doi:10.1038/nature11128

Supplementary Information
Table of Contents

Supplementary Information 1

Sequencing the Bonobo Genome

Anne Fischer[1,2], Susan Ptak[1], Kay Prüfer[1], Chinnappa Kodira[3], Janet Kelso[1] and Svante Pääbo[1,*]

1. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
2. International Center for Insect Physiology and Ecology, Nairobi, Kenya
3. 454 Life Sciences, Branford, CT, USA

* To whom correspondence should be addressed (paabo@eva.mpg.de)

We sequence the Bonobo genome of a female individual to a total depth of 26-fold coverage. Sequencing data consists of reads from the 454 GS FLX and GS FLX Titanium platform. A total of three-fold coverage of all sequence data is present in the form of paired end reads of insert sizes of 3 kilo bases, 9 kilo bases and 20 kilo bases.

Sequenced Individual and Library Preparation

We sequenced the genome of a female Bonobo (Ulindi, Leipzig Zoo, Studbook#183), born in captivity in Frankfurt (Germany) on the 10th of October 1993. A blood sample was drawn on the 18th of May 2001 during routine examination by the veterinarian of the Leipzig zoo. DNA was extracted from a cell culture obtained from lymphocytes transformed with Epstein-Barr virus using the Gentra-puregene kit from QIAGEN. The DNA was used for 454 shotgun libraries following manufacturer's instructions.

The 454 paired end libraries were prepared from the extracted DNA as follows: 1) High molecular weight genomic DNA was sheared to the desired size of 20kb, 9kb or 3kb span distance fragments using Hydroshear apparatus (Genomic Solutions, Ann Arbor, MI). 2) Fragment ends were made blunt by treating them with T4 DNA polymerase and T4 polynucleotide kinase (Roche Applied Science, Indianapolis, IN). 3) Circularization adaptors containing a loxP target sequencing DNA adaptors were blunt-end ligated onto fragment ends. 4) *Cre* recombinase was employed to create intra-molecular recombination circularizing the fragments. 5) The circularized material was nebulized to a size of 500bp and blunt-ended via T4 DNA polymerase and T4 polynucleotide kinase treatment. 6) Library fragments containing flanking linker sequences were selected via the biotinylated linker, using Dynal M-270 magnetic streptavidin beads (Invitrogen, Carlsbad, CA). 7) Double-stranded DNA adaptors were ligated to blunt fragment ends and the resulting libraries were amplified using 20 cycles of PCR and purified with AMPure beads (Agencourt Bioscience, Beverly, MA, USA). 8) Single stranded DNA libraries isolated

via Dynal M-270 bead binding followed by alkaline treatment were then quantified using Quant-iT RiboGreen (Invitrogen, Carlsbad, CA) prior to emulsion PCR amplification.

Shotgun Sequencing Data

All shotgun sequencing was carried out using the 454 GS FLX (hereon abbreviated to FLX) and GS FLX Titanium (Titanium) sequencing platforms. Reads were base called and filtered by the standard 454 processing software. A total of 58.1 million reads on 118 runs were acquired by the FLX and a total of 143.1 million reads on 143 runs by the Titanium platform. Runs had different plate-layout with 2 to 16 lanes per plate. FLX lanes gave an average read length between 150 to 250 base pairs and Titanium lanes varied from 150 to 410 base pairs. In total 63 billion bases were acquired. Table S1.1 summarizes the statistics over FLX and Titanium lanes and Figure S1.1 shows the distribution of average read length over sequenced lanes.

Quantity and Estimated Insert Sizes of Paired End Sequences

In addition to the shotgun sequencing libraries, we produced paired end libraries with estimated insert sizes of 3 kilo bases, 9 kilo bases and 20 kilo bases. A total of 68.4 million reads were produced from the paired end libraries (see Table S1.1 and Figure S1.1).

Typically, runs from 454 paired end libraries give a mixture of shotgun sequences and paired end sequences. Due to the paired end library preparation, true paired end sequences should carry a specific linker sequence marking the location of the insert. These linker-positive sequences, in turn, may originate in part from amplification duplicates and do not represent independent observations. Thus, the ratio of linker-positive to shotgun sequences and the number of unique molecules are important parameters to judge the quality of the paired end runs.

We tested for the presence of linker sequence and the uniqueness among linker positive-sequences in our paired end sequencing data. Uniqueness was determined by comparing the first 50 base pairs between linker-positive reads. Tables S1.2, S1.3 and S1.4 show the results for the three different insert size libraries. Generally, we observed around 70%, 60-70% and 50-60% linker positive sequences for 3, 9 and 20 kilo base paired end libraries, respectively. When subtracting redundant reads, the true paired end reads sum up to 12.7 million, 12.5 million and 6.5 million unique reads for the insert sizes 3, 9 and 20 kilo bases. The left and right sequence in unique paired end reads sums to a total of 10.4 giga bases. The clone coverage over all insert sizes is over 80-fold, assuming a genome size of 3.4 giga bases.

We additionally evaluated the insert sizes of the 9 kilo bases and 20 kilo bases paired-end libraries by aligning all unique linker-positive reads to the human genome (hg18). When calculating the

insert size, we only included reads for which both ends aligned uniquely and in correct orientation to each other. Estimated insert sizes varied from 8952 to 9825 base pairs for 9 kilo base libraries and from 18628 to 20414 base pairs for 20 kilo base libraries, consistent with expectation (see Tables S1.3 and S1.4). For assembly (SI 2), all paired-end sequences were used, regardless of their mapping to the human genome.

	Total Reads	Average Read Length per Lane	Total Bases x10^6
FLX Shotgun	58113888	151.6-246.6	13027
Titanium Shotgun	143137950	154.8-413.7	50511
3 kilo bases Paired End	27602635	281.3-370.9	8588
9 kilo bases Paired End	23903005	243.5-367.7	8097
20 kilo bases Paired End	16906997	284.9-373.2	5844
Total	269664475	-	86067

Table S1.1: Summary of 454 sequencing data. Numbers reflect the bases inside the clear ranges, as determined by the instrument's signal processing software and reported by the instruments' sffinfo software. Total bases include *N* bases and linker sequences.

Lib ID	Lanes Prefix	Reads	Linker+	Average Left Overhang	Average Right Overhang	Uniqueness	Unique PE Reads
Lib1	FPBIVZG, FPBVMLO	2,316,926	40%	166	177	62.36%	573,723
Lib2	FPCZ3PW, FQA8B1O, FQA82GB	3,309,791	43%	154	173	63.98%	903,923
9721	FQCV8NE, FQES10L	2,268,316	72%	139	143	71.97%	1,181,360
9722	FQCUZFX, FQF8HJA	2,318,153	75%	144	148	71.41%	1,233,626
9723	FQPUUWC, FQTGE7X	2,092,225	73%	147	154	65.70%	1,002,318
9724	FQPTE71, FQCVGOB	2,118,680	74%	152	157	60.67%	952,719
9725	FQGIZYV, FQ4ALJH	2,286,344	73%	139	144	71.47%	1,190,308
9726	FQH0JDV, FQH255V	2,049,612	72%	143	149	69.36%	1,026,153
9727	FQEN2RR, FQH0QQE	2,350,753	71%	143	148	73.33%	1,218,572
9728	FQIGH9C, FQRQLHM	2,083,852	68%	142	148	74.11%	1,055,034
9729	FQF8DOX, FQTIZJO	2,307,454	71%	145	149	74.08%	1,218,351
9730	FQF573J, FQRNL2W	2,100,323	73%	140	145	73.90%	1,128,672

Table S1.2: Summary of 454 3 kilo base paired end sequencing data. Each library was sequenced on two or three full Titanium runs. Linker+ gives the percentage of linker positive reads. No pUC vector sequences were found in the runs.

Lib ID	Lanes Prefix	Reads	Linker+	Average Left Overhang	Average Right Overhang	Uniqueness	Unique PE Reads	Insert Size
10626	FWXMPCL	1,045,929	67%	169	177	88.62%	619,778	8981
10627	FWT5D9A	630,373	45%	142	155	89.66%	257,144	8952
10628	FV38HK1	1,090,411	68%	176	184	86.37%	644,507	8973
10629	FV39B9G	995,349	62%	159	173	86.99%	540,429	9064
10630	FWT7XZW	1,101,196	67%	165	175	84.90%	630,358	9012
10631	FWXVGFT	807,450	62%	166	176	83.59%	419,638	9075
10633	FV9R4VY	1,146,761	68%	179	186	83.86%	653,294	9508
10635	FWT746V	722,813	65%	179	194	82.53%	384,799	9482
10636	FWKZNA3	584,171	55%	157	170	77.70%	249,016	9554
	FW6TEYC	829,470	59%	168	180	85.33%	414,506	9614
10637	FUVTY9H	1,118,557	63%	163	177	83.42%	587,935	9788
10638	FW6SSZY	1,017,705	67%	178	185	80.28%	549,658	9619

10639	FWZQ6VN	1,100,588	66%	175	183	82.01%	595,905	9595
10640	FWV5POX	933,413	64%	172	181	83.53%	502,123	9625
10641	FWZR29M	1,158,476	67%	174	184	78.04%	602,707	9669
10642	FW63VFL	1,112,222	66%	172	182	81.13%	594,594	9648
10692	FUVTY9H	1,067,757	65%	165	176	84.71%	588,271	9732
10693	FWVZKMY	1,043,721	67%	175	182	82.25%	577,495	9815
10694	FU1D09B	876,439	59%	153	166	83.79%	435,064	9825
10695	FVCI2AB	1,100,564	64%	167	181	85.89%	608,664	9785
10696	FV54Z6E	1,183,563	67%	175	182	74.63%	593,314	9747
10697	FV39CVK	1,086,639	59%	157	170	77.26%	493,303	9771
10698	FV39K9V	1,120,835	64%	161	172	82.77%	598,109	9717
10699	FWVZ31K	768,718	54%	156	167	84.28%	349,606	9744

Table S1.3: Summary of 454 9 kilo base paired end sequencing data. Each library was sequenced on one full Titanium run, except for Library 10636 which was sequenced on two runs. Linker+ gives the percentage of linker positive reads. No pUC vector sequences were found in the runs.

Lib. ID	Lane ID	Reads	Linker+	pUC	Average Left Overhang	Average Right Overhang	Uniqueness	Unique PE Reads	Insert Size
10197	FS03XYH01	467,569	66%	0%	181	192	53.07%	163,171	19959
10201	FS03XYH02	468,574	62%	0%	181	192	54.09%	157,636	18878
10202	FS07JTZ01	520,068	53%	0%	163	178	58.55%	160,973	18910
10198	FS07JTZ02	499,068	54%	0%	160	174	61.58%	165,512	20074
10195	FS206NH01	586,553	57%	0%	165	177	60.85%	202,924	20234
10200	FS206NH02	479,573	59%	0%	165	178	61.08%	172,164	20053
10196b	FS6Q6AR01	600,885	64%	0%	180	190	60.40%	232,857	20051
10204	FS6Q6AR02	452,881	59%	0%	172	183	63.18%	168,699	18863
10042	FSC5L8E01	573,876	62%	0%	177	187	61.39%	217,078	20123
10043	FSC5L8E02	529,076	63%	0%	180	189	63.07%	209,928	20279
10050	FSIMIN301	622,766	62%	2%	172	183	61.11%	236,488	18904
10051	FSIMIN302	481,465	61%	2%	173	185	62.13%	183,548	18901
10048	FSIN7GX01	496,559	65%	0%	184	195	64.63%	208,349	20182
10049	FSIN7GX02	541,329	63%	0%	186	198	65.45%	222,104	20226
10046	FSIQWOH01	563,137	64%	0%	180	189	67.20%	240,743	20227
10047	FSIQWOH02	499,666	63%	0%	178	188	71.17%	225,032	20179
10123	FSN5HHM01	560,123	52%	19%	170	185	54.43%	159,258	20110
10125	FSN5HHM02	522,488	63%	0%	175	189	62.67%	205,752	19934
10044	FSNVG7S01	503,662	57%	0%	174	189	68.18%	195,911	20209
10045	FSNVG7S02	484,059	59%	0%	177	190	62.42%	177,405	20414
10129	FSP2QZJ01	465,132	51%	0%	149	167	64.74%	152,313	20058
10122	FSP2QZJ02	485,340	48%	17%	162	178	60.33%	139,646	20065
10054	FSTN3BH01	595,544	61%	1%	171	182	68.30%	249,637	19159
10126	FSTN3BH02	583,553	65%	0%	173	186	52.52%	199,653	20119
10055b	FSTQL1V01	546,447	56%	1%	173	172	77.70%	236,214	19162
10124b	FSTQL1V02	576,807	59%	0%	176	180	64.64%	220,495	19999
10199b	FTD51JF01	587,345	62%	0%	173	182	62.87%	227,158	20116
10203	FTD51JF02	583,881	60%	0%	173	185	55.44%	194,913	18628
10052b	FXJ4M8I01	538378	64%	2%	181	188	78.82%	270,794	19000
10053b	FXJ4M8I02	478197	64%	3%	171	178	80.04%	243,654	18985
10127c	FXLXRWR01	488744	65%	0%	182	191	73.29%	233,424	20004
10128c	FXLXRWR02	524165	65%	0%	181	191	70.04%	238,745	20047

Table S1.4: Summary of 454 20 kilo base paired end sequencing data. Each library was sequenced on one half-plate Titanium lane. Linker+ gives the percentage of linker positive reads and pUC gives the percentage of vector sequence among all reads.

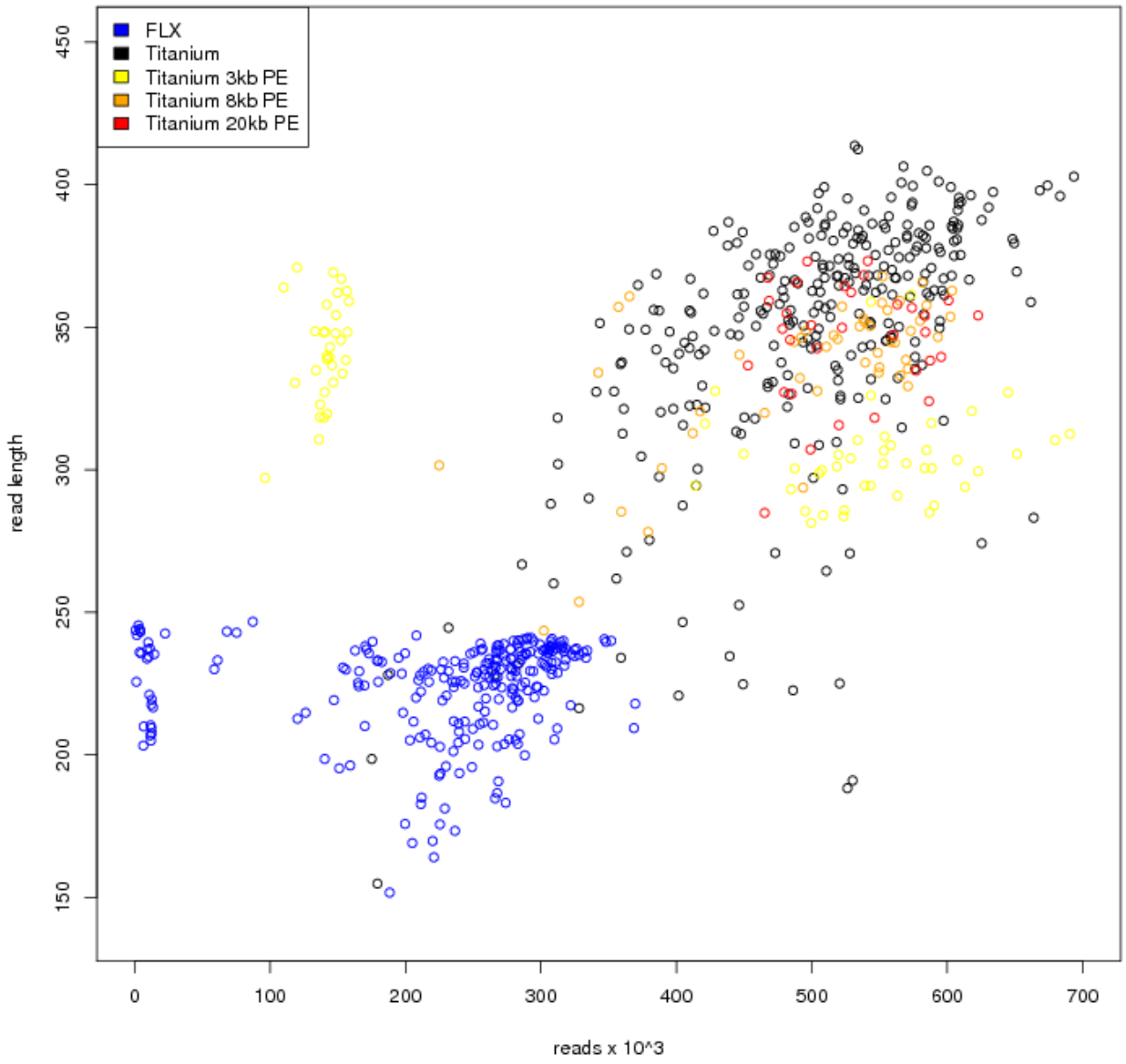


Figure S1.1: Scatterplot of number reads and average read length for all sff's. Differences in the number of reads may be caused by plate-layout, sequencing platform and sequencing yield. Read length differences are mainly due to difference in sequencing platform (FLX vs. Titanium).

Contaminant Screening

In order to detect sample mix-up, we sampled reads from each sequence format file (sff) by extracting every 299th read in the sff up to the maximum of 1000 samples per file. Sampled reads were aligned to the non-redundant GenBank database (nt, downloaded Jan 2007), the chimpanzee genome (pantro2), and the human genome (hg18) using blat [1] (Version 34; option –fastMap). The hits to the three different target databases were compared and the best hit was determined based on the bitscore. Over 70% of the sampled reads aligned best to the chimpanzee genome. Hits to other GenBank entries are rare and include the Epstein-Barr virus used in the transformation of the cell-lines and, in the case of the 20 kilo base insert libraries, other vector sequences used during the paired end library generation. However, we identified one Titanium sff (FLGU9LC01) with only 50% of the reads having the best alignment to the chimpanzee genome and a further 20% aligning to GenBank sequences of the plant-order *poales*.

