous primate alignments obtained from ENSEMBL and containing sequences from marmoset, macaque, gibbon, orangutan, gorilla, chimpanzee, and human was parsed to obtain only the 13,638 one-to-one orthologs that contained a gibbon sequence. The one-to-one alignments were filtered by the base quality score of the gibbon sequence, with any columns removed that had a score less than 20. To ensure the use of high quality alignments, a custom alignment masking algorithm was implemented, based upon the one employed by Han et al. (2009)[128]. The algorithm uses a sliding window of five codons to check each position of the alignment. Within each five-codon window, three sub-windows of three codons exist and if any of these sub-windows contained two or

more codons with two or more substitutions, the entire five codon window was masked. This effectively detects and removes runs of substitutions that are likely due to poor alignment, and that may lead to detection of positive selection where none exists. This algorithm was implemented in three separate instances, each defining a substitution differently. First, substitutions were defined by polarizing the gibbon branch to remove exons that were shown to be mis-aligned in gibbon only. Second, without restricting substitutions to a single species, any position in the alignment that contained less than 100% sequence similarity was considered to have a substitution. Finally, masking was done with substitutions defined using both approaches described above to capture any mis-alignments which may have been missed during the initial process.

Using the PAML software package, the branch-site test for positive selection was performed with gibbon as the foreground branch on each of the three masking datasets. Because not every alignment contained a sequence from each of the seven species, 44 variations of the primate tree were used, with the most commonly used tree being the full tree (n=11,263 trees). The next most used trees were missing just one species, with the tree missing only chimp used 361 times and the tree missing only gorilla being used 354 times and so forth. A conservative likelihood-ratio cutoff of 5.99 (p-value <0.01) was used to determine significance under the branch-site test. The branch-site test was run once on each of the three masking datasets in their entirety and then once again on each gene that was significant in the first run. Genes that were still significant in all three sets were then run again to ensure convergence by PAML, and this was the final set of genes determined to be under positive selection (Supplementary File 7 and 8). The multiple masking implementations and PAML runs ensure a conservative estimate of genes under positive selection.

10.4 Gene ontology (GO) term analysis

GO terms for all 13,638 genes were retrieved from ENSEMBL and enrichment analysis was done using Fisher's exact test. No GO terms were found to be enriched after correcting for the number of tests (based on a Dunn-Sidak correction) in the set of genes under positive selection. We report the top 5

(not significant) enrichment functional categories and compare our findings with what has been published about other primates. Specifically, Supplemental File 7 shows a table published as part of the gorilla genome in which we have included findings from this project. In addition, seven genes whose proteins localize to the centrosome have been under positive selection in gibbons (e.g., CSK (c-Src tyrosine kinase), p-value=2.42E-06 and KIAA1731, p-value=0.0019) (Supplementary Files 8 and 9). We also looked across functional categories but found no significant enrichment.

	Human vs. Orangutan	Human vs. Gibbon	Human vs. Macaque
1-to-1 orthologues	16,843	16,431	16,158
All orthologues	18384 / 18494	18010 / 17632	19232 / 18314
Average % of identity	89.1655	89.4731	85.6799

Table ST10.1 - Comparison of orthologous between the human genome and the orangutan, gibbon and macaque genomes. On the "all orthologous" line, the two numbers in each cell refer to each species (human first) and are different because of the 1-to-many and many-to-many relationships.

Supplemental Section S11 – Gene family analysis

The gene family analysis was performed using CAFE 3.0 and relied on the Ensembl gene annotations of *Nomascus leucogenys* together with the Ensembl gene sets from 10 other mammalian genomes: human, chimpanzee, orangutan, rhesus macaque, common marmoset, mouse, rat, dog, horse and cow. CAFE 3.0 infers rates of gene gain and loss using a two-phase model: in the first phase, errors in the genome assembly and annotation are taken into account by averaging over possible errors in gene family sizes at the tips of the phylogenetic tree. In the second phase, unknown ancestral states are taken into account by averaging over possible states using a birth-death stochastic model[129].

We applied the following ultrametric phylogenetic tree:

(((((((chimp:6,human:6):7,orang:13):6,gibbon:19):5,macaque:24):16,marmoset:40):47,(mouse:17,rat:17):70):6,((dog:74,horse:74):9,cow:83):10) wherein the numbers represent million years. A total of 210,853 genes grouped in 8.803 gene families were included in this comparative analysis (Table ST11.1).

	All Ensembl genes	Ensembl Genes used in the CAFE analysis
Gibbon	**18,267**	**17,248**
Human	20,424	19,083
Chimpanzee	18,592	17,540
Orangutan	19,900	18,305
Rhesus	21,260	19,526
Marmoset	20,687	19,767
Mouse	21,947	20,436
Rat	22,604	21,773
Dog	19,694	18,951
Horse	20,129	18,814
Cow	19,839	19,410

Table ST11.1 Genes retrieved from the Ensembl database and genes used in the CAFE analysis for the eleven mammalian species

Gene families were required to contain at least one gene in one or more species of the two orders primates and rodents, and in the laurasiatheria (dog, horse and cow).