We further investigated the contaminated sff by aligning more sensitively to the *Zea mays* [2] and chimpanzee genome [3] using megablast [4] (Version: 2.2.14; options: –W 16 –F F –U F –e 0.001). Over 99% of all reads have a hit to one of the two databases, with 50.8% aligning best to *Zea mays* and 48.8% aligning best to the chimpanzee sequence. *Zea mays* reads from this sff segregated well into distinct contigs in our assembly (see SI 2 for details) and these contigs were excluded from the AGP file. The sff was omitted from all analysis based on mapping of bonobo genome reads.

Sequence Read Archive Accessions and Data Availability

All sequences have been made available through the Sequence Read Archive under the study accession ERP000601 (http://www.ebi.ac.uk/ena/data/view/ERP000601). The contaminated sff has been deposited as part of this study under experiment id ERX012382. GS FLX shotgun, Titanium shotgun, 3kb paired end, 9kb paired end, 20kb paired end were deposited under experiment ids ERX012380, ERX012381, ERX012383, ERX012384, ERX012385, respectively.

Supplementary Information 2

The Bonobo Shotgun Assembly

Jason Miller[1,*], Brian Walenz[1], Sergey Koren[1,2], Granger Sutton[1], James Knight[3], Chinnappa Kodira[3], James Mullikin[4], Kay Prüfer[5]

1. J. Craig Venter Institute, Rockville MD USA
2. University of Maryland, College Park, MD USA
3. 454 Life Sciences, Branford CT USA
4. National Human Genome Research Institute, NIH, Bethesda MD USA
5. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

* To whom correspondence should be addressed (jmiller@jcvi.org)

Abstract

The Bonobo Genome Consortium generated DNA sequencing reads representing the genome of a single bonobo individual. The data consisted of almost 270 million fragment sequences generated on FLX machines from 454 Life Sciences. The fragments derived from FLX standard and Titanium chemistries, and from paired and unpaired protocols. The data was assembled at the J. Craig Venter Institute with the open-source Celera Assembler software including the CABOG variant designed for pyrosequencing data. The assembly process combined reads and paired end constraints into contigs and scaffolds. Considering all reads that survived minimum length and quality filters, the assembly incorporated 88% of reads and satisfied 86% of the usable mate pair constraints while violating only 0.12%. The assembled scaffolds had a combined length that approaches the expected 3 Gbp genome size.

General Assembly Parameters

Two assemblies were generated with the Celera Assembler software, also known as CABOG [5-9]. The specific software version, 5.4.3, is available from the Source Forge web site (http://wgs-assembler.sourceforge.net) as a packaged release. It is also tagged VERSION-5_43-RELEASE in the cvs source code repository on Source Forge. Celera Assembler was run with the algorithmic parameter settings given in Table S2.1. The expected sequencing error rate (utgErrorRate, default=0.015) was adjusted upwards based on preliminary analysis of other Titanium reads, not shown. The limit on iterations of scaffold operations (doExtendClearRange, default=2) was chosen to reduce run time.

The assembly process ran on a compute grid running Linux and SGE at the J. Craig Venter Institute. The total assembly pipeline used about 2.5 TB of disk. The parallel computes ran on grid nodes with 2 to 4 core and 8 to 16 GB RAM. The non-parallel computes ran on a single 16-core node with 96GB shared RAM. The performance-related parameter settings are given in Table S2.2.

Stage	Parameter	Value	Explanation
OBT			Calculate partial overlaps for trimming
	obtOverlapper	ovl	Use the Sanger-era overlap algorithm
	obtMerSize	22	Seed on uncompressed K-mers with **K**=22
	obtMerThreshold	721	Seed overlaps on K-mers with **T**<= 721
MER			Calculate dovetail overlaps for unitigs
	ovlOverlapper	mer	Use the 454-era overlap algorithm
	ovlMerSize	29	Seed on compressed K-mers with **K**=29
	ovlMerThreshold	300	Seed on K-mers with **T**<=300
	ovlErrorRate	0.06	Retain overlaps up to 6% before correction
BOG			Calculate untigs with the Best Overlap Graph
	utgErrorRate	0.03	Maximum 3% error in corrected overlaps
CNS			Calculate consensus on unitigs and scaffolds
	cnsErrorRate	0.06	Maximum 6% error between read & consensus
CGW			Calculate contigs and scaffolds
	cgwErrorRate	0.10	Maximum 10% error for contig merges
	doExtendClear-Ranges	1	Two rounds of CGW and one round of ECR

Table S2.1. Algorithm parameters used with Celera Assembler. The parameter/value pairs were passed to Celera Assembler with the 'runCA –s <spec file>' pipeline executive. Default values were used for all parameters not shown. The quantity **T** is the number of observations of a single K-mer sequence across all the reads.

Stage	Parameter	Value	Explanation
ALL			Grid usage configuration
	useGrid	1	Use grid for parallel computations
	scriptOnGrid	0	Non-grid for other computes
	sgeOverlap	-pe threaded 2 -l memory=4g	Two threads, 8GB total
	sgeMerOverlapSeed	-pe threaded 4 -l memory=4g	Four threads, 16GB total
	sgeMerOverlapExtend	-pe threaded 2 -l memory=6g	Two threads, 12GB total
	sgeFragmentCorrection	-pe threaded 2 -l memory=4g	Two threads, 8GB total
	sgeOverlapCorrection	-pe threaded 1 -l memory=8g	One thread, 8GB total
MER			Calculate dovetail overlaps
	merOverlapperThreads	4	
	merOverlapper-SeedBatchSize	1500000	
	merOverlapper-ExtendBatchSize	1000000	

Table S2.2. Performance parameter settings used with Celera Assembler. The "-pe" and "-l" values were automatically passed to the SGE grid controller as parameters to the "qsub" job submission command. These directives are outside the SGE standard but SGE was configured to recognize them. The two values served to reserve a minimum number of CPU core and a minimum amount of RAM per thread, respectively, on the grid.

Sequence Data Processing and Overlap-Based Trimming

The sequencing data had been generated at 454 Life Sciences on 454 GS FLX instruments. Sequencing had used both the FLX standard chemistry and the FLX Titanium chemistry with XLR protocols. The Titanium library construction had used unpaired protocols plus protocols for 3Kbp, 8Kbp, and 20Kbp paired ends. See Table 1.1 for details on the input data. SFF files were pre-processed with Celera Assembler's sffToCA program. Each read's clear range was taken from the SFF file to exploit quality value (QV) trimming performed by the 454 base calling software. FLX standard fragments were filtered if they contained the ambiguous base call N inside or outside the clear range. FLX Titanium reads were not filtered based on N-content. Technical replicate fragments, a problem described elsewhere [10], were filtered if they formed a perfect prefix of any other fragment from the same library. Fragments from the paired end libraries were examined for presence of the 44 bp FLX standard linker sequence or the 42bp Titanium linker sequence, as provided by 454 Life Sciences. The software trimmed fragments with partial or multiple linker sequences and fragments with linker at one fragment end. It converted remaining linker-positive fragments to Sanger-style mate pairs of reads. It filtered the remaining reads to enforce a 64 bp read length minimum. The paired end processing is explained in Figure S2.1

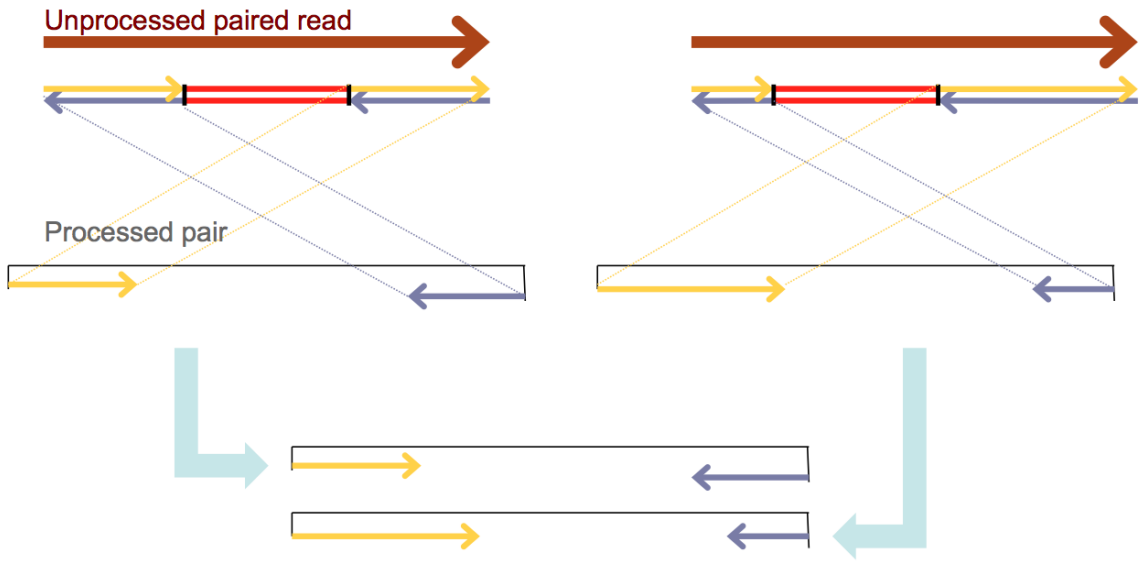


Figure S2.1. Schematic view of paired end processing. Top: The fragment sequences (brown) present two observations of the same double-stranded linker-insertion event (left and right). Software detects and removes the linker sequence (red) and generates Sanger-style mate pairs by swapping the fragment prefix and suffix (yellow) after reverse-complementing the fragment prefix (blue). The sffToCA software implemented this algorithm. Bottom: The two fragments contribute mate pairs whose linker-ligation points align perfectly. Software detects the redundancy from partial overlaps and removes all but one redundant mate pair. The OBT module implemented this algorithm.

Celera Assembler's OVL overlap module calculated partial overlaps of read pairs. Partial overlaps were required to span at least 40bp and have error<=6%. In contrast to the dovetail overlaps described elsewhere, partial overlaps were not required to span any read ends. Overlaps were calculated on all read pairs sharing at least one K-mer where **K**=22 and **T**<=721; **T** is the

K-mer frequency observed in all untrimmed reads. The **T** threshold was chosen to capture 99.99% of all distinct K-mers with **T**>1. This value of **T** captured 87.26% of the sequence data. The parallel overlap computation required 8,681 CPU hours and wrote 70.34 billion overlaps in 623 GB (after compression) of disk space.

Celera Assembler's OBT (Overlap-Based Trimming) program filtered and trimmed the reads based on the partial overlaps. OBT details include:

1. OBT used pre-computed partial overlaps to recognize near-perfect replicate reads from the same library and to remove all but the longest such replicate. This was designed as a more rigorous filter than the perfect-prefix filter applied earlier.
2. OBT detected mate pairs whose pair-wise alignments indicated agreement to within 1bp at both linker removal sites. This filter was designed to remove replicate mate pairs, that is, multiple observations of the same linker insertion event. See Figure S2.1. The filter was restricted to mate pairs from the same library. It removed 3.65 million pairs, or 7.3 million reads.
3. OBT trimmed reads whose sequence could not be confirmed by other reads. Of the 70.34 billion partial overlaps, this filter used the 38.41 billion overlaps that had length>=75bp and error<=2%. It examined each candidate read in the presence of its overlaps. This process trimmed 262,888,933 reads. It deleted 22.2 million reads whose resulting clear range had length<64bp. The deleted data included 19.4 million unpaired reads and 2.8 million paired reads. Both reads were deleted from 69 thousand pairs.
4. Using the same framework, OBT trimmed reads deemed to be spurs, i.e. those reads with error-prone sequence at one end. It tested whether the set of overlaps to a read defined one point on the read where many alignments ended. It trimmed sequence after such a point. The 2,976,650 reads flagged as spur were trimmed and about 699,000 were filtered due to remaining length<64bp.
5. Using the same framework again, OBT trimmed reads deemed to be chimera. It tested whether the set of overlaps to a read defined multiple distinct spans of confirmed sequence. In this case, it retained only the largest confirmed span. The 3,130,444 reads flagged as chimera were trimmed and about 173,000 were filtered due to remaining length<64bp.

The fragment processing is summarized in Table S2.4.

Process	Detail	ΔReads	Remaining Reads
sffToCA			
	Input from SFF files		269,676,092
	Detect linker, split fragments into mate pairs	+25,965,786	
	Filter imperfect linker, perfect prefix, and length<64bp	-13,567,894	
	Output for assembly		282,073,984
obt			
	Remove near-perfect replicate reads	-1,889,332	
	Remove both reads of replicate mates	-7,302,194	

	Trim low-quality bases	-22,206,193	
	Trim chimera	-172,666	
	Trim spurs	-699,025	
	Double-counted reads	+297,372	
QC	Usable reads		250,101,946

Table S2.4. Results of initial read processing. The initial increase in read count was due to the splitting of linker-positive fragments into mate pairs of reads. The overall reduction in read count was due to quality filtering. The table includes a correction for double counting because version 5 of Celera Assembler reported all applicable filters per deleted read.

After processing, the read population was characterized by the histograms in Figure S2.2. All reads have length between 64 and 2047 bp, reflecting the minimum and maximum enforced by Celera Assembler filters. Reads from unpaired libraries have distinct peaks with most Titanium reads longer than FLX Standard reads. The reads from paired end libraries have many short reads, an expected result of the linker-removal process. The paired-end libraries have smaller peaks near the Titanium unpaired peak as expected from the number of linker-negative fragments.

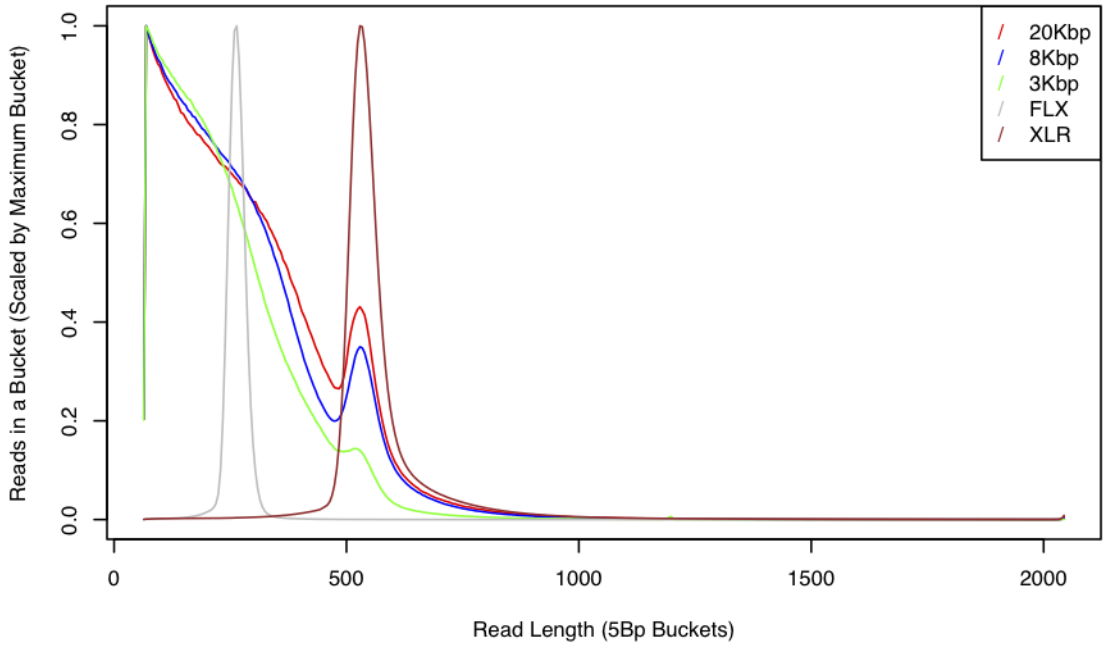


Figure S2.2. Read length histogram by library type. The figure reflects reads in Celera Assembler's internal database after trimming, filtering, and processing to generate mate pairs. FLX indicates unpaired FLX standard; XLR indicates unpaired Titanium. The others indicate paired end Titanium sequencing with 3 insert sizes. Bucket size is 5 bp. Curves were normalized by their maximum count in any bucket.

Overlap, Layout, Consensus Steps

Celera Assembler's 454-specific MER overlap module was used to generate a list of dovetail and containment overlaps between reads. These overlaps were required to span two of the four read ends for each read pair. MER compressed consecutive identical bases to 1 base before tabulating

K-mers in reads. For reads that shared at least one compressed K-mer, MER used the uncompressed sequences to generate an overlap, if possible. MER parameters were selected in order to reduce overall run time and disk usage. MER was deployed with **K**=29 and **F**<=300; **F** is the K-mer frequency observed in the trimmed and filtered reads. The **F** threshold was designed to capture 99.96% of all distinct K-mers. This value of **F** captured 87.06% of the sequence data. The parallel computation required 2,471 CPU hours. The fraction of overlaps representing a containment overlap, defined as spanning both ends of at least one read, was 32%. Overlaps at 5' ends of reads were slightly underrepresented, accounting for 48% of overlaps per read. Overlaps per read correlated with read length (not shown) but this measure was consistently higher in unpaired libraries. The average number of overlaps per read exceeds expected coverage, probably due to genomic repeats. See Table S2.5.

Library Type	Average read length	Reads	Overlaps	Overlaps per read
Standard unpaired	264	45,590,642	1,786,365,438	39.2
Titanium unpaired	493	128,765,534	5,802,370,158	45.1
Ti 3Kbp paired	239	29,347,965	829,457,192	28.3
Ti 8Kbp paired	277	28,322,324	856,293,356	30.2
Ti 20Kbp paired	292	18,075,481	565,518,636	31.3
Total		250,101,946	9,840,004,780	

Table S2.5. Dovetail and containment overlap statistics before error rate correction.

Celera Assembler adjusted the error rates in overlaps to reduce the effect of sequencing error. Specifically, each read was examined in the context of its overlaps. Any base call contradicted consistently across overlaps was virtually corrected. Overlap error rates were recalculated based on the virtual corrections. Error rate correction increased by 26% the number of overlaps that satisfied the minimum 3% error rate for consideration by the unitig module.