Estimated rates of gene duplication and loss are best fit by a CAFE model with three levels of gene family evolution (Fig. SF11.1). At a rate of ~0.00596 gene duplications and losses/million years, the human and chimpanzee lineages showed the fastest evolution of gene family, which has been previously reported[16,130]. Gene families in the gibbon branch appeared to have evolved at an intermediate rate of 0.00344 duplications and losses/My, the same estimated for the remaining primates except marmoset, which shared with the non-primate species a slower rate of 0.00129 duplications and losses/My.

Overall, we observed 707 rapidly evolving families across the mammal tree (P<0.01), 135 of which were rapidly evolving on the gibbon lineage. However, only 133 of them represented family contractions, suggesting that several functional genes have been not annotated in the NLE genome assembly. A closer inspection of the human-gibbon syntenic region corresponding to potentially missing gibbon genes revealed several instances where such genes might be entirely or largely contained in gaps present in the Nleu1.1 assembly. Accordingly, the number of genes used in the CAFE analysis after applying the aforementioned filters is the lowest in *N. leucogenys* (Table ST11.1).

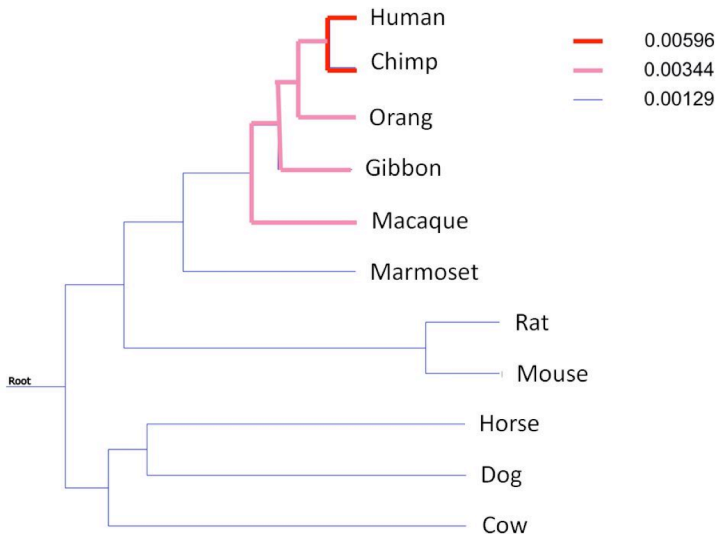


Figure SF11.1 Rates of gene duplication and loss in gene families of eleven mammals, including the gibbon Nomascus leucogenys.

Supplemental Section S12 – Gibbon accelerated regions (gibARs)

12.1 Determining conserved elements / alignment filtering

Alignment files

In order to detect gibbon-specific accelerated regions (gibARs) we first obtained unusually conserved sequence elements. To that end, conserved genomic regions based on phylogenetically deep alignments of 46 vertebrates were downloaded from the UCSC genome browser (phastConsElements46wayPlacental track, www.genome.ucsc.edu). Gibbon is not part of the 46way alignment, so no gibbon sequence was used to infer these conserved regions. For further analysis we used a gorilla (gorGor3) referenced alignment of eleven primates, including Gibbon. To obtain gorGor3 coordinates for conserved sequence elements GRCh37-referenced regions (50 bp or longer) were mapped to gorilla using liftOver with standard parameters; mapped regions less than 5 bp apart were merged. This resulted in 490,194 gorilla-reference conserved elements, for each of which we obtained alignment files from the UCSC genome browser (http://hgdownload.soe.ucsc.edu/goldenPath/gorGor3/multiz11way/).

Alignment processing

Alignments corresponding to conserved regions were extracted, and bases with a quality score less than 50 in Gibbon (Nleu1.0) were masked. Alignments with blocks not containing all of GRCh37, panTro3, gorGor3, ponAbe2, nomLeu1 and rheMac2 were discarded. Additionally, alignments with blocks containing sequence from either of the above-mentioned species that mapped to more than one location in gorGro3 (the reference) were discarded. This ensures a 1:1:1:1:1:1 relationship between all regions we considered in human, chimpanzee, orangutan, gorilla, gibbon and macaque and only left high-quality bases in gibbon.

Additional filters

To additionally guard for alignment artifacts we utilized the filters used for human accelerated regions[131], which include human-macaque-dog-mouse synteny requirements. To do so, we excluded alignments in which the included human sequence did not pass these filters. Additionally, we excluded regions overlapping genomic coordinates annotated as:

- Part of WGAC or WSSD in Gibbon (from Section 3).
- Repetitive sequence in the Nleu1.0/nomLeu1 repeat masker (rmsk) or simple repeat tracks from the UCSC genome browser.
- Gibbon pseudogenes annotated in Ensembl v69.