Unitigs are preliminary contigs built from the graph of reads and overlaps. Celera Assembler's 454-specific unitig module is called BOG. As described previously [8], BOG built a "best overlap graph" in which each graph node represented a read end. Pairs of nodes representing the same read were joined by an undirected edge. The best overlap at each read end was represented by a directed edge. Best was defined as having the most bases in an alignment whose corrected error rate was 3% or less. BOG built "promiscuous" unitigs by greedy path following. It split promiscuous unitigs at path intersections. As a guard against chimera, it split unitigs that incorporated at least 7 mate constraint violations at one point. Reads with a containment overlap to the interior of some other read were excluded during the unitig construction phase but included during the mate-based splitting phase. Reads with multiple containment overlaps were placed according to the overlap with the lowest alignment error. BOG required 4 hours on one CPU using 50 GB RAM.

Celera Assembler's scaffold module, called CGW, constructed contigs and scaffolds from the unitigs, unitig overlaps, and mate pairs. The scaffold phase required 18 days on one CPU with 96 GB RAM. (The long run time was attributed to a costly cache flush that has since been reconfigured in the software.) Libraries with sufficient unitig co-placement of mate pairs had their insert size mean and standard deviation re-estimated. CGW re-estimated insert size for 68

of the 359 libraries. Average adjustments were -215bp (maximum +814bp) for 20 Kbp libraries, -1176 bp (maximum -1451 bp) for 8 Kbp libraries, and +777bp (maximum +920 bp) for 3 Kbp libraries.

CGW built a scaffold graph that had one node for each unitig. It added one edge for each between-unitig relationship that was supported either by two or more agreeing mate constraints or by one mate constraint and a consensus sequence alignment with compatible coordinates. For unitigs containing a single-read-coverage region that was not covered by satisfied mate constraints, CGW split the unitig and its node into two. CGW labeled nodes as repeat if it detected multiple paths in and out of the node, such that the paths were mutually exclusive based on node sizes (unitig lengths) and edge lengths along the paths. Additionally, CGW labeled nodes whose unitig A-stat coverage statistic [5] was less than unity. CGW reserved all repeat nodes for late stages of contig extension and gap filling.

CGW ran algorithms that merged unitigs into contigs and contigs into scaffolds. CGW promoted unitigs with A-stat>= 5 to contigs. Then it used unitigs with 1<=A-stat>=5 to merge remaining unitigs into contigs and contigs into scaffolds. CGW used algorithms to reduce portions of the graph into linear scaffolds. These algorithms included: greedy path merging [11]; transitive reduction [12]; and gap length estimation by least squares linear regression over mate constraints, modified to allow at most 20bp overlap between contigs with no sequence alignment. CGW ran the algorithms iteratively until the graph had stabilized.

The first round of CGW was followed by one round of the ECR (Extend Clear Range) module. ECR looked for trimmed sequence that could now be confirmed by nearby reads given the scaffold layout. It extended the clear ranges of confirmed reads. ECR closed 52,452 (32%) of the intra-scaffold gaps. Of gaps estimated to be 100 bp or less, ECR closed 50.6% whereas it closed only 18.6% of larger gaps. After ECR, the second round of CGW closed additional gaps and incorporated additional reads. The effects of ECR are shown in Table S2.6.

Statistic	i6 (no ECR)	i7 (with ECR)
Total scaffolds	11,891	12,032
Scaffolds with length >= 2 Kbp	2,798	2,695
Contigs in scaffolds	175,200	122,889
Contigs per scaffold	14.73	10.21
Scaffold N50	9,857,453	9,621,714
Contig N50	37,954	66,721
Bases in contigs (or scaffolds)	2,724,727,262	2,727,546,803
Reads in contigs (or scaffolds)	219,669,140	220,135,457

Table S2.6. The impact of Extend Clear Range (ECR) is evident in the comparison of the i6 and i7 assemblies. The Bonobo i6 assembly used parameter doExtendClearRanges=0. The Bonobo i7 assembly was generated by re-starting at the CGW stage with parameter doExtendClearRanges=1. (The i7 run exploited minor code adjustments that reduced CGW run time.) Thus, i6 ran CGW once with no ECR while i7 ran CGW, ECR, and CGW again. The i7 contig N50 is nearly double that of the i6 assembly. The i7 scaffold N50 is lower than the i6, probably due to merges of partially redundant contigs.

Finally, CGW addressed the unitigs it had labeled as repeat. CGW placed each unitig in zero, one, or multiple scaffold locations according to sequence overlaps and mate constraints by the process of "throwing rocks and stones" [5, 6]. Reads within the repeat unitigs were given a

specific contig and scaffold location if their mate was compatibly located within the same scaffold. Regardless of whether repeat reads could be so resolved, the repeat unitig consensus sequence was promoted into the contig sequence. Unitigs placed at this stage were called surrogates because their consensus had become a surrogate for their reads. See Table S2.7.

Statistic	Value
Surrogate unitig count	
Surrogates	187,254
Surrogate placements in scaffolds	202,046
Reads in surrogates	3,764,661
Surrogate reads placed in scaffolds	331,435
Surrogate unitig length (bp)	
Mean length	505
Maximum length	46,769
Combined length of surrogates	94,643,319
Combined length of surrogate placements	108,285,829

Table S2.7. Surrogate unitig statistics. A surrogate is a unitig that was deemed repetitive and was placed in scaffolds at one or more locations. The reads in a surrogate could be placed at most once in scaffolds, and then only by a mate constraint.

Any unitigs that were not incorporated into scaffolds were called “degenerates.” Any individual reads that were not incorporated into unitigs, contigs, or scaffolds were called “singletons.” See Table S2.8 for an accounting.

Statistic	Value
Degenerates	
Unitigs	1,221,536
Consensus bases	482,765,688
Loci with variant consensus	1,639,099
Reads in degenerates	18,420,765
Reads with mate also in degenerates	1,537,450
Average unitig length (bp)	395
Minimum unitig length (bp)	63
Unitigs with length<100 bp	50,902
Unitigs with length>=10 Kbp	147
Unitigs with length>=20 Kbp	12
Maximum unitig length (bp)	40,707
Singletons	
Reads	8,104,340
Bases	1,514,233,029
Average read length (bp)	186
Reads whose mate is also singleton	437,564

Table S2.8. Analysis of the unassembled sequence. Degenerates are unassembled unitigs of 2 or more reads. Singletons are unassembled single reads.

The Celera Assembler’s consensus module, CNS, ran on every contig in every scaffold. CNS used a progressive pair-wise approach to achieve a multiple sequence alignment. Especially in regions of low sequence identity, this approach can yield sub-optimal alignments due to propagation of gaps. To remove such artifacts, CNS applied its ‘abacus’ algorithm [5] to merge gap-rich regions within sliding windows along each multiple sequence alignment. Next, CNS used a column-wise voting algorithm to decide each consensus base call. Separately, it output

variant sequence at columns with sufficiently strong evidence of polymorphism, as described [9]. CNS reported alternates at 6,228,575 scaffold positions. Variant columns were not phased by read or mate. The variant phasing algorithm in CNS [9] was disabled out of concern that the Sanger-era implementation had not been tuned for 454 Titanium data. Finally, CNS assigned a consensus quality value to every scaffold base. 94% of bases received QV=60. The Ns in gaps received QV=0. The distribution of intermediate values is shown in Figure S2.3. CNS used a variant of the Churchill Waterman algorithm [13] as shown in Figure S2.4.

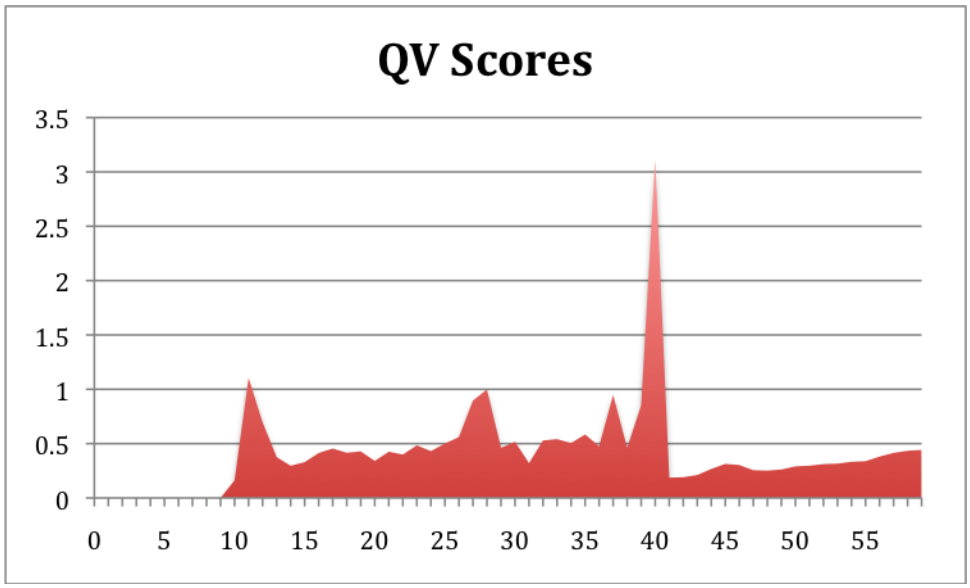


Figure S2.3. Distribution of the 1% of QV scores that were in the 1 to 59 range. X-axis: QV score. Y-axis: Number of bases with that score, in millions. Not shown: the 5.0% of scores at QV=0 and the 94% of scores at QV=60.

Initial conditions

Let D = the depth (number of underlying reads) at position i.
Let B[j=1..D] = the column of input base calls in D reads at position i.
Let Q[j=1..D] = the column of input quality values in D reads at position i.
Let C = the output consensus base at position i.
Let QV = the output quality value, an integer between 0 and 60 inclusive, for position i.
Let A[k] = members of the 5-letter alphabet A={A,C,G,T, -}. Dash indicates a gap inserted by aligner.

Assumptions

Assume every input B[j] is an element of A.
Assume every Q[j] is an integer in the range 0-60.

Special case behavior

If Q[J]==0 then Q[j]=5.
If the read depth D= 0 at column i, then C=gap and QV=0. Note this can happen due to surrogate unitigs in repeats.
If the read depth D = 1, then QV = Q[1]. That is, the input base and QV are promoted to the consensus.
C is chosen to maximize QV. In case of a tie, C is chosen randomly from the tied bases.

Formula

Here is intuition. Consider only column i. For read j, we will calculate the amount of support for all k=1..5 possible consensus base calls. We will use the given quality value Q[j]. Since Q indicates the probability of an error, we will subtract from unity to get probability of correctness. The fraction 1/4 is a prior; assuming all 5 base calls are equally likely, the 4 bases not used are each responsible for 1/4 of the residual probability. Note that every read offers some support even for a consensus base that does not match the read.

$$Pr(B_j == A[k]|Q_j) = \begin{cases} 1 - 10^{(-Q_j/10)} & B_j == A[k] \\ (\frac{1}{4})10^{(-Q_j/10)} & \text{otherwise} \end{cases}$$

For all 5 possible consensus base calls (k=1..5), take the product of the support from all (j=1..D) reads.

$$Q_k = \prod_{j=1}^{D} Pr(B_j == A[k])$$

Choose the base call with the maximum support. Break ties randomly.

$$Q_{max} = \max_{k=1}^{5} Q_k$$

Normalize the quality value such that the sum of all probabilities equals unity.

$$Q_{norm} = Q_{max}(\frac{1}{\sum_{k=1}^{5} Q_k})$$

Convert the probability of correctness to probability of error. Convert that to a phred-style quality score.

$$QV = round((-10)log_{10}(1 - Q_{norm}))$$

Figure S2.4. The Celera Assembler algorithm for consensus QV scores.

Celera Assembler's terminator module generated FASTA files representing the assembled scaffolds, the contigs, and the unitigs. It listed the unassembled singletons and degenerate unitigs. It gave the mappings of reads to the untigs, contigs, and scaffolds. It also generated a summary report called the QC file. These outputs are available upon request.

Analysis of Contigs and Scaffolds

The Bonobo i7 assembly process generated scaffolds whose combined span approaches the expected ~3 Gbp genome size. The assembly put 88% of reads in contigs. The assembly satisfied 86% of the mate constraints while violating only 0.12%. See Table S2.9.

Category	Statistic	Value
Scaffolds		
	Big scaffolds (length>=2 Kbp)	2,695
	Span of big scaffolds (including gaps)	2,857,577,652
	Total scaffolds	12,032
	Span of all scaffolds	2,869,032,589
	Total bases in scaffolds	2,727,546,803
	Scaffold N50 (bp)	9,621,714
Contigs		
	Big contigs (length>=10 Kbp)	52,061
	Total bases in big contigs	2,485,807,477
	Total contigs	122,889
	Total bases in contigs	2,727,546,803
	Contig N50 (bp)	66,721
	Loci with variant consensus	6,228,575
Reads		
	Average clear range (bp)	302
	Average read coverage in contigs	25.05X
	Usable reads	250,101,946
	Fraction placed in big contigs	81.73%
	Fraction placed in contigs	88.02%
	Fraction in placed repeats	1.51%
	Fraction in unassembled unitigs	7.37%
	Fraction as unassembled single reads	3.24%
Mates		
	Usable reads with a mate constraint	39,594,346
	Fraction with satisfied mate constraint	86.16%
	Fraction with violated mate constraint	0.12%
	Fraction with mate in different scaffold	2.29%
	It or its mate is in a repeat or unassembled sequence	11.58%

Table S2.9. Contig and scaffold statistics. All statistics are derived from the QC report generated by Celera Assembler. Both N50 statistics are based on total bases in scaffolds.

The read coverage in scaffolds has mean=25X and mode=26X. See Figure S2.5. The coverage distribution has a small tail of high-coverage positions, probably representing collapsed repeats. There is an excess of positions at low coverage. Some of this may be due to low coverage associated with 454 sequencing at %GC extremes. Some low coverage is certainly an artifact of the assembly process. Positions with 0X include the Ns in gaps between scaffolds as well as placements of repeat unitigs whose consensus was used as surrogate for reads that could not be resolved to a particular repeat instance.

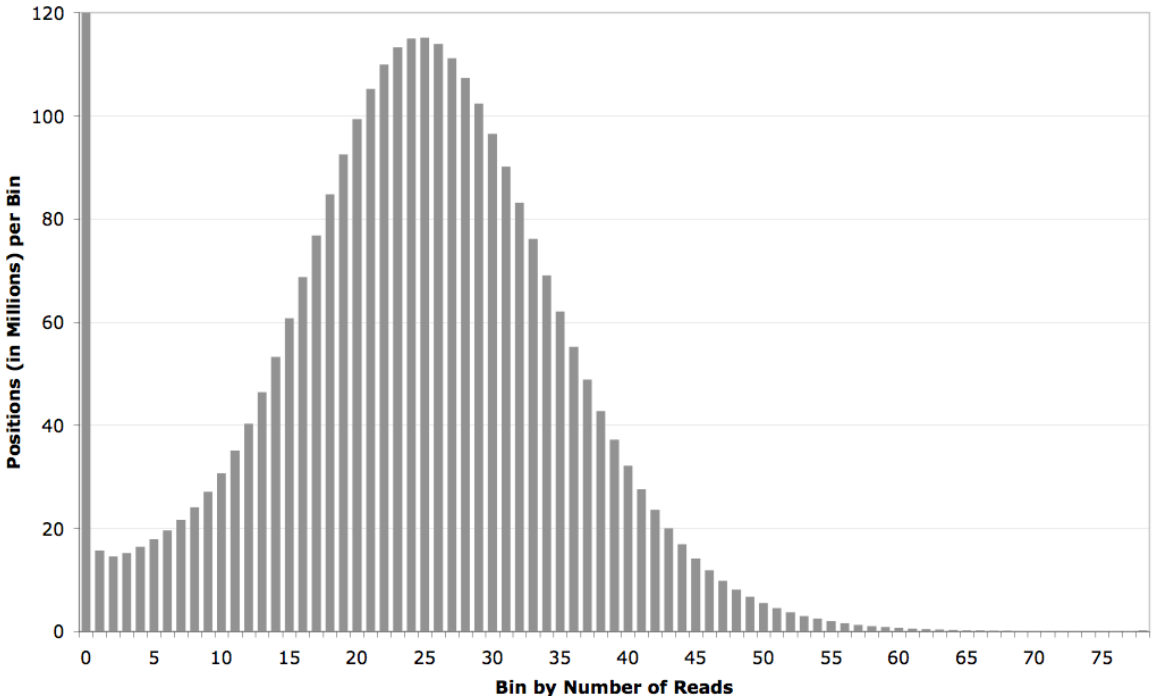


Figure S2.5. Coverage plot for scaffolds in the Bonobo i7 assembly. Each bar height indicates number of consensus columns whose read coverage matches its X-axis value. Bar at 0X is truncated from 152.9 million. All 0X positions are associated with repeats for which a consensus sequence was inserted as surrogate for reads that could not be individually resolved.

The Bonobo contigs were assessed for gene content using human transcript sequences. All available human mRNA transcript and exon sequences were downloaded from NCBI RefSeq and aligned to i6 contigs by BLAT [1] at 90% minimum identity. Of the 27,280 human transcripts, 27,218 (99.77%) mapped. Of the 207,009 human exons, 200,980 (97.09%) aligned by the same criteria. With the additional requirement of 90% transcript length coverage, 90% of transcripts had a mapping. The following gene IDs indicate transcripts not mapped to the contigs though their sequence could be at least partially recovered from the reads by alternate assembly techniques (not shown): C2orf85, CLN3, DEFB134, DEFB135, DEFB4, IGFL1, KDM5D, KRTAP19-2, MEIG1, NOMO1, NOMO2, NOMO3, OR10A4, OR4F17, OR4F4, OR4F5, PDE4DIP, PROP1, RPS4Y2, SEC22B, SPAG11A, SPAG11B, TEKT4, UTY, WASH1.

Analysis of Unitigs

Celera Assembler generated unitigs as seeds for its contig and scaffold construction. The unitigs were analyzed by alignment to a reference sequence. Bonobo unitigs were aligned to the NCBI hg19 human genome reference sequence. Human was selected because it is a close relative of bonobo and offers a high-quality genome reference. Alignments were generated with the ATAC software [7], which calculates the maximal, disjoint alignments that can be formed by chaining seeds of maximal, one-to-one, indel-free alignments. ATAC was chosen for its efficiency at human-scale whole-genome alignment. One drawback of ATAC is, relying on one-to-one alignments as seeds, it would not align any bonobo unitigs with end-to-end alignment to two or more repeats in the human genome. Bonobo unitigs of length 10Kbp or greater (average length = 19089bp) were selected for alignment. Alignments covered 99.5% of the unitigs and 98.1% of

the unitig consensus sequence. See Table S2.11.