Impact of filtering steps on conserved regions

After filtering we retain 115,623 candidate regions for substitution rate acceleration in gibbon. Below the effects of different filtering steps are summarized:

- 490,194 phastCons regions are longer than 49 bp and map to gorGor3.
- -349,726regions additionally meet the species and 1:1 constraints
- 349,133 of these do not overlap pseudogenes
- -315,937 are not annotated as repeats
- -314,412 don't overlap WGACs or WSSDs
- -128,161 pass the HAR[131] filters

12.2 Identification of gibbon accelerated regions (gibARs)

To identify gibARs we used the framework of likelihood ratio tests (LRTs) to identify lineage specific substitution rate acceleration[132]. We processed the alignment file for each of the 115,623 candidate regions and used the phylofit function of the rphast package[133] to fit the corresponding null and an alternative models to this data. We downloaded the generalized reversible model used to generate conservation tracks from the alignments underlying our analysis from the UCSC genome browser. The null model then consisted of this neutral phylogenetic model plus a global scaling parameter for all branches in the species tree (to account for the fact that the candidate regions were conserved across

mammals), while the alternative model additionally included a scaling parameter larger than one for the gibbon branch (to allow for a possibly accelerated rate of substitutions on the gibbon lineage since divergence from the common ancestor of gibbons and great apes). After model fitting, the respective likelihoods *L1* (null) and *L2* (alternative) were used to calculate the likelihood test statistic $T = 2(L2–L1)$ and p-values were obtained from the asymptotic null distribution[134]. Correction for multiple testing was performed and q-values were calculated with the multtest package in R using the Benjamini-Hochberg (BH) procedure. To assess if the excess substitutions in gibARs are consistent with the process of GC-biased gene conversion (gBGC)[135], which can mimic positive selection, we used a phylogenetic modeling approach[136]. This method essentially substitutes the parameter for unbiased, accelerated substitutions in the gibbon lineage with a gene conversion disparity parameter. In our results we report adjusted p-values from both analyses, substitution rate acceleration and gBGC.

Masking of clustered substitutions

Analyzing parsimony-inferred substitutions in conserved elements we observed that in Gibbon many lineage-specific substitutions are closer to another lineage-specific substitution than we observe in parallel analyses of human-specific substitutions. Further, if we annotate substitutions with LRT p-values (see above) of the genomic regions they are in, we observe that clustered substitutions have an excess of low p-values for rate acceleration in Gibbon. These observations suggest that the LRT analysis is picking up some regions with sequencing, alignment or assembly errors in gibbon that are not genuinely fast evolving. Therefore, in addition to quality masking, we masked all gibbon-specific substitutions that had other substitutions less than two base pairs away, and then we re-calculated LRT statistics and raw p-values based on these newly masked alignments. This prevented tight clusters of substitutions to contribute to the inference of gibbon-specific substitution rate acceleration. We then discarded all regions with new raw p-values larger than the previous un-masked maximum raw p-value obtaining a FDR of 15%. This filter is a conservative approach to avoiding false positive gibARs.

12.3 Analysis of gibARs

Gibbon accelerated regions

Applying the above procedure we identified 240 Gibbon accelerated regions (gibARs), (Supplementary File 9). On average gibARs are 153 bp long (114 bp is the median), with an IQR between 67 bp and 212 bp and a minimum and maximum of 50 bp and 715 bp, respectively. This is longer than the conserved candidate regions, (median = 80 bp and and IQR = 61 bp to 122 bp), which is expected because the LRT has increased power for longer sequences.

GC-biased gene conversion (gBGC)[135] is unlikely to be a major confounder to the gibARs we report, because of the 240 gibARs only 21 (less than 10%) have significant evidence for gBGC in our LRT (FDR<10%)[136]. Multiple testing adjusted p-values for the gBGC-LRT are available for all gibARs in Supplementary File 9. No masking of clustered substitutions was performed prior to these gBGC-LRTs, so these q-values are less conservative than the corresponding quantities for substitution rate acceleration (on which the gibARs are based). This, in turn implies, that 10% is a conservative estimate for the prevalence of gBGC amongst the gibARs we report.

Genic distribution of gibARs

We identified the nearest protein-coding gene for each gibAR. For 240 gibARs, the nearest gene was annotated on the same contig as the gibAR. For gibAR_227, we used the UCSC genome browser xenoRefGene table (based on refSeq genes of non-Gibbon species), because gibAR_213 is located on a contig (GL397671 in Nleu1) that contains no Ensembl v69 gene predictions. The xenoRefGene table showed that gibAR_213 is located in the second exon of the *Robo2*-gene.