Statistic	Value
Sequence in unitigs tested for alignment	1.557 Gbp
Sequence in alignments	1.528 Gbp
Unitigs tested for alignment	81,588
Unitigs aligned in 1 segment	81,140
Unitigs aligned in 2 segments	18
Unitigs aligned in 3 segments	1
Unitigs with no alignment	429

Table S2.11. Alignment of bonobo unitigs to the human reference genome. Unitigs aligned in at most 3 segments. Alignments spanned nearly all the unitig sequence.

Of the 19 unitigs with multi-segment alignments, 6 align to 2 human chromosomes or 2 distant loci in the human genome, and 13 align to two or three tandem loci in human such that one segment is inverted. The segmented alignments could be due to evolutionary differences or bonobo mis-assemblies. Celera Assembler's best.edges output file was searched for inter-unitig overlaps that qualified as the best pair-wise overlap from either read involved. No such overlaps were found within 1 Kbp of any of the 20 alignment breakpoints. This indicates that the unitigs with segmented alignments are at least consistent with the bonobo best overlap data.

Discussion: Problems with the Assembly

The CNS consensus module had been unable to calculate a consensus for five contigs. These five contigs were inspected manually and then adjusted such that consecutive unitigs abutted with no overlap. CNS ran to completion after the adjustments. The five contig IDs are 1120238064707, 1120238064708, 1120238064709, 1120238064710, and 1120238064711.

The coverage plot was examined at a fine scale. In regions of 70X or less, there were 560 pairs of successive positions that had a coverage difference of 40 or more. These coverage spikes were attributed to replicate mate pairs. Their existence in the assembly was attributed to a software bug in the Celera Assembler 5.4.3 mate filter that is fixed in higher-number versions.

Celera Assembler outputs two categories of sequence from the usable reads that was not incorporated in scaffolds. Its "degenerate" category refers to unitigs containing two or more reads, and its "singleton" category refers to individual reads. Most degenerates were small but a few were large. The longest degenerate had a full-length BLASTN alignment to human chromosome 18. With 4998 reads, its 43X coverage would have contributed to its exclusion from scaffolds. Overall, the degenerates had one variant locus per 295 bases compared the contig average of one per 438 bases; this is consistent with degenerates being enriched for collapsed repeats. The singletons are enriched for short reads, having a 186 bp average length well below the 302 bp average for usable reads.

One scaffold had invalid gap lengths. The maximum expected gap size is 20 Kbp based on the input paired end insert size estimates. Scaffold 1120238076601, with 1740 contigs, had 27 gaps larger than 30 Kbp and three gaps larger than 10 Mbp. Its large gaps may have been the result of a software failure to re-estimate gap sizes after a scaffold merge operation was attempted,

rejected, and reverted.

Seven small scaffolds have no read assignments. These are composed entirely of repeat unitigs, an unexpected result. These seem to be artifacts of incomplete scaffold splitting due to an unidentified software bug. The scaffolds are 1120388623473, 1120388623481, 1120388623521, 1120388623526, 1120388623536, 1120388623539, and 1120388623550. Their size range is 582 to 6936 bp.

A quality control screen revealed the presence of contaminant DNA from another species known to be present simultaneously in the sequencing laboratory (see SI 1). The contamination was detected during the Bonobo i7 assembly computation, too late to prevent its inclusion. Sequence analysis limited the apparent contamination to the FLGU9LC01.sff file, representing all of one Titanium unpaired library. Reads from this file were compared to genomic sequence with BLASTN [14]. Reads were designated ZM or PT based on whether BLAST returned a higher scoring alignment to *Zea mays* (corn) or *Pan troglodytes* (chimpanzee). This classifier identified 50.8% ZM reads and 48.8% PT reads with 0.4% unclassified.

The assembly was tested for clustering of the ZM-contaminant reads. The Celera Assembler read-to-scaffold mapping file (i7.posmap.frgscf) was searched with the ZM read IDs. Of ~12000 scaffolds in the assembly, 2500 included reads from the contaminated run. Of contaminated scaffolds, nearly all were composed of exactly 100% ZM reads or exactly 0% ZM reads. Of the 48 scaffolds that mixed ZM and non-ZM reads, 36 were over 90% ZM and the rest contained exactly one or two ZM reads. Thus, the ZM reads appear highly clustered and separate from bonobo sequence in the Bonobo i7 assembly. See Table S2.12.

Data set	Reads	ZM Reads	Scaffolds
Reads			
The FLGU9LC01.sff file	630,543	320,156	
Scaffolds			
All scaffolds	220,135,457	267,034	12,032
Scaffolds with FLGU9LC01 content	219,401,203	15,418	2,446
… and 100% ZM content	11,633	11,633	893
… and 90%<= ZM content < 100%	3,860	3,772	36
… and 1% <= ZM content < 90%	0	0	0
… and 0% < ZM content < 1%	7,885,721	13	12
… and ZM content = 0%	211,499,989	0	1,505

Table S2.12. Clustering of contaminant reads in the Bonobo i7 assembly. FLGU9LC01 is the putatively contaminated sequencing run. ZM Reads are reads presenting *Zea mays* (corn) sequence.

The scaffolds with ZM content of 90% or more were excluded from the assembly.

Acknowledgements.

The authors are grateful to the J. Craig Venter Institute and the JCVI IT department, particularly Eddy Navarro, Hank Wu, John Bury, Darnell Edwards, and Marty Stout. Authors JM, BW, SK, and GS received funding from the National Institutes of Health via grant 2R01GM077117-04A1 from the National Institute of General Medical Sciences.

Supplementary Information 2a

Initial Assembly QC, SNP Calling and Chromosomal Assignment of Scaffolds

James C. Mullikin[1,*], Kay Prüfer[2,*] and Ines Hellmann[3]

1. National Human Genome Research Institute, NIH, Bethesda MD USA
2. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
3. Max F. Perutz Laboratories, University Vienna, Vienna, Austria

* To whom correspondence should be addressed (mullikin@mail.nih.gov, pruefer@eva.mpg.de)

We use alignments of the bonobo i7 scaffolds to the chimpanzee genome to detect breakpoints in the contiguity of scaffolds. The total count of breakpoints indicates a low amount of misassemblies in the bonobo genome. Based on the mapping to chimpanzee, we assign and order bonobo scaffolds to provide a predicted chromosomal location.

Furthermore, we use 22kb of previously published Sanger sequencing data from Ulindi, the individual sequenced in this genome project, to assess the quality of the assembled sequence. We use known SNPs in the Sanger data to gauge how effective our criteria discriminate between true and false heterozygous sites.

Inter-Chromosomal Misjoins and Chromosomal Assignments of Scaffolds

In order to quantify breakpoints in the bonobo assembly, we aligned scaffolds divided into 2040bp sequence segments every 2000 bases to the panTro2 assembly using ssahaSNP [15]. SsahaSNP parameters allowed up to 4.5% sequence divergence between the 2040bp segments and the panTro2 assembly. To reduce mapping errors, segments were removed if they aligned to multiple places and the best alignment contained more than 30% of the sequence divergence (measured in single nucleotide differences per kilobases aligned) observed in the second best alignment. 1.6% of the 2040bp sequence segments were dropped because of this mapping similarity exclusion criterion. Out of 1505 scaffolds greater than twice the segment length (>4080 bp), this alignment process placed 1,216 scaffolds onto the panTro2 assembly. Of these, a total of 110 scaffolds aligned to more than one chromosome with at least 10 segments each when ignoring mappings to panTro2 chr_random and chrUn alignments. These criteria give an upper limit of 141 misjoins, some of which may be due to true rearrangements between bonobo and chimpanzee or due to misjoins in the chimpanzee assembly.

To place a scaffold onto the chimpanzee genome, at least twenty-five 2040bp segments had to map in proper order and orientation. Any scaffolds not meeting this minimum were grouped into chrUn

for unknown placement, including regions within a scaffold not meeting this minimum. A total of 534 scaffolds, spanning 1,690,681,950 bases, were placed onto chimpanzee chromosomes as complete scaffolds. Another 86 scaffolds were broken into 200 fragments through this placement process, spanning a total of 1,038,686,943 bases. In the chrUn bin, there are 10,035 scaffolds spanning 139,308,438 bases. One scaffold accounts for the majority of these bases, scf1120388623559, spanning 92,723,997 bases of which 90,985,837of these are Ns (see also SI 2). The other 10,034 scaffolds span 46,594,475 bases with an average size of 4.6kb. The largest unmapped scaffold is scf1120388622904 with 3,709,346bp and it was grouped into chrUn because it mapped almost entirely to chr8_random from panTro2. The PpV07.agp file reports the above scaffolds as mapped with this procedure and is available through http://bioinf.eva.mpg.de/bonobogenome/PpV07.agp.

Comparison with Sanger Sequencing Data

We used sequencing data from a previous study [16] to test the accuracy of the assembled bonobo genome sequence (henceforth called Fischer dataset). These high quality sequences span a total of 22 kilo bases in 26 separate regions on autosomes and stem from the same donor ("Ulindi"). SNPs in the Sanger sequence have been called manually by inspection of trace files. A total of 35 SNPs have been detected in the heterozygous sequence.

We mapped all 26 sequences to the bonobo i7 scaffolds using *blat* [1] (option –fastMap). Each sequence had one clearly best alignment spanning the entire length of each region. We realigned bonobo scaffold sequences and the Fischer sequences with *muscle* [17] to generate pairwise alignments for each region. In these alignments, we detected no mismatches between the Fischer sequences and assembly sequence. However, we found one erroneous gap in the assembly which is located at a position of a SNP.

Quality Scores at SNP Positions

Typically Quality Scores are assigned to each consensus base in an assembly. The value of the quality score is determined by the quality scores of the reads and their base calls. At heterozygous positions in the genome, quality scores tend to be lower since reads are disagreeing in base calls. Therefore, quality score filtering to increase the quality of the assembled sequence can have the side-effect of excluding SNP positions. This side-effect can for instance lead to an apparent shorter lineage length in genome comparisons. Here, we investigate how the quality scores calculated by the Celera Assembler are distributed at heterozygous genomic positions in the bonobo genome (see also SI 2 for details on how the quality scores are calculated).

We tested the effect of heterozygous sites on quality scores using the 35 known SNP positions in

the 22 kilo bases of Sanger sequenced regions from the Fischer dataset. Table 2a.1 shows the location, base and quality scores for 34 of the 35 positions (one SNP aligned to a gap in the consensus; see also previous section). With a quality score cutoff of 30, we would have included 28 SNP positions and excluded 6.

Scaffold	Location	SNP	Consensus base	Quality
scf1120388623507	14807702	Y	C	60
scf1120388623513	10125354	Y	C	60
scf1120388623383	4586925	R	G	60
scf1120388623469	13201743	R	A	60
scf1120388623538	3048472	Y	C	48
scf1120388623058	339706	K	G	60
scf1120388623058	339776	R	G	60
scf1120388623058	339906	R	G	60
scf1120388623462	9802502	Y	C	24
scf1120388623422	4732519	R	G	60
scf1120388623422	4732762	Y	C	60
scf1120388623468	14210707	M	A	60
scf1120388623408	1982847	Y	C	4
scf1120388623408	1983292	R	G	48
scf1120388623459	14827418	K	T	10
scf1120388623363	277133	S	C	60
scf1120388623363	277146	Y	T	15
scf1120388623363	277192	Y	C	19
scf1120388623363	277213	W	A	60
scf1120388623363	277371	K	T	60
scf1120388623363	277380	Y	C	57
scf1120388623363	277393	Y	T	60
scf1120388623363	277433	Y	T	60
scf1120388623363	277544	M	C	60
scf1120388623443	1667524	Y	C	60
scf1120388623443	1667904	R	A	60
scf1120388623443	1668330	Y	T	60
scf1120388623427	7368895	R	A	60
scf1120388623427	7369322	R	A	60
scf1120388623427	7369445	R	G	17
scf1120388623420	7369794	Y	C	60
scf1120388623420	7369888	R	A	60
scf1120388623420	7370574	M	C	60
scf1120388623465	1043852	R	G	60

Table S2a.1: 34 known SNP positions with consensus base and quality score in the bonobo assembly. SNPs are given as ambiguity codes (Y=C,T; R=A,G; W=A,T; S=C,G; K=G,T; M=A,C).

SNP Calling in 454 Sequencing Data

In order to call heterozygous sites in Ulindi, we remapped all 454 reads to the human genome (hg18) using *BWA* [18] (Version: 0.5.7; long read option: *bwasw*, otherwise standard parameters). We filter the alignments according to the following criteria:

- Each alignment has a mapping quality (see [19]) of at least 30
- Each base has a minimum quality of at least 20 for the middle base and 15 for the five neighboring bases on each side (NQS, see: [20, 21])
- Bases within a distance of five base pairs to an gap or insertion in the alignment are ignored
- Bases within a distance of five base pairs of two or more differences to the target database are ignored

After filtering alignments, heterozygous positions are identified with an additional set of filters:

- At most 50 reads are covering the position
- At least 3 reads are supporting both alleles
- With n reads supporting the highest allele and m reads supporting the second highest allele, we exclude sites if $\sum_{k=0}^{m}\binom{n+m}{m}0.5^{n+m} < 0.025$ (i.e. when the ratio of major to minor allele deviates significantly from the expected 0.5 ratio).

We compared all called SNPs (according to the above criteria) to the known 35 heterozygous positions from the Fischer dataset (see Table S2a.2 and S2a.3). With the above criteria, we detected 31 SNPs corresponding to the previously known SNP positions and no additional SNPs. Two true SNPs were not found because less than 3 reads supported each allele; the remaining two SNPs were not reported because the second highest allele was much more rare than expected when assuming equal chance of picking reads from each allele.

True SNPs	35
Detected SNPs	31
Additional SNPs	0

Table S2a.2: Evaluation of SNP calling procedure against a set of known SNPs. Detected SNPs give the number of SNPs detected out of all SNPs present in the Fischer dataset. Additional SNPs give the number of SNPs detected in addition to the SNPs in the Fischer dataset.

Chr (hg18)	Location	#A	#C	#G	#T	Passing SNP Criteria
chr4	104993460	0	3	0	7	Yes
chr11	29517116	8	0	11	0	Yes
chr17	12060099	4	0	15	0	No
chr5	62685444	10	0	10	0	Yes
chr14	61854333	9	0	8	0	Yes
chr2	107639985	0	13	0	8	Yes
chr2	107640114	0	15	0	7	Yes
chr2	107640184	10	14	0	0	Yes
chr12	46058026	0	8	0	9	Yes
chr22	34643495	10	0	13	0	Yes
chr22	34643738	0	10	0	6	Yes
chr18	57111474	8	11	0	0	Yes
chr8	129374631	0	3	0	7	Yes
chr8	129375076	9	0	12	0	Yes
chr3	119987542	0	0	11	8	Yes
chr2	151420353	0	0	9	7	Yes
chr2	151420354	8	0	7	0	Yes
chr2	151420465	8	0	9	0	Yes
chr2	151420505	10	0	8	0	Yes
chr2	151420518	6	0	8	0	Yes
chr2	151420527	9	8	0	0	Yes
chr2	151420686	8	0	0	6	Yes
chr2	151420707	9	0	7	0	Yes
chr2	151420753	8	0	6	0	Yes
chr2	151420766	0	5	7	0	Yes
chr6	14870719	0	2	0	2	No
chr6	14871104	8	0	5	0	Yes
chr6	14871530	0	8	0	11	Yes
chr5	10037990	0	2	0	5	No
chr5	10038113	0	4	0	6	Yes
chr5	10038540	0	9	0	12	Yes
chr20	7621759	0	12	0	4	No
chr20	7621853	5	0	7	0	Yes
chr20	7622538	9	6	0	0	Yes
chr5	128233846	0	17	0	9	Yes

Table S2a.3: Known SNPs and SNP calling based on differences between reads. #{A,C,G,T} give the number of reads showing base A, C, G, T at the given position.

Supplementary Information 3

Genome Alignments and Quality Control for the Bonobo Genome Assembly

Kay Prüfer*, Susan Ptak, Janet Kelso and Svante Pääbo

Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

* To whom correspondence should be addressed (pruefer@eva.mpg.de)

For subsequent analyses and quality assessment, we use all scaffolds of the i7 assembly (see SI 2). As a first step, we produce pairwise whole genome alignments of the bonobo, chimpanzee [3], orangutan and rhesus macaque [22] genomes to the human genome [23]. From these pairwise alignments we generate various multiple sequence alignments. We use both pairwise and multiple sequence alignments with the human and chimpanzee genome to assess sequence accuracy and the completeness of the bonobo genome sequence.

Whole Genome Alignments

In a first step, we aligned scaffolds of the i7 bonobo assembly and the chromosome-assigned sequence of chimpanzee (panTro2), orangutan (ponAbe2) and rhesus macaque (rheMac2) to the human genome (hg18) using *lastz* (Version 1.01.50) [24]. All alignments used an identical set of parameters (--gap=600,150 --hspthresh=4500 --gappedthresh=2200 --inner=2000 --seed=12of19 --notransition --ydrop=15000 –masking=254; scoring matrix identical to the matrix used at UCSC for hg18 pantro2 alignments: http://hgdownload.cse.ucsc.edu/goldenPath/panTro2/vsHg18/). The aligner will not seed alignments in lower-case masked regions. In order to avoid biases due to different repeat databases, we converted all query genome sequences to uppercase, thus removing the repeat masking information. We kept lowercase masking for the target genome (human genome version hg18, using lowercase repeat masking from RepeatMasker and TandemRepeatsFinder as available from the UCSC Genome Browser under URL: http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/). With these precautions we aim to make the pairwise sequence alignments as comparable as possible.

Pairwise alignments were post-processed following closely the UCSC Genome Browser pipeline [25, 26]. In particular, we use the programs *axtChain* (with parameters: -minScore=5000 -linearGap=medium), *chainAntiRepeat, chainMergeSort, chainPreNet*, chainNet and *netSyntenic* to produce chained and netted alignments; *netChainSubset* and *chainStitchId* were used to generate *liftover* files. Files were converted to maf and axt format using *netSplit, netToAxt, axtSort* and *axtToMaf*. All

programs were compiled from jksrc v130 downloaded 2009-07-07. See Table S3.1 for an overview of all pairwise whole genome alignments prepared.

We produced three multiple sequence alignments. The first consists of the genomes of human, chimpanzee and bonobo; the second adds the orangutan genome, and the third adds the rhesus macaque and orangutan genomes to these three species. Chained and netted pairwise whole genome alignments from the previous section were used as the input for generating whole genome alignments using multiz (Version: 012109)[27]. All pairwise alignments were pre-processed with *single_cov2*. The program *roast* was used to join pairwise alignments. Table S3.2 gives the input files and parameters for the three multiple sequence alignments prepared for this study.

Genome Alignment	Query	Target
Human-Bonobo	bonobo i7	hg18
Human-Chimpanzee	panTro2	hg18
Human-Orangutan	ponAbe2	hg18
Human-Macaque	rheMac2	hg18

Table S3.1: Pairwise whole genome alignments prepared for this study

Multiple Sequence Alignment	Input alignments	Tree Topology
HCB	Human-Bonobo; Human-Chimpanzee	(hg18 (panTro2 bonobo))
HCBO	Human-Bonobo; Human-Chimpanzee; Human-Orangutan	((hg18 (panTro2 bonobo)) ponAbe2)
HCBOM	Human-Bonobo; Human-Chimpanzee; Human-Orangutan; Human-Macaque	(((hg18 (panTro2 bonobo)) ponAbe2) rheMac2)

Table S3.2: Whole genome multiple sequence alignments prepared from pairwise alignments

Genome Completeness

The phylogenetic relationship between bonobo, chimpanzee and human, in which chimpanzee and bonobo are equidistant to the human genome sequence, gives us the opportunity to put the completeness of the bonobo genome sequence in perspective to the draft chimpanzee assembly. Assuming that no large scale duplications or deletions happened on either the bonobo or chimpanzee lineage, the number of bases in alignment with the human genome sequence can be assumed to be equal when the chimpanzee and bonobo assemblies accurately depict the complete sequence of both species. Here, we use the pairwise chimpanzee-human and bonobo-human whole genome alignments from the previous section to test for

genome completeness. These pairwise alignments use identical parameters and seeding filters, so that we can expect equal amounts of human bases in alignment if both genomes are equally complete.

Figure S3.1 and Table S3.3 show the number of human bases in alignment to the bonobo assembly normalized by the number of bases in alignment to the chimpanzee assembly. Genome-wide coverage by bonobo bases is lower with 99.6% the value of the chimpanzee covered human bases. A total of 9 of 22 autosomes yield more aligned bases in bonobo than chimpanzee. The deviation from the chimpanzee aligned bases is around 1% in either direction with the exception of the chromosome 9 and 16. These results show that the autosomal sequence of the bonobo assembly is largely as complete as the chimpanzee assembly.

Chromosome 9 and 16 are outliers compared to the rest of the autosomes. For these two chromosomes, 3-4% less bonobo sequence aligns to human as compared to the chimpanzee assembly. Part of this discrepancy may be explained by an overcollapse of segmental duplications in the bonobo assembly (see SI 4). Both chromosomes are known to be enriched for segmental duplications [28-30].

In contrast to the autosomes, chromosome X shows an excess of roughly 17% more sequence alignable from the bonobo assembly. Since the chimpanzee genome was assembled from sequencing data of a male individual, this difference can be explained by the lower coverage of chromosome X compared to the other autosomes in the chimpanzee assembly.

We further test the potential effect of overcollapsed repetitive sequence in the bonobo genome assembly by dividing the human genome sequence in repeat-masked and not repeat-masked sequence (according to UCSC Genome Browser lower-case masking for hg18 from tandem-repeat-finder and RepeatMasker annotation). Figure S3.2 and Table S3.3 show the results. We observe that chromosome 9 and 16 show that the unaligned sequence is more often repetitive in the human genome. The additional coverage on chromosome X for the bonobo assembly stems primarily from repetitive sequence.

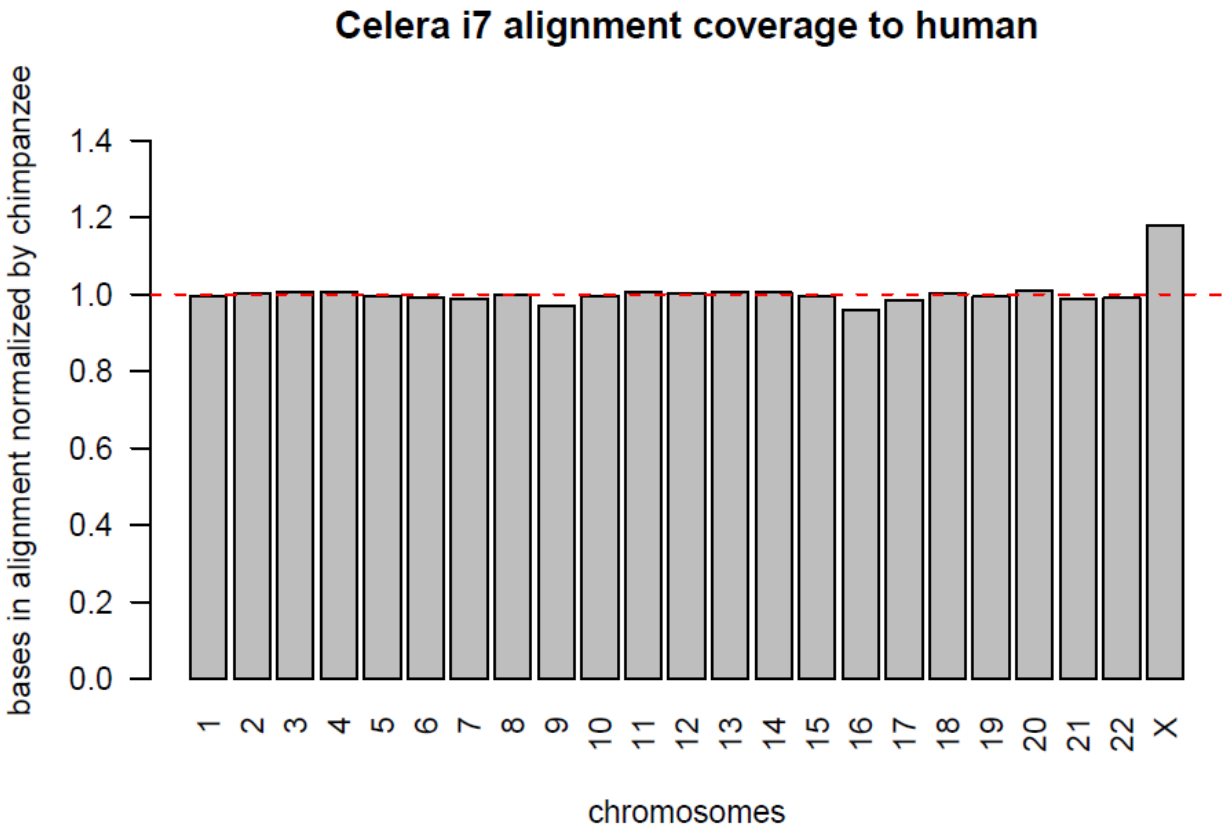


Figure S3.1: Human bases in alignment to bonobo i7 assembled sequence normalized by the chimpanzee aligned bases. The red dashed line gives the expected coverage if bonobo sequence aligns as well as chimpanzee sequence.

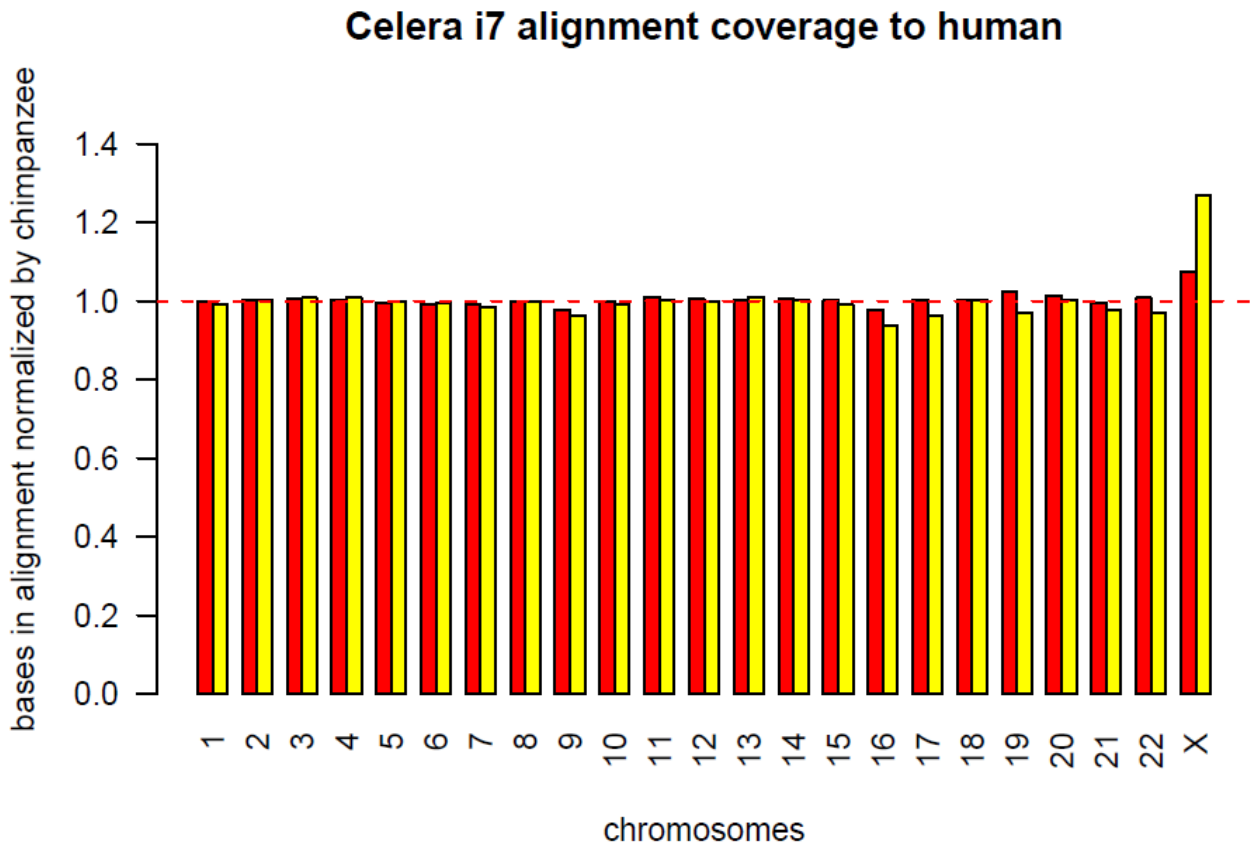


Figure S3.2: Human bases in alignment to bonobo i7 assembled sequence normalized by the chimpanzee aligned bases. Red bars show the measure for human non-repetitive sequence and yellow bars for repeatmasked sequence. The red dashed line gives the expected coverage if bonobo sequence aligns as well as chimpanzee sequence.

Chromosome	Coverage (all bases)	Coverage (repeatmasked)	Coverage (non-repetitive)
1	99.5%	99.8%	99.1%
2	100.3%	100.2%	100.4%
3	100.6%	100.4%	100.8%
4	100.5%	100.2%	100.8%
5	99.6%	99.5%	99.8%
6	99.2%	99.1%	99.4%
7	98.9%	99.3%	98.4%
8	99.9%	99.9%	99.9%
9	97.1%	97.8%	96.3%
10	99.7%	100.0%	99.3%
11	100.6%	100.9%	100.3%
12	100.2%	100.4%	100.0%
13	100.6%	100.3%	100.9%

14	100.4%	100.5%	100.4%
15	99.6%	100.0%	99.0%
16	95.7%	97.6%	93.7%
17	98.4%	100.4%	96.2%
18	100.2%	100.2%	100.0%
19	99.4%	102.4%	96.9%
20	100.8%	101.3%	100.2%
21	98.7%	99.6%	97.6%
22	99.0%	100.8%	97.0%
X	117.8%	107.4%	126.8%

Table S3.3: Human bases in alignment to bonobo i7 assembled sequence normalized by the chimpanzee aligned bases.

Sequence Accuracy

The phylogenetic relationship of chimpanzee and bonobo to human can also be used to compare the sequence accuracy between chimpanzee and bonobo. Here we use the HCB multiple sequence alignment to parsimoniously assign differences between bonobo and chimpanzee to the bonobo and chimpanzee lineages using the human genome as an outgroup sequence. Assuming no difference in rate of change between the bonobo and chimpanzee lineage the number of fixed differences is expected to be very similar on both lineages. However, a difference in sequencing error will inadvertently increase the number of assigned differences. Thus a difference in assigned changes can be interpreted as a sign of difference in sequencing error.

When analyzing all multiple sequence alignments including human autosomal sequence, we are able to assign a total of 5.67 million differences to the chimpanzee and 5.71 million differences to the bonobo lineage. If all additional differences on the bonobo lineage are regarded as error, the bonobo sequence would contain an error of 1.5 differences per 100,000 base pairs in excess of the chimpanzee genome.

The comparison of the human chromosome X gives 288 thousand assigned changes on the chimpanzee lineage compared to 199 thousand on the bonobo lineage. This difference can again be explained by the lower coverage of chromosome X sequence in the chimpanzee assembly, but could also be further exacerbated by chromosome Y misalignments.

For the next comparison we only considered changes assignable with human chromosome 21. The chimpanzee version of chromosome 21 is of finished quality, with an estimated error rate of less than two errors in 100,000 basepairs [31]. Thus, a higher number of assigned changes are expected for the bonobo lineage. Indeed, we see over 8000 changes more on the bonobo lineage than on the chimpanzee

lineage (chimpanzee=76146; bonobo=82761, expectation of equal rate, binomial test p-value < 10^{30}). Assuming no error in the assignment of differences, no mutation rate difference between chimpanzee and bonobo, and no error in the chimpanzee chromosome 21 sequence, we can estimate the error of the bonobo genome sequence based on the excess of changes. This way, we estimate the error in the bonobo scaffolds to be 2.1 per 10,000 base pairs. When considering human repeatmasked and non repetitive sequence separately, we see that sequencing errors are increased in the repeats. We estimate the error to be 0.9 per 10,000 base pairs for non-repetitive sequence and 3.6 per 10,000 base pairs for repetitive sequence.

Sequence	**Estimated Error Rate**
All chromosome 21	2.1×10^{-4}
Repeatmasked chromosome 21	0.9×10^{-4}
Repetitive chromosome 21	3.6×10^{-4}

Table S3.4: Error rate estimates based on parsimony assignment of differences to the chimpanzee and bonobo lineages using human as outgroup.

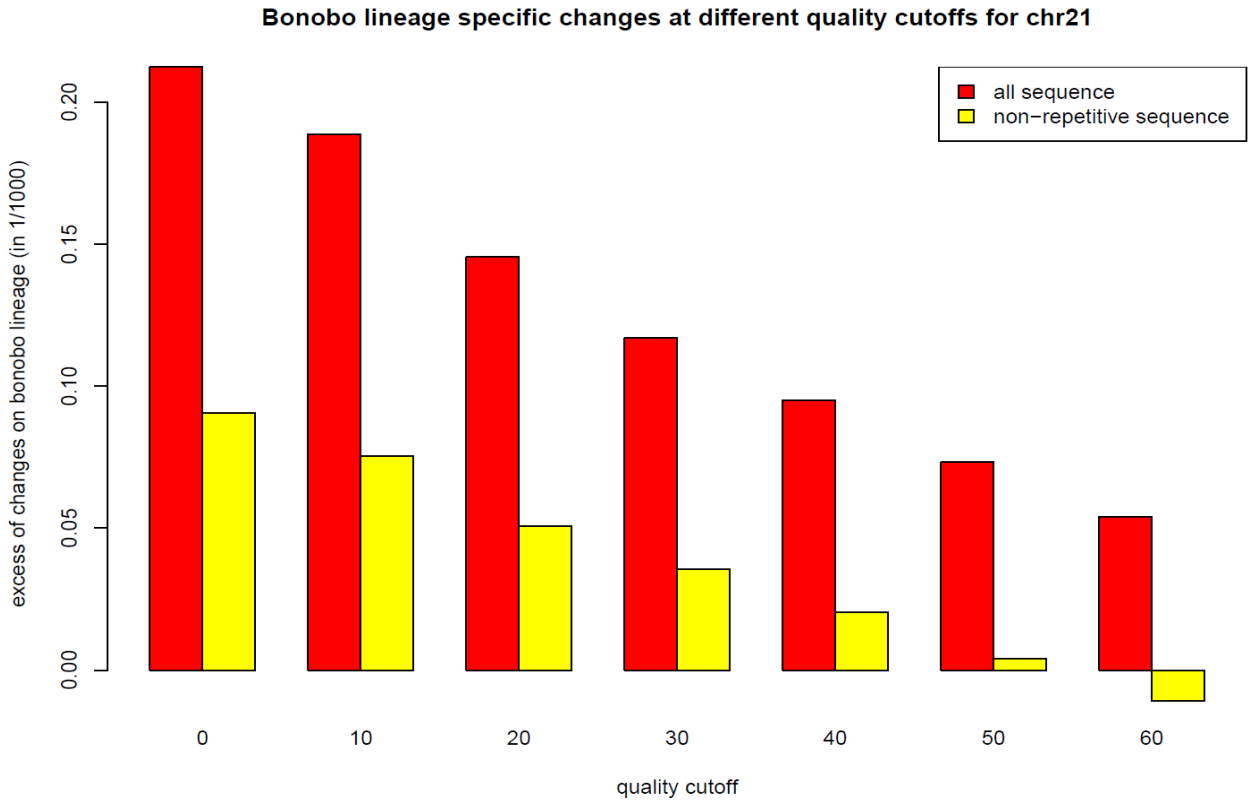


Figure S3.3: Estimated error rate at different quality score cutoffs for the comparison with chromosome 21. For each cutoff all bases with at least this quality score are counted. Yellow bars represent repeatmasked sequence.

Quality Scores and Sequence Accuracy

The Celera Assembler calculates phred-like quality scores for each assembled base, indicating the confidence for each consensus base dependent on the quality scores of the raw sequence reads. For analyses, we can use quality score cutoffs to further improve the accuracy of the bonobo genome assembly sequence. Using the multiple sequence alignment with the human chromosome 21, we can test to what degree quality score cutoffs (i.e. bases at a certain quality score or with higher score) improve the sequence quality in relation to the finished chimpanzee chromosome 21. Figure S3.3 shows the estimated error rate for all compared bases and human repetmasked sequence. At quality score cutoff of 40, the error rate over all sequence is estimated to less than one error per 10,000 base pairs. Interestingly, the non-repetitive sequence is estimated to contain less than 2 errors in 100,000 bases at the same quality score cutoff. We observe a similar difference between average error rate between all bases and non-repetitive bases over all different quality score cutoffs. This shows that more error is present in repetitive sequence and that this error is not reflected in the assembly quality scores. Instead the higher error rate may be caused by misassembly or misalignment affecting repetitive sequence stronger than non-repetitive.