The majority of gibARs are intergenic (158 out of 240, i.e. 66%), which is a significant enrichment compared to the candidate regions (49%, p=5.99E-8, Fisher's exact test). Correspondingly exonic regions are depleted amongst gibARS (29 out of 240, i.e. 12%), p=4.1E-8 Fisher's exact test, (Extended Data Fig. 6). UTRs and upstream/downstream regions are similar in gibARs compared to the candidate regions.

Genes containing gibARs

Of the 240 gibARs 82 are located inside protein coding genes, and 29 of those 82 are located in the exons of 28 corresponding genes with hgnc symbols: AMN1 ATXN1 CDH11 CHCHD3 CRH DENND5B DLK1 DNM1L DPH1 DUS2 EIF3D ELAC2 EPC1 GABRA1 GBF1 GPR180 HPX INHA LRRC7 MNS1 OFCC1 PKNOX2 PLD1 POLR2B SPG11 SRSF7 WDFY2 XXYLT1. Two genes, CNTN4 and MAGI1, each harbor two gibARs, but not in exonic sequence. The closest protein-coding gene for each gibAR (and its distance) can be found in Supplementary File 9.

Gene Ontology Annotation of gibARs

For Gene Ontology enrichment testing we looked for genes within +/- 100kb of each accelerated region (and candidate regions for the background set) and then compared GO-terms annotated to the background-selected genes with those of the ones selected by the gibARs using the hypergeometric distribution. =A TreeMap from the REVIGO[137] website (http://revigo.irb.hr) for GOslim Biological Process terms with a Benjamini-Hochberg FDR of 5% or less is shown in Extended Data Fig.6 and we observe categories like chromosome organization..

gibARs are enriched nearby LAVA-associated genes

We mapped both gibARs and LAVA elements to their nearest protein-coding Gibbon Ensembl genes (Nleu1.0). We then performed a hypergeometric test for enrichment of shared genes (using all protein coding Gibbon Ensembl genes as a background set), which is significant with a p-value of 8.1E-06 and an odds ratio of 2.74 (1.79–4.07, 95% CI). Therefore gibARs and LAVA elements preferentially co-occur nearby the same genes.

References

1 Gnerre, S., Lander, E. S., Lindblad-Toh, K. & Jaffe, D. B. Assisted assembly: how to improve a de novo genome assembly by using related species. *Genome Biol* **10**, R88, doi:10.1186/gb-2009-10-8-r88 (2009).

2 Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res* **13**, 103-107, doi:10.1101/gr.809403 (2003).

3 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 (2003).

4 Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87, doi:nature04072 (2005).

5 Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-1067, doi:btm071 [pii]

10.1093/bioinformatics/btm071 (2007).

6 Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195, doi:10.1371/journal.pcbi.1002195 (2011).

7 Church, D. M. *et al.* Modernizing reference genome assemblies. *PLoS Biol* **9**, e1001091, doi:10.1371/journal.pbio.1001091 (2011).

8 Wilming, L. G. *et al.* The vertebrate genome annotation (Vega) database. *Nucleic Acids Res* **36**, D753-760, doi:gkm987 (2008).

9 Carbone, L. *et al.* A high-resolution map of synteny disruptions in gibbon and human genomes. *PLoS Genet* **2**, e223, doi:06-PLGE-RA-0357R3 (2006).

10 Girirajan, S. *et al.* Sequencing human-gibbon breakpoints of synteny reveals mosaic new insertions at rearrangement sites. *Genome Res* **19**, 178-190, doi:10.1101/gr.086041.108 (2009).

11 Flicek, P. *et al.* Ensembl 2013. *Nucleic Acids Res* **41**, D48-55, doi:10.1093/nar/gks1236 (2013).

12 Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* **42**, D764-770, doi:10.1093/nar/gkt1168 (2014).

13 Severin, J. *et al.* eHive: an artificial intelligence workflow system for genomic analysis. *BMC Bioinformatics* **11**, 240, doi:10.1186/1471-2105-11-240 (2010).

14 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).

15 Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003-1007, doi:10.1126/science.1072047 (2002).

16 Marques-Bonet, T. *et al.* A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**, 877-881, doi:10.1038/nature07744 (2009).

17 Ventura, M., Archidiacono, N. & Rocchi, M. Centromere emergence in evolution. *Genome Res* **11**, 595-599, doi:10.1101/gr.152101 (2001).

18 Ventura, M. *et al.* Recurrent sites for new centromere seeding. *Genome Res* **14**, 1696-1703, doi:10.1101/gr.2608804 (2004).

19 Scally, A. *et al.* Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169-175, doi:10.1038/nature10842 (2012).

20 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57, doi:10.1038/nprot.2008.211 (2009).

21 Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641-646, doi:10.1126/science.1197005 (2010).

22 She, X., Cheng, Z., Zollner, S., Church, D. M. & Eichler, E. E. Mouse segmental duplication and copy number variation. *Nat Genet* **40**, 909-914, doi:10.1038/ng.172 (2008).

23 Nicholas, T. J. *et al.* The genomic architecture of segmental duplications and associated copy number variants in dogs. *Genome Res* **19**, 491-499, doi:10.1101/gr.084715.108 (2009).

24 Liu, G. E. *et al.* Analysis of recent segmental duplications in the bovine genome. *BMC Genomics* **10**, 571, doi:10.1186/1471-2164-10-571

1471-2164-10-571 [pii] (2009).

25 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-580, doi:gkc131 [pii] (1999).

26 Hach, F. *et al.* mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat Methods* **7**, 576-577, doi:10.1038/nmeth0810-576

nmeth0810-576 [pii] (2010).

27 Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41**, 1061-1067, doi:10.1038/ng.437

ng.437 [pii] (2009).

28 Dumas, L. *et al.* Gene copy number variation spanning 60 million years of human and primate evolution. *Genome research* **17**, 1266-1277, doi:10.1101/gr.6557307 (2007).

29 Nzounza, P. *et al.* The scaffolding protein Dlg1 is a negative regulator of cell-free virus infectivity but not of cell-to-cell HIV-1 transmission in T cells. *PLoS One* **7**, e30130, doi:10.1371/journal.pone.0030130 (2012).

30 Karsten, S. L. *et al.* A genomic screen for modifiers of tauopathy identifies puromycin-sensitive aminopeptidase as an inhibitor of tau-induced neurodegeneration. *Neuron* **51**, 549-560, doi:10.1016/j.neuron.2006.07.019 (2006).

31 Rannala, B. & Yang, Z. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**, 1645-1656 (2003).

32 Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nat Genet* **43**, 1031-1034, doi:10.1038/ng.937 (2011).

33 Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* **12**, 756-766, doi:10.1038/nrg3098 (2011).

34 Matsudaira, K. & Ishida, T. Phylogenetic relationships and divergence dates of the whole mitochondrial genome sequences among three gibbon genera. *Mol Phylogenet Evol* **55**, 454-459, doi:10.1016/j.ympev.2010.01.032 (2010).

35 Chatterjee, H. J., Ho, S. Y., Barnes, I. & Groves, C. Estimating the phylogeny and divergence times of primates using a supermatrix approach. *BMC Evol Biol* **9**, 259, doi:10.1186/1471-2148-9-259 (2009).

36 Hudson, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337-338 (2002).

37 Jukes, T. H., Cantor, C. R. Mammalian Protein Metabolism. 21–132 (1969).

38 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

39 Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423-3424, doi:10.1093/bioinformatics/btr539 (2011).

40 Neph, S. *et al.* BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919-1920, doi:10.1093/bioinformatics/bts277 (2012).

41 Benjamini, Y. Y., D., . The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 1165–1188 (2001).

42 Schmidt, D. *et al.* ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. *Methods* **48**, 240-248, doi:10.1016/j.ymeth.2009.03.001 (2009).

43 Xu, H. *et al.* A signal-noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics* **26**, 1199-1204, doi:10.1093/bioinformatics/btq128 (2010).

44 Schwalie, P. C. *et al.* Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes. *Genome biology* **14**, R148, doi:10.1186/gb-2013-14-12-r148 (2013).

45 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).

46 Smit, A., Hubley, R & Green, P. RepeatMasker Open-3.0. (1996-2010).

47 Morgulis, A., Gertz, E. M., Schaffer, A. A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J Comput Biol* **13**, 1028-1040, doi:10.1089/cmb.2006.13.1028 (2006).

48 Down, T. A. & Hubbard, T. J. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome research* **12**, 458-461, doi:10.1101/gr.216102 (2002).

49 Davuluri, R. V., Grosse, I. & Zhang, M. Q. Computational identification of promoters and first exons in the human genome. *Nat Genet* **29**, 412-417, doi:10.1038/ng780 (2001).

50 Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research* **25**, 955-964 (1997).

51 Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**, 78-94, doi:10.1006/jmbi.1997.0951 (1997).

52 Ongoing and future developments at the Universal Protein Resource. *Nucleic acids research* **39**, D214-219, doi:10.1093/nar/gkq1020 (2011).

53 Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic acids research* **38**, D5-16, doi:10.1093/nar/gkp967 (2010).

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:10.1016/S0022-2836(05)80360-2 (1990).

55 Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31, doi:10.1186/1471-2105-6-31 (2005).

56 Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome research* **14**, 988-995, doi:10.1101/gr.1865504 (2004).

57 Carbone, L. *et al.* Centromere remodeling in Hoolock leuconedys (Hylobatidae) by a new transposable element unique to the gibbons. *Genome Biol Evol* **4**, 648-658, doi:10.1093/gbe/evs048 (2012).