Supplementary Information 3a

Indel Error Assessment of the Bonobo Genome Sequence Assembly

Stephen J. Meader[1]*, Chris P. Ponting[1] and Gerton Lunter[2]

1. MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, South Parks Road, Oxford, OX1 3QX. United Kingdom.
2. The Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, United Kingdom

* To whom correspondence should be addressed (stephen.meader@dpag.ox.ac.uk)

The fine scale accuracy of the bonobo genome sequence assembly was assessed by examining the distribution of insertion and deletion mutations (indels) in comparisons to both chimpanzee and human genome sequence assemblies. The Neutral Indel Model [32] can be exploited to quantify the numbers of erroneous gaps within a genome sequence alignment [33]. These gaps represent nucleotides which were wrongly inserted or deleted during the sequencing or assembly process. For high quality assemblies, the frequency of short aligned blocks between adjacent alignment gaps (inter-gap segments) is well approximated by a geometric distribution. As assembly quality decreases in one or both of the aligned assemblies, often as a result of a lower read coverage, we observe an excess of short inter-gap segments over the predictions of the Neutral Indel Model. This excess of short segments over the predictions of the model reflects clusters of gaps in the alignment which result from indel errors [33].

In an initial analysis, the autosomal and non-repetitive component of the bonobo genome sequence assembly was analysed. The bonobo genome sequence was compared to that for chimpanzee using alignments of the two assemblies with the human genome sequence assembly (**Figure S3a.1,** see SI 3 for details on the preparation of pairwise alignments). Any differences in indel errors between the two alignments reflect differences in the quality of the bonobo and chimpanzee genome assemblies. Of the 1070Mb of the non-repetitive bonobo genome sequence alignable to the human assembly, it was estimated that 11.8% of observed indels were errors (0.276 errors per kb, 95% c.i. 0.268-0.282) (**Table S3a.1**), a slightly increased error frequency when compared to a similar alignment of human and chimpanzee assemblies (0.232 errors per kb, 95% c.i. 0.225-0.238). Errors were greatest in G+C-rich or G+C-poor sequence (**Figure S3a.2**), as has been observed previously with other great ape genome alignments [33].

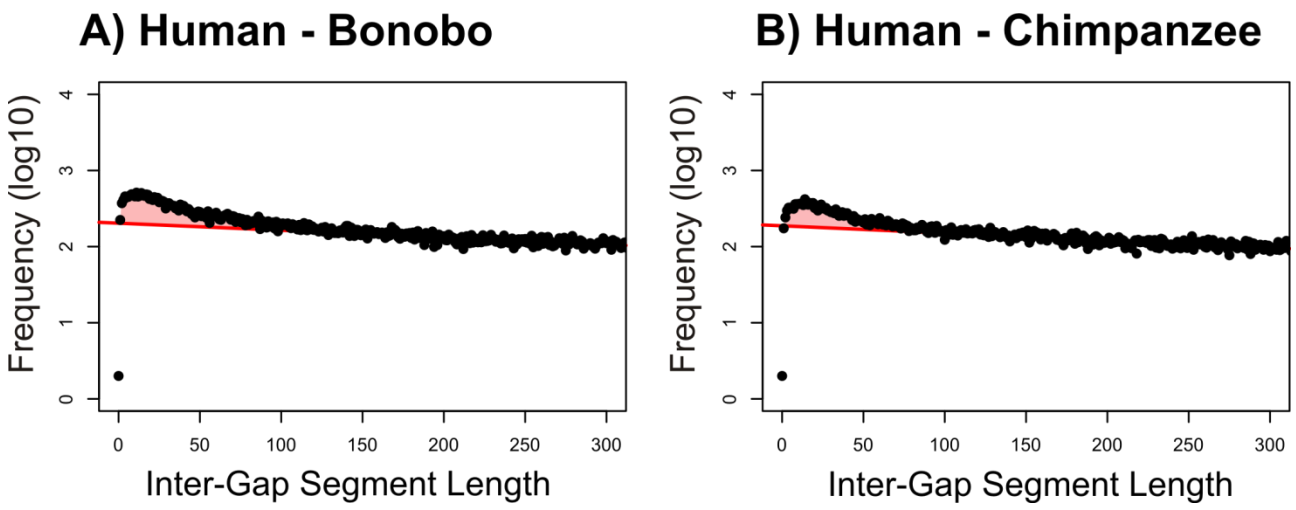


Figure S3a.1: Quantifying gap errors in the bonobo genome sequence assembly for a single G+C bin (G+C content: 0.368 - 0.380). Frequency histograms (log_{10} scale) of inter-gap segment lengths for a single G+C bin representing whole genome alignments of (A) human and bonobo assemblies, and (B) human and chimpanzee assemblies

Primary Species	Secondary Species	Alignable Sequence	Excess indels (*Ng*)*	Percentage of Total Indels (ε)*	Indel Error Rate (*D* per kb) *
Human	Bonobo	1072 Mb	296k (288k-303k)	11.8 (11.5–12.1)	0.276 (0.268-0.282)
Human	Chimpanzee	1052 Mb	243k (236k-250k)	10.4 (10.1–10.7)	0.232 (0.225-0.238)
Orangutan	Bonobo	1344 Mb	969k (955k-981k)	19.0 (18.7-19.3)	0.769 (0.758-0.779)
Human	Orangutan	1260 Mb	890k (875k-905k)	17.1 (16.8-17.4)	0.661 (0.651-0.674)

*values in brackets represent 95% confidence intervals

Table S3a.1: Indel error estimates for the bonobo genome assembly based upon analyses of pairwise alignments. *Ng* is the total number of indels in excess of the predictions of the Neutral Indel Model. ε represents the percentage of all indels within the alignment analysed that are estimated to be errors. The indel error rate *D* was calculated by dividing *Ng* by the total number of aligning bases [33].

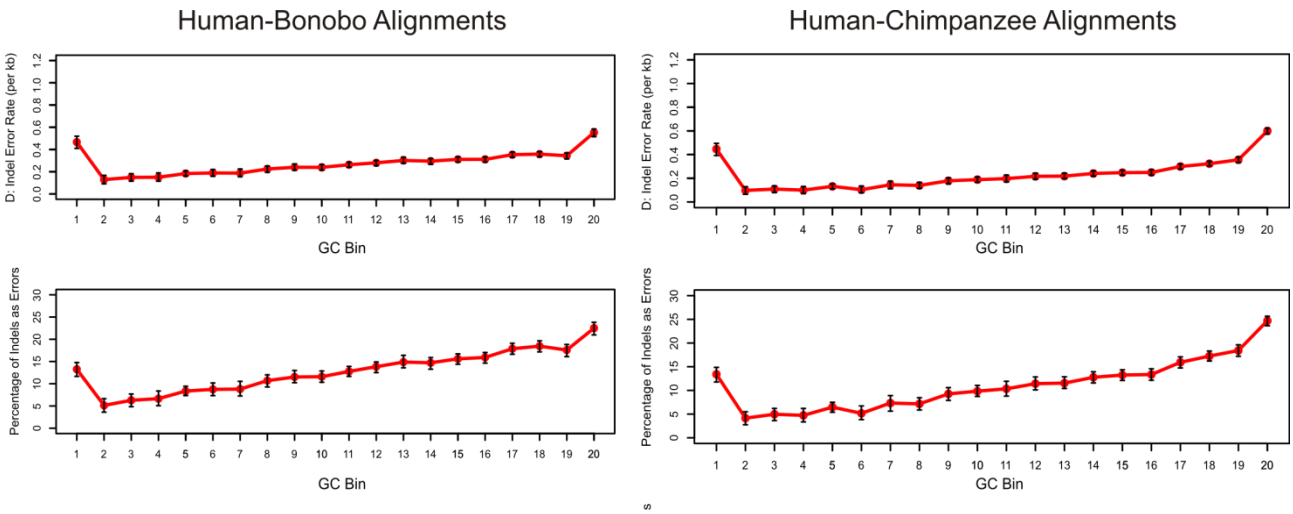


Figure S3a.2: Estimates of indel errors in 20 equally populated G+C bins for alignments of genome sequence assemblies for human-bonobo (left) and human-chimpanzee (right). Estimates of indel errors are presented as the indel error rate *D* (top) and the percentage of all indels that are errors ε (bottom).

The assembly for the X chromosome of the chimpanzee genome is of lower quality than the autosomal sequence, presumably a result of reduced read coverage for sex chromosomes when sequencing a male individual [34]. For the bonobo genome sequence, the X chromosome assembly is believed to be of comparable quality to the autosomal assembly, as the genome of a female bonobo was sequenced. Estimates of indel rates on the X chromosome are consistent with this: for the human-bonobo alignment the X chromosome has a significantly lower indel rate (0.291 errors per kb, 95% c.i. 0.248-0.326) compared to the human-chimpanzee alignment (0.445 errors per kb, 95% c.i. 0.404-0475). In contrast to this, the quality of chromosome 21 for chimpanzee is believed to be higher than the assembly average, while chromosome 21 for bonobo is considered to be in line with the assembly average. This is consistent with estimates of indel errors for chromosome 21 for alignments of human-bonobo (0.310 errors per kb, 95% c.i. (0.241-0.376) and human-chimpanzee (0.257 errors per kb, 95% c.i. 0.193-0.317). Confidence intervals for these last estimates are wide as a result of the small amount of sequence examined.

Finally, lineage specific error rates were investigated using a three-way alignment of the human, bonobo and chimpanzee assemblies (**Figure S3a.3)** created by merging pairwise alignments [33] for human-bonobo and human-chimpanzee. For each indel event, parsimony was used to infer along which branch the indel had occurred. As expected, the human genome assembly has the lowest indel error rate (0.092 errors per kb, 95% c.i. 0.086-0.096). Estimates for the bonobo (0.135 errors per kb, 95% c.i. 0.129-0.141), and chimpanzee (0.128 errors per kb, 95% c.i. 0.123-0.133) were higher than in human, with the error rate of the bonobo assembly only slightly higher than that for chimpanzee.

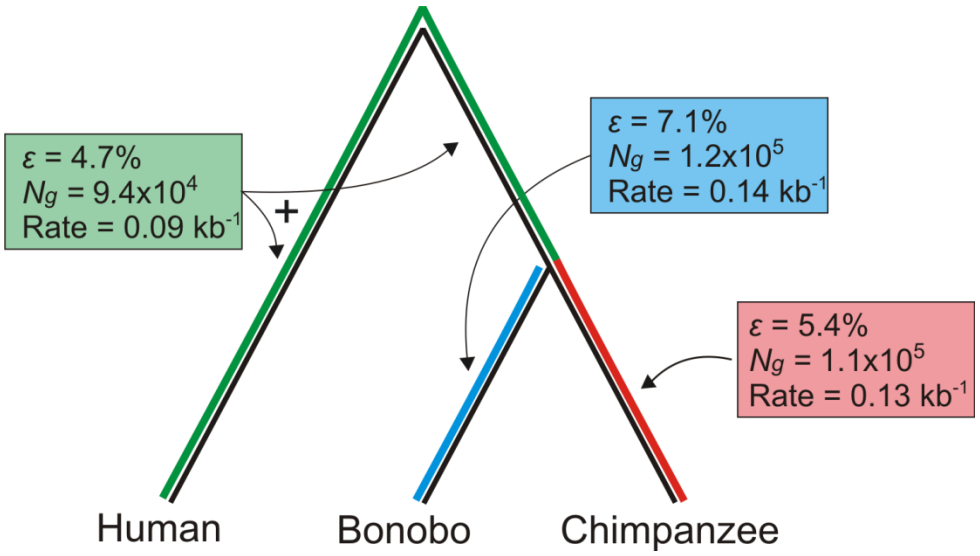


Figure S3a.3: Estimates of lineage-specific indel errors (ε = percentage of all indels that are errors, *Ng* = total number of indel errors in the alignment) in great ape genome assemblies. Using a three-way alignment of the human, bonobo and chimpanzee genome assemblies, it is possible to estimate the quantity of lineage-specific indel errors for the bonobo and chimpanzee assemblies. We infer that the remaining indel errors are present in the human genome assembly.

Supplementary Information 4

Segmental Duplication Analysis of the Bonobo Genome

Emre Karakoc[1], Can Alkan[1], Saba Sajjadian[1], Claudia Rita Catacchio[2], Mario Ventura[1,2], Tomas Marques-Bonet[1,3], Evan E. Eichler[1]*.

1-Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA

2-Sezione di Genetica-Dipartimento di Anatomia Patologica e Genetica, University of Bari, 70125 Bari, Italy

3-Institut de Biologia Evolutiva (UPF-CSIC), 08003 Barcelona, Catalonia, Spain. Institut Catala de Recerca Avancats (ICREA).

4-Howard Hughes Medical Institute, Seattle, Washington 98195, USA

* To whom correspondence should be addressed (eee@gs.washington.edu)

1-Duplications on the Bonobo Genome (QC check)

We analyzed segmental duplications (SDs) in the bonobo genome using two computational methods: 1) a heuristic for genome-wide self-alignment (whole-genome assembly comparison, WGAC [35]) and 2) the assessment of excess depth-of-coverage by whole-genome shotgun sequence detection (WSSD) on the bonobo assembly [36]. We performed all analyses on the bonobo genome assembly version i7, which contains 12,038 scaffolds and a total of 2.87 Gb (see also SI 2). After the analysis, we assigned these scaffolds to their respective chromosome using the AGP file. 139 Mbp of the reference genome, which were not represented in the AGP file, were assigned to an unknown chromosome. All scaffolds were repeat masked [37] using a common repeat library composed of both human and chimpanzee retrotransposons in addition to other low complexity sequences. The first method, WGAC, identifies pairwise alignments >1 kbp and >90% identity. We identified 49 Mbp (nonredundant base pairs or 1.7% of the whole genome) as "duplicated" between and within scaffolds. This is substantially less than what was observed in chimpanzee or human genomes (Human Consortium [23] and Chimpanzee Consortium [3, 38]) (~5% of the genome assembly). Of the 49 Mbp detected, 37 Mbp (76%) map to autosomes and sex chromosomes while the remaining 12 Mbp remain unmapped (chrUn). Based on scaffold assignment, we classify 12,733 duplications as interchromosomal and 5,332 as intrachromosomal. (Note: 2,193 of these map to the "unknown" chromosome (41% of the intrachromosomal))

Next, we assessed the SD content in the bonobo assembly (version i7) by measuring excess of read-depth against the bonobo genome (WSSD). This method identifies SDs >10 kbp in length and >94% identity. To make the data comparable with other human datasets generated from

Illumina [39], we fragmented 143 million 454-FLX generated sequences (~27X coverage) from the reference bonobo sample ("Ulindi") into 2.2 billion short (36 bp) reads. We mapped reads using the mrFAST [39] aligner, which places reads to all possible locations in the reference genome within a given edit distance of 2 (>94% sequence identity). We predicted a total of 39.1 Mbp of duplicated sequences (>94%; >20 kbp and 70.5 Mbp at >10 kbp), which is substantially less than what was previously reported for the chimpanzee genome (70.59 Mbp, >94%; >20 kbp [38]), suggesting a possible collapse of highly identical duplications in the assembly. 65.2 Mbp map to chromosomes whereas 5.3 Mbp remain unassigned. We compared WGAC and WSSD estimates by focusing on SDs >10 kbp and >94%sequence identity (**Table S4.1**). Only 14.6 Mbp of duplications were common toboth (6.96 Mbp > 10Kbps) which would be our conservative true positive(ancient duplications, WGAC 90-94%, are not included in the intersection),while 58.64 Mbp (> 10 Kbp) were predicted by WSSD and 3.46 Mbp (> 10 Kbp) were predicted by WGAC methods alone.As expected from *de novo* sequence assembly with next-generation data, an overwhelming majority of highly identical SDs were collapsed. In this case, as the assembly is done using the 454 reads, the assembly was created with relatively low thresholds for sequence identity to compensate for the higher error rate than Sanger sequencing reads. This is supported by the observations that most high-identity duplications (>94% identity) are missing from the WGAC pairwise predictions (**Figure S4.1**).