58 Bandelt, H. J., Forster, P. & Rohl, A. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16**, 37-48 (1999).

59 Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451-1452, doi:10.1093/bioinformatics/btp187 (2009).

60 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome research* **12**, 656-664, doi:10.1101/gr.229202. Article published online before March 2002 (2002).

61 Ray, D. A., Xing, J., Salem, A. H. & Batzer, M. A. SINEs of a nearly perfect character. *Syst Biol* **55**, 928-935 (2006).

62 Konkel, M. K. & Batzer, M. A. A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. *Semin Cancer Biol* **20**, 211-221, doi:10.1016/j.semcancer.2010.03.001 (2010).

63 Marchani, E. E., Xing, J., Witherspoon, D. J., Jorde, L. B. & Rogers, A. R. Estimating the age of retrotransposon subfamilies using maximum likelihood. *Genomics* **94**, 78-82, doi:10.1016/j.ygeno.2009.04.002 (2009).

64 Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297-304 (2000).

65 Boissinot, S., Chevret, P. & Furano, A. V. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Mol Biol Evol* **17**, 915-928 (2000).

66 Cordaux, R., Hedges, D. J., Herke, S. W. & Batzer, M. A. Estimating the retrotransposition rate of human Alu elements. *Gene* **373**, 134-137, doi:10.1016/j.gene.2006.01.019 (2006).

67 Smit, A., Hubley, R & Green, P. RepeatMasker Open-3.0. (1996-2010).

68 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26, doi:10.1038/nbt.1754 (2011).

69 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

70 van de Lagemaat, L. N., Medstrand, P. & Mager, D. L. Multiple effects govern endogenous retrovirus survival patterns in human gene introns. *Genome biology* **7**, R86, doi:10.1186/gb-2006-7-9-r86 (2006).

71 Zhang, Y., Romanish, M. T. & Mager, D. L. Distributions of transposable elements reveal hazardous zones in mammalian introns. *PLoS computational biology* **7**, e1002046, doi:10.1371/journal.pcbi.1002046 (2011).

72 Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. & Morris, Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome biology* **9 Suppl 1**, S4, doi:10.1186/gb-2008-9-s1-s4 (2008).

73 Kamburov, A., Stelzl, U., Lehrach, H. & Herwig, R. The ConsensusPathDB interaction database: 2013 update. *Nucleic acids research* **41**, D793-800, doi:10.1093/nar/gks1055 (2013).

74 Kamburov, A. *et al.* ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic acids research* **39**, D712-717, doi:10.1093/nar/gkq1156 (2011).

75 Kamburov, A., Wierling, C., Lehrach, H. & Herwig, R. ConsensusPathDB--a database for integrating human functional interaction networks. *Nucleic acids research* **37**, D623-628, doi:10.1093/nar/gkn698 (2009).

76 Lower, R. *et al.* Identification of human endogenous retroviruses with complex mRNA expression and particle formation. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 4480-4484 (1993).

77 Damert, A. *et al.* 5'-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. *Genome research* **19**, 1992-2008, doi:10.1101/gr.093435.109 (2009).

78 Hayashi, S., Hayasaka, K., Takenaka, O. & Horai, S. Molecular phylogeny of gibbons inferred from mitochondrial DNA sequences: preliminary report. *J Mol Evol* **41**, 359-365 (1995).

79 Takacs, Z., Morales, J. C., Geissmann, T. & Melnick, D. J. A complete species-level phylogeny of the Hylobatidae based on mitochondrial ND3-ND4 gene sequences. *Mol Phylogenet Evol* **36**, 456-467, doi:S1055-7903(05)00109-0 (2005).

80 Monda, K., Simmons, R. E., Kressirer, P., Su, B. & Woodruff, D. S. Mitochondrial DNA hypervariable region-1 sequence variation and phylogeny of the concolor gibbons, Nomascus. *Am J Primatol* **69**, 1285-1306, doi:10.1002/ajp.20439 (2007).

81 Whittaker, D. J., Morales, J. C. & Melnick, D. J. Resolution of the Hylobates phylogeny: congruence of mitochondrial D-loop sequences with molecular, behavioral, and morphological data sets. *Mol Phylogenet Evol* **45**, 620-628, doi:S1055-7903(07)00290-4 (2007).

82 Kim, S. K. *et al.* Patterns of genetic variation within and between Gibbon species. *Mol Biol Evol* **28**, 2211-2218, doi:10.1093/molbev/msr033 (2011).

83 Meyer, T. J. *et al.* An Alu-based phylogeny of gibbons (hylobatidae). *Mol Biol Evol* **29**, 3441-3450, doi:10.1093/molbev/mss149 (2012).

84 Mootnick, A. R. Species Identification Recommended for Rescue or Breeding Centers. *Primate Conservation*, 103-138 (2006).