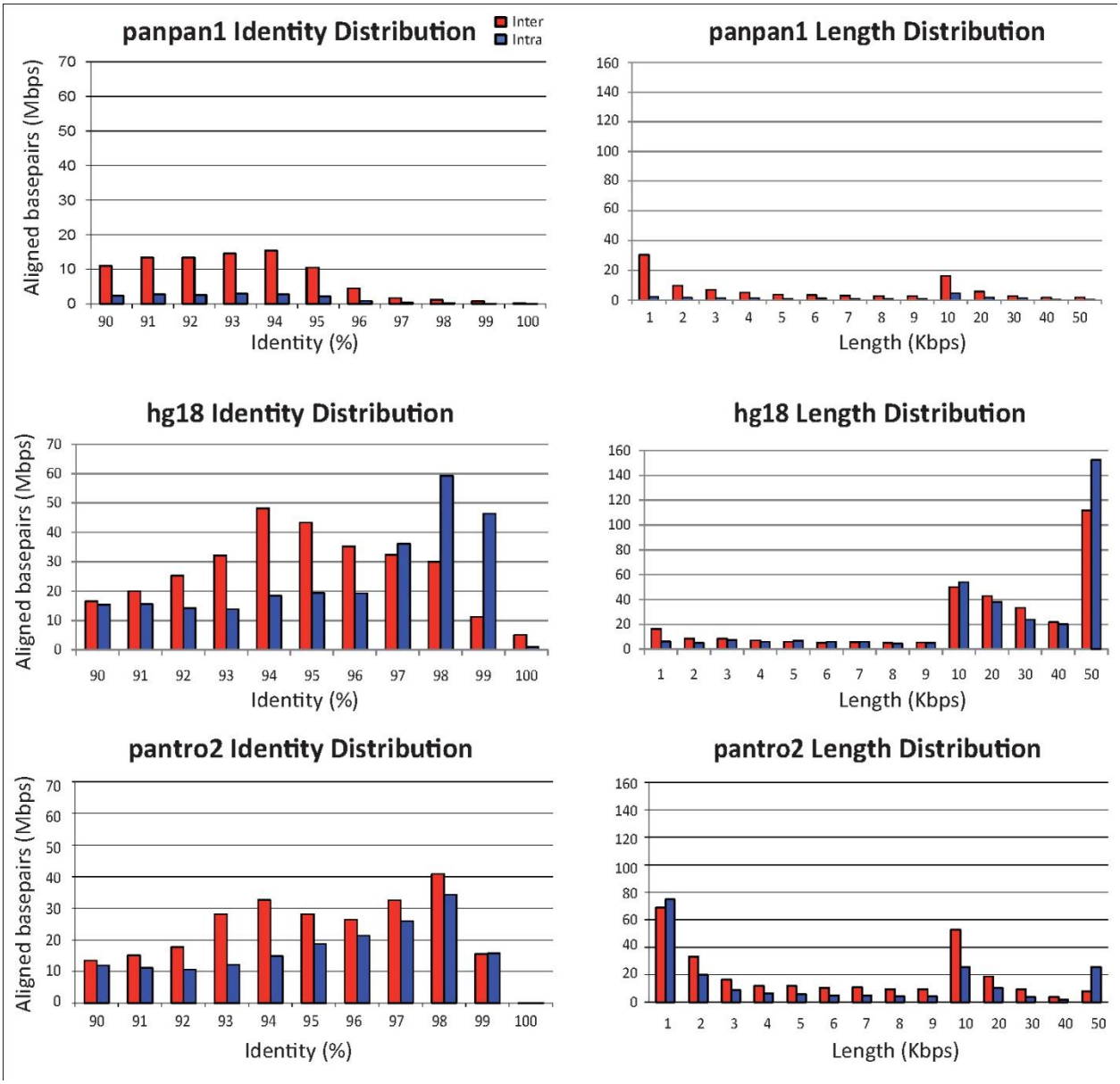


Figure S4.1. Pairwise identity and length distributions of segmental duplications in the bonobo assembly (panpan1 i7). Note the reduced number of duplications with sequence identity >95% in comparison to other assemblies.

Table S4.1. Summary statistics of duplication analysis (>10 kbp) on the bonobo assembly (scaffolds) using both methods, WGAC and WSSD, for detecting duplications.

identity	WGAC	WSSD	shared	WGAConly	WSSDonly
>=90%	24,344,827	65,615,058	14,615,625	9,729,202	50,999,433
>=94%	10,433,949	65,615,058	6,969,782	3,464,167	58,645,276

WSSD =whole-genome shotgun sequence detection [36], a method of detecting duplications based on excess of depth-of-coverage. WGAC =whole-genome assembly comparison [35]and compares the assembly against itself.

2-Comparative SD Analyses

In order to compare the duplication pattern between humans and great apes, we created a second read-depth duplication map by mapping bonobo reads to the human reference genome. This provided a common set of coordinates for comparing to other great ape duplication maps [40]. As described above, we "Illuminized" 143 million of 454-FLX generated sequences mapping the derived sequence reads (36 bp) to the human genome reference assembly (hg17, Build35) based on the following criteria (94% sequence identity and 5 kbp of non-repeat masked) [39] (**Table S4.2**). For this comparative analysis, we focused on regions >20 kbp in length and defined SDs using a threshold of 4 standard deviations beyond the autosomal mean (within well-known single-copy control regions). We considered regions only if the sequence consists of less than 80% of common repeats [37, 41].

Table S4.2. *Summary statistics of bonobo read mapping on human assembly (hg17).*

454 reads	# basepairs	Illuminized Reads	Mapped to hg17	Avg DoC	StdDev DoC
143,137,950	81,958,072,022	2,201,761,240	501,537,341	1921.81	305.43

Excluding sex and random chromosomes, we predict 77.2 Mbp (>20 kbp) of SD content in the bonobo genome. We repeated the same procedure described above using whole-genome shotgun sequence data generated for the chimpanzee [3]. Comparing the common chimpanzee and bonobo, we find remarkably similar duplication content (76.5 Mbp, >20 kbp). We note that this estimate is slightly larger than previously reported due to methodological differences in the use of next-generation sequence data as opposed to capillary sequence data [40] (chimpanzee 65 Mbp, SDs >20 kbp). Using these read-depth methods, we find that 66.3 Mbp are shared by both common chimpanzee and bonobo, 5.2 Mbp are specific to chimpanzee, and 4.9 Mbp are specific to bonobo (**Figure S4.2**). It is important to note that this analysis is based on the data generated from a single individual for each species (*Ulindi* and *Clint*) and does not take copy number polymorphism among the species into account, which is likely to be substantial.

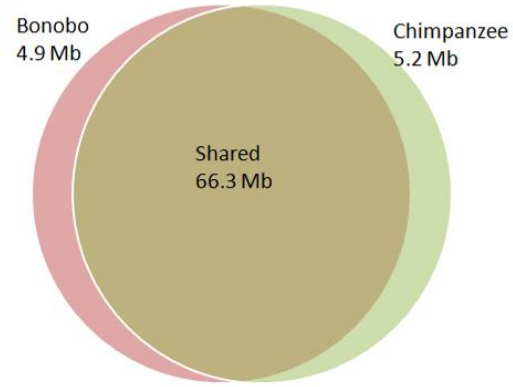


Figure S4.2. Comparison of chimpanzee and bonobo SD content. SDs (>20 kbp) predicted by

regions of excess read-depth to the human reference genome (hg17, Build35).

We next compared the extent of overlap with human using a single Yoruban African genome (*NA18507*) to represent the human genome [39, 42] (**Figure S4.3; Table S4.3**). We applied copy number correction to compensate the bias associated to mapping to the human assembly. In short, we used the depth-of-coverage as a surrogate of copy number, allowing us to recalculate the amount of specific duplications [39].

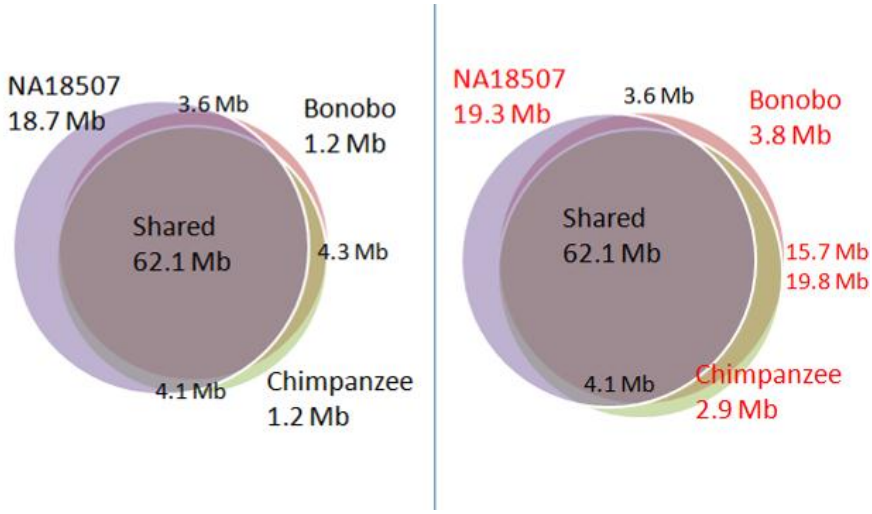


Figure S4.3. Comparison of human, chimpanzee, and bonobo SD content. Read-depth estimation of large SDs (>20 kbp) based on mapping to human reference: before (left) and after (right) lineage-specific copy number correction to account for the human bias.

Table S4.3. Duplication in NA18507 (HSA), Clint (PTR) and Ulindi (PPA). Lineage-specific SDs are copy number corrected to account for the human bias in the mapping.

Species	>20 kbp	Copy number corrected	ALL
HSA	18,691,233	19,330,883	25,656,268
HSA_PPA	3,584,367		7,531,879
HSA_PTR	3,960,229		5,919,765
HSA_PTR_PPA	61,455,014		59,254,577
PPA	1,224,698	3,801,687	4,415,018
PTR	1,226,463	2,936,510	4,424,359
PTR_PPA	4,268,176	15,747,639/19,759,322	4,989,433
Grand Total	**94,410,180**		**112,191,299**

3-Experimental Validation

We validated and reclassified the duplication maps of human, chimpanzee, and bonobo by using array comparative genomic hybridization (CGH). Two custom microarrays were implemented. The first set of experiments (Ape chip) was based on a customized oligonucleotide microarray (NimbleGen, 385,000 isothermal probes) targeted specifically to the great ape SDs ([40], GEO13934). The second microarray was designed to specifically target new regions detected in this study (Bonobo chip).

Table S4.4. *Summary of arrayCGH experiments performed (HSA, Homo sapiens; PPA, Pan paniscus; PTR, Pan troglodytes).*

HSA(NA15510)/PPA(ULINDI) => Bonobo chip[1]
PTR(CLINT)/PPA(ULINDI) => Bonobo chip[1]
HSA(NA15510)/PPA(LB502) => Ape chip[2] (Marques-Bonet, 2009)
HSA(NA15510)/PPA(LB501) => Ape chip[2]
PTR(CLINT)/PPA(LB502) => Ape Chip[2]
HSA(NA15510)/PTR(CLINT) => Ape Chip[2]

[1] *Design with sites > 10Kbps only;*[2]*Bonobo-specific sites not covered.*

We performed interspecific hybridization experiments based on one chimpanzee (*Clint*), one human (*NA15510*), and three bonobos (*Ulindi*, *LB501*, and *LB502*). All experiments were performed in replicate where test and reference labels were swapped. Log2 relative hybridization intensity was calculated for each probe. We restricted our analysis to regions greater than 20 kbp in length with at least 20 probes [40]. After normalization, copy number variable regions were detected based on an HMM segmentation [43] of the probe log2 intensity. CNV calls (centered at 1 standard deviation) were compared against the original interval query set. An interval was considered as validated when there was more than one-third overlap with the HMM calls, or the median log2 of the region was beyond 1 standard deviation of the hybridization (log2 = ~0.3). Since multiple individuals were used in this study, we reported as duplication those computational predictions where the event was confirmed in at least one individual from each species.

Using this approach, we validated 249 of the 338 human-specific SD intervals (14.6 Mbp), 11/29 chimpanzee (371 kbp), 16/28 bonobo predictions (264 kbp), and 55/56 *Pan*-shared duplications (5.1 Mbp). According to the new array CGH results, we reclassified our species categories (**Table S4.5;Table S4.6; Figure S4.4**).

Table S4.5. Validated duplication in NA18507 (HSA), Clint (PTR) and Ulindi (PPA). Lineage-specific SDs are copy number corrected to account for human bias (in case of shared duplications, both corrections based on each species are applied).

Species	Validated	Copy number corrected
HSA	14,650,923	14,580,753
HSA_PPA	458,530	
HSA_PTR	800,945	
HSA_PTR_PPA	68,227,013	
PPA	264,263	703,886
PTR	371,105	883,275
PTR_PPA	5,151,688	17,263,592/22,317,789

Table S4.6. validated regions of bonobo specific duplications (Ulindi).

Chr	Start	End	Description NGenes Completed	Description NGenes Partial
chr1	193.575.000	193.648.069	NM_006684-NP_006675-CFHR4-complement factor H-related 4|	NM_005666-NP_005657-CFHR2-H factor (complement)-like 3|
chr16	66.592.000	66.618.397	NM_018380-NP_060850-DDX28-DEAD (Asp-Glu-Ala-Asp) box polypeptide 28|	NM_017803-NP_060273-DUS2L-dihydrouridine synthase 2-like, SMM1 homolog|
chr17	21.349.319	21.390.000		NM_001113434-NP_001106905-C17orf51-hypothetical protein LOC339263|
chr17	21.856.819	21.878.000		
chr5	17.449.000	17.531.000		
chr8	11.837.065	11.858.000		
chr9	138.250.000	138.273.000		NM_000718-NP_000709-CACNA1B-calcium channel, voltage-dependent, N type,|

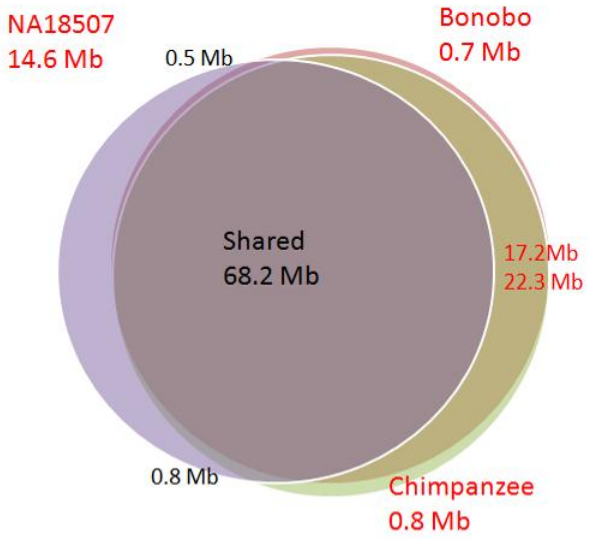


Figure S4.4. Validated human, chimpanzee, and bonobo SDs (>20 kbp) using WGS depth-of-coverage in the human assembly (hg17). For shared SDs, the copy number is adjusted separately for the chimpanzee (17.2 Mbp) and the bonobo (22.3 Mbp) copy.

To eliminate recurrent duplications in the primate lineage and ensure specificity in the lineage-specific events, we also compared our results to duplication maps established for macaque, orangutan, and gorilla [40]. This further refined the set to 371 kbp of bonobo-specific SDs and 225 kbp of chimpanzee-specific duplications. Two complete genes (*CFHR4* and *DDX28*) and three partial genes (*CFHR2*, *DUS2L*, and *CACNA1B*) overlap with bonobo-specific SDs. Only one complete gene (*HLA-DRB1*) and three partial genes (*SLC24A4*, *BMPR2*, and *HLA-DRB5*) overlap with chimpanzee-specific SDs.

If we focus on SDs that validate with either increased copy in chimpanzee or bonobo (irrespectively of lineage specificity), we identify 179 sites (16.5 Mbp) where bonobo has more copies than chimpanzee and 165 sites (15.8 Mb) where chimpanzee has increased in copy with respect to bonobo (see **Supplementary Table S4.6** for gene list). This includes a specific cluster of beta-defensin genes (**Figure S4.5**), which are duplicated in both chimpanzee and bonobo (not in human). Bonobo shows increased copy of this gene family (n = 8) when compared to chimpanzee (n = 5 copies). However, this difference between Clint and Ulindi is due to individual differences rather than fixed species difference (**Table S4.6; Figure S4.6**).

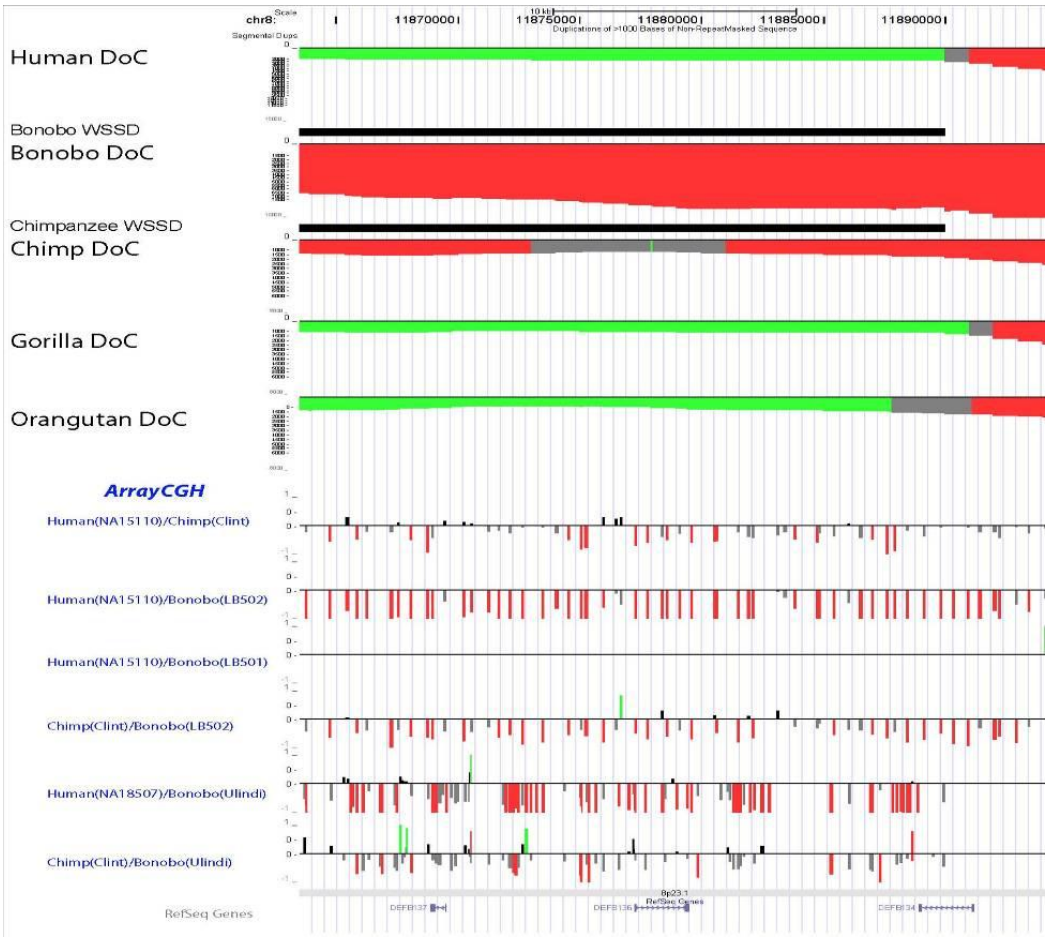


Figure S4.5. Pan-shared duplication on the beta-defensin cluster. We estimated that the bonobo genome has more copies than chimpanzee (eight and five, respectively).

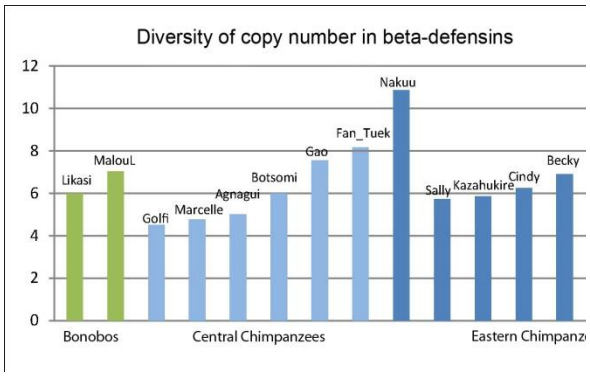


Figure S4.6. Individual copy number in the beta-defensin cluster. The median and range of each individual copy number shows no fixed differences between chimpanzee and bonobo.