85 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10-12 (2011).

86 Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome research* **21**, 936-939, doi:10.1101/gr.111120.110 (2011).

87 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

88 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498, doi:10.1038/ng.806 (2011).

89 Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* **30**, 2725-2729, doi:10.1093/molbev/mst197 (2013).

90 Felsenstein, J. PHYLIP (Phylogeny Inference Package).

91 Wall, J. D. *et al.* Incomplete lineage sorting is common in extant gibbon genera. *PLoS One* **8**, e53682, doi:10.1371/journal.pone.0053682 (2013).

92 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591, doi:msm088 (2007).

93 Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci* **17**, 57–86 (1986).

94 Lischer, H. E., Excoffier, L. & Heckel, G. Ignoring heterozygous sites biases phylogenomic estimates of divergence times: implications for the evolutionary history of Microtus voles. *Mol Biol Evol*, doi:mst271 (2013).

95 Yoshioka, T. *et al.* A new combination chemotherapy with cis-diammine-glycolatoplatinum (Nedaplatin) and 5-fluorouracil for advanced esophageal cancers. *Intern Med* **38**, 844-848 (1999).

96 Kubatko, L. S. & Degnan, J. H. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst Biol* **56**, 17-24, doi:771129708 (2007).

97 Yang, Z. Maximum-Likelihood Models for Combined Analyses of Multiple Sequence Data. *J Mol Evol* **42**, 587-596 (1996).

98 Degnan, J. H. & Rosenberg, N. A. Discordance of species trees with their most likely gene trees. *PLoS Genet* **2**, e68, doi:10.1371/journal.pgen.0020068 (2006).

99 Orlando, L. *et al.* Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**, 74-78, doi:10.1038/nature12323 (2013).

100 Brudno, N. D. a. M. Hapsembler : An Assembler for Highly Polymorphic Genomes. *Research in Computational Molecular Biology*, 38–52 (2011).

101 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780, doi:10.1093/molbev/mst010 (2013).

102 Chan, Y. C. *et al.* Mitochondrial genome sequences effectively reveal the phylogeny of Hylobates gibbons. *PLoS One* **5**, e14419, doi:10.1371/journal.pone.0014419 (2010).

103 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792-1797, doi:10.1093/nar/gkh340 (2004).

104 Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540-552 (2000).

105 Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* **9**, 772, doi:10.1038/nmeth.2109 (2012).

106 Zwickl, D. J. *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence data sets under the maximum likelihood criterion*, (2006).

107 Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310-2314, doi:10.1126/science.1065889 (2001).

108 Swofford, D. L. PAUP*: phylogenetic analysis using parsimony (* and other methods), version 4. *Sunderland: Sinauer Associates* (2003).

109 Gelman A, R. D. Inference from iterative simulation using multiple sequences. *Stat Sci* **7**, 457–511 (1992).

110 Nylander, J. A., Wilgenbusch, J. C., Warren, D. L. & Swofford, D. L. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics* **24**, 581-583, doi:btm388 [pii]

10.1093/bioinformatics/btm388 (2008).

111 Drummond, A. J., Ho, S. Y., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol* **4**, e88, doi:05-PLBI-RA-0392R4 (2006).

112 Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7**, 214, doi:10.1186/1471-2148-7-214 (2007).

113 Vignaud, P. *et al.* Geology and palaeontology of the Upper Miocene Toros-Menalla hominid locality, Chad. *Nature* **418**, 152-155, doi:10.1038/nature00880 (2002).

114 Brunet, M. *et al.* New material of the earliest hominid from the Upper Miocene of Chad. *Nature* **434**, 752-755, doi:nature03392 (2005).

115 Lebatard, A. E. *et al.* Cosmogenic nuclide dating of Sahelanthropus tchadensis and Australopithecus bahrelghazali: Mio-Pliocene hominids from Chad. *Proc Natl Acad Sci U S A* **105**, 3226-3231, doi:10.1073/pnas.0708015105 (2008).

116 Kelley, J. (ed The primate fossil record) 369–384 (Cambridge University Press, 2002).

117 Zalmout, I. S. *et al.* New Oligocene primate from Saudi Arabia and the divergence of apes and Old World monkeys. *Nature* **466**, 360-364, doi:10.1038/nature09094 (2010).

118 Pozzi, L., Hodgson, J. A., Burrell, A. S. & Disotell, T. R. The stem catarrhine Saadanius does not inform the timing of the origin of crown catarrhines. *J Hum Evol* **61**, 209-210, doi:10.1016/j.jhevol.2011.02.008 (2011).

119 Fabre, P. H., Rodrigues, A. & Douzery, E. J. Patterns of macroevolution among Primates inferred from a supermatrix of mitochondrial and nuclear DNA. *Mol Phylogenet Evol* **53**, 808-825, doi:10.1016/j.ympev.2009.08.004 (2009).

120 Springer, M. S. *et al.* Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. *PLoS One* **7**, e49521, doi:10.1371/journal.pone.0049521 (2012).

121 Perelman, P. *et al.* A molecular phylogeny of living primates. *PLoS Genet* **7**, e1001342, doi:10.1371/journal.pgen.1001342 (2011).

122 Raaum, R. L., Sterner, K. N., Noviello, C. M., Stewart, C. B. & Disotell, T. R. Catarrhine primate divergence dates estimated from complete mitochondrial genomes: concordance with fossil and nuclear DNA evidence. *J Hum Evol* **48**, 237-257, doi:S0047-2484(04)00166-6 (2005).

123 Goodman, M. *et al.* Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. *Mol Phylogenet Evol* **9**, 585-598, doi:S1055-7903(98)90495-X [pii] 10.1006/mpev.1998.0495 (1998).

124 Israfil, H., Zehr, S. M., Mootnick, A. R., Ruvolo, M. & Steiper, M. E. Unresolved molecular phylogenies of gibbons and siamangs (Family: Hylobatidae) based on mitochondrial, Y-linked, and X-linked loci indicate a rapid Miocene radiation or sudden vicariance event. *Mol Phylogenet Evol* **58**, 447-455, doi:10.1016/j.ympev.2010.11.005 (2011).

125 Thinh, V. N. *et al.* Mitochondrial evidence for multiple radiations in the evolutionary history of small apes. *BMC Evol Biol* **10**, 74, doi:10.1186/1471-2148-10-74 (2010).

126 Wallace, I. M., O'Sullivan, O., Higgins, D. G. & Notredame, C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res* **34**, 1692-1699, doi:34/6/1692 (2006).

127 Howe, K., Bateman, A. & Durbin, R. QuickTree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics* **18**, 1546-1547 (2002).

128 Han, M. V., Demuth, J. P., McGrath, C. L., Casola, C. & Hahn, M. W. Adaptive evolution of young gene duplicates in mammals. *Genome research* **19**, 859-867, doi:10.1101/gr.085951.108 (2009).

129 Han, M. V., Thomas, G. W., Lugo-Martinez, J. & Hahn, M. W. Estimating Gene Gain and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using CAFE 3. *Mol Biol Evol* **30**, 1987-1997, doi:10.1093/molbev/mst100 (2013).

130 Hahn, M. W., Demuth, J. P. & Han, S. G. Accelerated rate of gene gain and loss in primates. *Genetics* **177**, 1941-1949, doi:10.1534/genetics.107.080077 (2007).

131 Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476-482, doi:10.1038/nature10530 (2011).

132 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:10.1101/gr.097857.109 (2010).

133 Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief Bioinform* **12**, 41-51, doi:10.1093/bib/bbq072 (2011).

134 Self, S. G. L., K.-Y. Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions. *Journal of the American Statistical Association* **82**, 605–610 (1987).

135 Duret, L. & Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* **10**, 285-311, doi:10.1146/annurev-genom-082908-150001 (2009).

136 Kostka, D., Hubisz, M. J., Siepel, A. & Pollard, K. S. The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. *Mol Biol Evol* **29**, 1047-1057, doi:10.1093/molbev/msr279 (2012).

137 Supek, F., Bosnjak, M., Skunca, N. & Smuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**, e21800, doi:10.1371/journal.pone.0021800 (2011).

138 L., B. The genome as a life-history character: why rate of molecular evolution varies between mammal species. Philos Trans R Soc Lond B Biol Sci. 366, 2503-2513 (2011).

139 Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet 13, 745-753, doi:10.1038/nrg3295 (2012).

140 Elango, N., Thomas, J. W. & Yi, S. V. Variable molecular clocks in hominoids. Proc Natl Acad Sci U S A 103, 1370-1375, doi:10.1073/pnas.0510716103 (2006).

141 Kong, A. et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature 488, 471-475, doi:10.1038/nature11396 (2012).

142 Stevens, N. J. et al. Palaeontological evidence for an Oligocene divergence between Old World monkeys and apes. Nature 497, 611-614, doi:10.1038/nature12161 (2013).

143 1Green, R. E. et al. A draft sequence of the Neandertal genome. Science 328, 710-722, doi:10.1126/science.1188021 328/5979/710 [pii] (2010).

144 Veeramah, K. R. & Hammer, M. F. The impact of whole-genome sequencing on the reconstruction of human population history. Nat Rev Genet 15, 149-162, doi:10.1038/nrg3625 (2014).

145 Takahata, N. & Satta, Y. Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. Proc Natl Acad Sci U S A 94, 4811-4815 (1997).

146 Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. Genetics 156, 297-304 (2000).

147 Harvey P. H., M. R. D., Clutton-Brock T. H., in Primate Societies, (University of Chicago Press., 1987).