4-Segmental Duplications and ILS

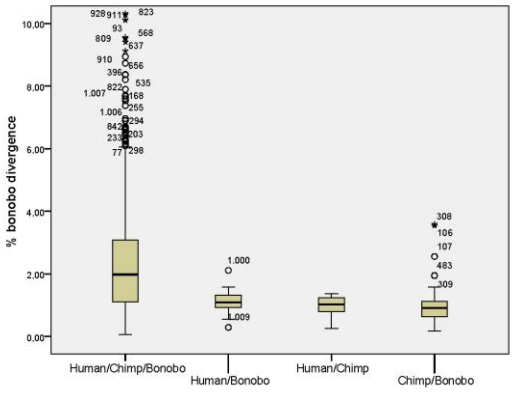
The methods to detect CNVs based on depth of coverage have a strong limitation in terms of the location of paralogous copies. They can add little about the structural conformation or paralagous variants of the region. Previous analyses of duplications that do not follow the consensus phylogeny suggested that coinciding independent events are more often the explanation. The expectation is ~10% between human and chimp with an upper bound of 30% (Marques-Bonet et al. 2009).

In this type of analyses, divergence is only used as a surrogate of experimental validation when this is not possible (Neandertal genome project). In this case, we evaluate again the accumulated divergence between human and bonobo of the regions with excess of depth of coverage (single nucleotide variants, supported by at least 2 reads, not-repeat masked with basepair quality>=20) and we found that almost all potential ILS SDs are indeed really young (lower divergence, similar to single copy bonobo regions) suggesting potential recurrent events rather than ILS. Similarly, 21% of the older shared SDs (shared by human/bonobo/chimp) have similar cumulative lower divergence (< 1%) which fits well with the expected rate of recurrent events/gene conversion.

Table4.7 and Figure4.7. Percentage of cumulative bonobo divergence in regions duplicated in bonobo, chimpanzee and human.

Class of SD	N	Mean divergence (%)	Std Dev
Human/bonobo	16	1,13	0,43
Chimp/bonobo	76	0,98	0,57
Human/chimp*	20	1,00	0,30
Human/chimp/bonobo	908	2,48	2,37

** Notice that bonobo is single copy in this category*



Acknowledgements

Thanks to Arcadi Navarro, Elodie Gazave, and Carl Baker for performing the array CGH hybridizations. This work was supported by a Ramón y Cajal (MICINN-RYC 2010) and ERC Starting Grant (StG_20091118) [to T.M.-B.], and by the National Institutes of Health [HG002385 to E.E.E.]. E.E.E. is an investigator of the Howard Hughes Medical Institute.

Supplementary Information 5

Additional Chimpanzee and Bonobo Sequencing and Initial Processing

Anne Fischer[1,2], Kay Prüfer[1,*], Janet Kelso[1], Susan Ptak[1] and Svante Pääbo[1,*]

1. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
2. International Center for Insect Physiology and Ecology, Nairobi, Kenya

* To whom correspondence should be addressed (paabo@eva.mpg.de, pruefer@eva.mpg.de)

Additional to the data of the bonobo genome, we sequenced a total of 19 individuals: three bonobos, two western chimpanzees, seven eastern chimpanzees and seven central chimpanzees. All individuals originate from the wild and all but one individual were sampled in bonobo and chimpanzee sanctuaries.

Sequences were generated by Illumina sequencing. All sequenced individuals have also been deeply sequenced by 454 sequencing for autosomal regions summing to 150 kilo bases [44]. We use the 454 data to estimate error rates in the illumina sequences.

Sampling of Individuals

The blood samples from bonobos, central and eastern chimpanzees and bonobos were collected by Michel Halbwax and Anne Fischer in 2005, 2007 and 2009 during regular health checks. The lymphocytes were extracted from blood samples using a Ficoll gradient and frozen.

Most of these apes were confiscated by various officials from individuals selling these animals for trade or who kept them as pets, and then were brought to the sanctuaries. Samples from eastern chimpanzees (*Pan troglodytes schweinfurthii*) were collected in the Chimpanzee Sanctuary and Wildlife Conservation Trust (CSWCT), Ngamba Island, Entebbe, Uganda, central chimpanzees (*Pan troglodytes troglodytes*) in Tchimpounga chimpanzee rehabilitation center (TCRC), Pointe-Noire, Republic of Congo, bonobos (*Pan paniscus*) in Lola ya bonobo sanctuary, Kinshasa, Democratic Republic of Congo and western chimpanzees (*Pan troglodytes verus*) from Sierra Leone (one from Tacugama chimpanzee sanctuary, Freetown, one currently at the BPRC in Netherlands but wild born in Sierra Leone).

Id	Individual	Species/Subspecies	Sex	Sampled at
B1	Likasi	Bonobo	F	Lola Ya Bonobo
B2	MalouL	Bonobo	F	Lola Ya Bonobo
B3	Lodja	Bonobo	F	Lola Ya Bonobo

CC1	FanTuek	Central Chimpanzee	F	TCRC
CC2	Marcelle	Central Chimpanzee	F	TCRC
CC3	Agnagui	Central Chimpanzee	F	TCRC
CC4	Gao	Central Chimpanzee	F	TCRC
CC5	Botsomi	Central Chimpanzee	F	TCRC
CC6	Golfi	Central Chimpanzee	F	TCRC
CC7	Bayokele	Central Chimpanzee	F	TCRC
EC1	Nakuu	Eastern Chimpanzee	F	CSWCT
EC2	Sally	Eastern Chimpanzee	F	CSWCT
EC3	Becky	Eastern Chimpanzee	F	CSWCT
EC4	Kidogo	Eastern Chimpanzee	F	CSWCT
EC5	Cindy	Eastern Chimpanzee	F	CSWCT
EC6	Kazakuhire	Eastern Chimpanzee	F	CSWCT
EC7	Katie	Eastern Chimpanzee	F	CSWCT
WC1	Small Lucie	Western Chimpanzee	F	Tacugama
WC2	Louise	Western Chimpanzee	F	BPRC

Table S5.1: Origin of Sequenced Individuals.

Library Preparation, Sequencing and Basecalling

DNA was extracted from 50ml cell cultures (5-50 x10^6 cells) obtained from lymphocytes transformed with Epstein-Barr virus using the Gentra-puregene kit from QIAGEN and following manufacturer's instructions. Indexed Illumina libraries were prepared from the extracted DNA, giving each individual a unique index. All libraries were sequenced on the Illumina Genome Analyzer II as either 76 cycle or 101 cycle paired end run with additional cycles to read the index. A low percentage of indexed phiX library was sequenced alongside the libraries on all lanes. The index sequence for the spiked-in phiX sequences was "TTGCCGC". Index sequences for the individuals are shown in Table 5.2.

The *Ibis* program [45] was used for basecalling. *Ibis* was run for each lane separately and trained with the phiX reads of this lane. After base calling, reads were filtered for the correct index read. Paired reads were merged when both reads shared significant similarity (at least 11 basepairs overlap, 90% identity). The base with the highest quality is chosen as consensus base in overlapping parts and quality scores are recalculated as sum of both quality scores when the bases in both reads match and the maximum quality score when bases differ.

The resulting files of merged and paired-end reads were used as input for mapping. Mapping results show that around one fold genome coverage was reached per individual. Table S5.2 gives an overview of the data acquired.

Illumina Sequencing can give different error profiles between runs and lanes. When possible, individuals were sequenced on the same run. In two cases, we sequenced mixed libraries over lanes, ensuring identical error profile for this data: One lane was sequenced with a mixed library of individuals

B1 and B2, and the individuals WC1, WC2, B3, EC7 and CC7 were sequenced as a mixed library over 3 lanes. Table S5.3 summarizes the lanes sequenced for this project.

Id	Individual	Index	Read length	Mapped reads (merged) in 10^6	Mapped forward + reverse reads in 10^6	Bases in MQ30 reads
B1	Likasi	AACCGCC	76	0.3	51.0	2.7G
B2	MalouL	AACGAAC	76	1.8	89.2	5.0G
B3	Lodja	ATGGTAT	101	9.2	24.6	2.7G
CC1	FanTuek	ACGACCT	101	13.4	31.0	3.7G
CC2	Marcelle	ACGGAGG	76	4.7	52.3	3.2G
CC3	Agnagui	ACCTCAT	101	13.6	29.6	3.7G
CC4	Gao	ACGATTC	101	11.5	35.1	3.7G
CC5	Botsomi	ACCTTGC	101	12.8	24.3	3.2G
CC6	Golfi	ACGCGGC	101	11.5	32.1	3.5G
CC7	Bayokele	GCCAAT	101	6.8	20.6	2.2G
EC1	Nakuu	GAGTGG	76	2.1	45.9	2.8G
EC2	Sally	ATAGAAG	76	2.0	52.1	3.1G
EC3	Becky	ACGTAAC	101	11.9	33.3	3.7G
EC4	Kidogo	ACTCGTT	101	12.2	28.1	3.4G
EC5	Cindy	TCTACC	101	4.9	43.0	3.4G
EC6	Kazakuhire	ACTACTG	101	14.5	30.4	3.7G
EC7	Katie	AGCGCTG	101	6.0	26.3	2.4G
WC1	Small Lucie	GGTAGC	101	8.9	28.0	2.9G
WC2	Louise	CAACCGG	101	7.4	19.7	2.1G

Table S5.2: Total sequencing data and read length for chimpanzee and bonobo individuals. Total bases and mapped reads are determined after mapping to the hg18 genome using BWA [46] (Version: 0.5.7). All bases in reads mapping with quality score of at least 30 are counted for bases in reads.

Run	#Lanes	Individuals
091105	1	B1, B2
100119	1	B1
100119	2	B2
100519	1	CC2
100426	1	EC1
100217	1	EC2
100322	1	EC5
100506	1	CC1
100506	1	CC3
100506	1	CC4
100506	1	CC5
100506	1	CC6
100506	1	EC3
100506	1	EC4
100506	1	EC6
101015	3	B3, CC7, EC7, WC1, WC2

Table S5.3: Assignment of sequencing data to Illumina Sequencing runs and lanes.

Data Filtering and Alignment

We aligned all Illumina sequencing data using BWA [46] (Version: 0.5.7; default parameters) to the human genome (hg18). Only reads with a mapping quality of at least 30 were considered for further analysis. We then filtered the sequences using the following criteria:

- Bases below a base quality of 30 were filtered.
- Bases that are within 5 base pairs of an insertion/deletion difference to the reference were filtered
- Bases were filtered if they are within 5 basepairs of two or more base differences.

Comparison with 454 Sequencing Data and Sequencing Error Estimates

454 sequencing data is available for 15 regions summing to ca. 150 kilo bases for 16 out of 19 Illumina-sequenced individuals. The 15 regions were amplified by long-range PCR and sequenced on the GS FLX sequencing platform to a 30x average coverage for each individual. The regions have an average GC content of 38% (in humans, 34% in the chimpanzee genome) and were sampled to be mostly neutrally evolving, but are on average low in repeats compared to the genome average (see [44]).

We used this deep 454 sequencing dataset to estimate the Illumina sequencing error rate for individuals. For this, we first mapped the 454 sequencing data to the human genome (version hg18) using BWA [18] (version 0.5.7; default parameters). The resulting alignments were filtered as described in SI 2 for the 454 genome data. In order to estimate the error in the Illumina sequencing data, we compared each filtered Illumina read (as described in the previous section) to the 454 data at positions where 454 reads exceed a coverage of 20x. A base in the filtered Illumina read was then counted as a potential sequencing error when no 454 read showed this variant.

The resulting error estimates are summarized in Table 5.4 and indicate a generally low sequencing error after filtering, with a point estimate of less than 1 error in 1000 base pairs for all individuals. Figure 5.1 shows the error estimates for each individual. The confidence intervals for the error estimates are overlapping for many individuals. Given that Illumina reads tend to accumulate errors towards the ends of reads, one may expect a consistent difference between sequence data of different cycle number. However, we observe no consistent effect of cycle number; some 76 cycle sequenced individuals exhibit a higher error rate of 101 cycle sequenced individuals. The variance of sequence errors over different Illumina runs thus exceeds the influence of read-length.

The error rate estimates based on the comparison with 454 data makes the assumption that the 454 data depicts the true sequence and that the 150 kilo bases are representative for the genome. However, several other factors not taken into account in the comparison with 454 sequencing data may lead to an underestimate of the error rate.

First, error rates may be underestimated due to coinciding errors in 454 and Illumina sequences.

In order to correct for this problem, we regarded any Illumina basecall as erroneous when not at least five 454 reads supported the allele. Figure 5.1 shows that the more restrictive criterium changes the results only marginally.

Second, both 454 and Illumina data were mapped using the BWA algorithm. Similar to coinciding sequencing errors, a consistent mapping error or bias affecting both mapped datasets would be masked in our comparison. However, our 454 data has been prepared from a PCR product and can thus only cause alignment errors within the amplified parts of the genome. The shotgun data from Illumina, on the other hand, is sampled from the entire genome. Thus, the Illumina data is expected to be affected much stronger by any potential alignment problems and the masked error from alignment errors within the 150 kilo bases can be safely ignored.

Third, we only use a subset of the entire genome sequence for error estimation. Differences in the features of the 150 kilo bases to the genome average may thus limit our ability to accurately estimate sequence error genome wide. Some of the features have been picked carefully to match genome average. However, the 150 kilo base regions are depleted in repetitive elements. This may lead to an underestimate of sequencing error when repetitive regions are for instance enriched in false alignments. The error estimates should thus be rather interpreted as representative for the non-repetitive fraction of the genome.

In summary, our results indicate that our filtering procedures lead to highly accurate sequences. However, this result may only hold for regions that are of low repeat content.

Individual/Lane	Bases compared	Bases different
B1 mixed	4273	2
B2 mixed	4230	0
B1 separate	18908	13
B2 separate	41632	21
B3	56114	19
CC1	72983	6
CC2	46150	6
CC3	86400	10
CC4	70175	7
CC5	71397	15
CC6	49484	3
EC1	16759	3
EC2	31102	15
EC3	55600	15
EC4	68939	7
EC5	59008	4
EC6	78629	5
WC2	41332	7

Table S5.4: Comparison of Illumina to 454 sequence data. Data for B1 and B2 is present twice since one lane was sequenced with both libraries mixed, while the remaining data was sequenced on separate lanes.

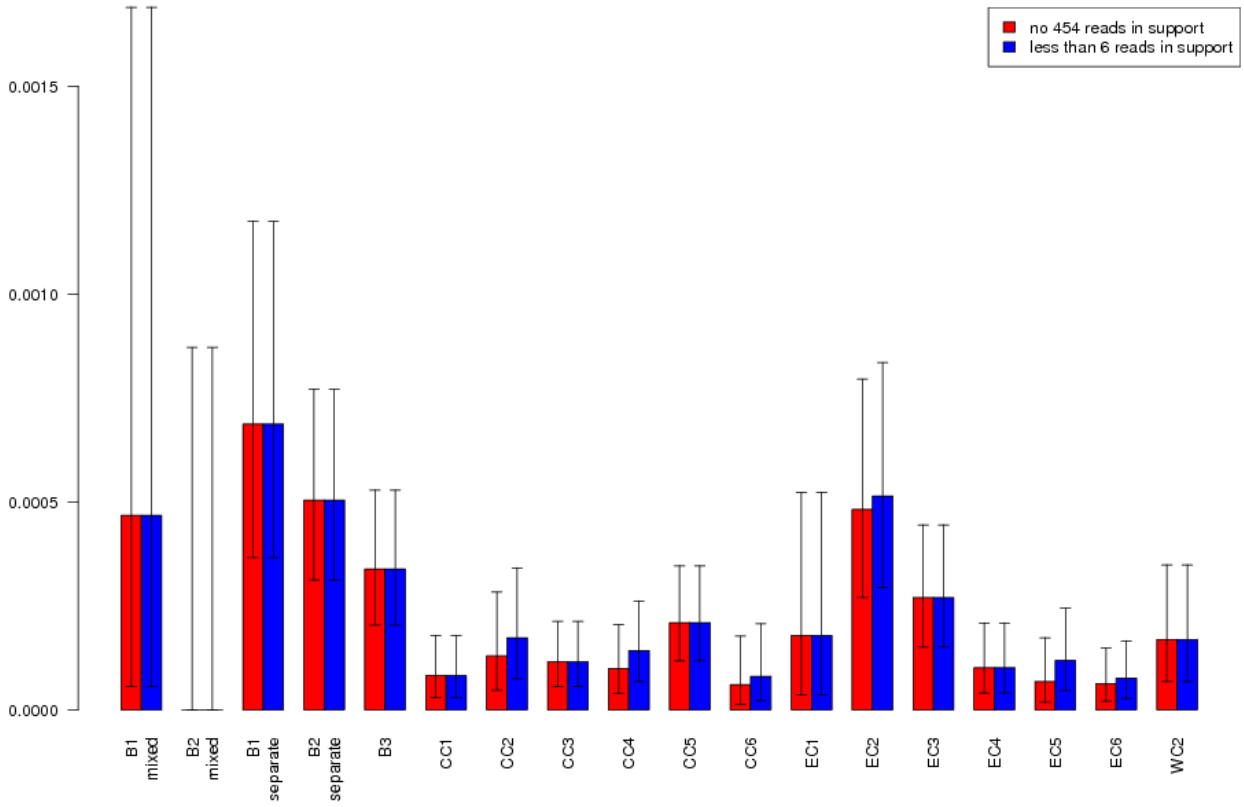


Figure S5.1: Comparison of Illumina to 454 sequence data. Red bars represent error-estimates requiring one 454 read to differ; blue bars require five 454 reads to differ. Error bars are calculated as 95% confidence intervals of a binomial distribution.

Mitochondrial Genome Sequences

In order to test for potential sample mix up, we collect all sequences aligning to the mitochondrial sequence (chrM from panTro2 release for chimpanzee, Ulindi mt-genome for bonobos) using BWA. A simple mapping consensus is called for each individual using the samtools pileup option and the consensi of all individuals are aligned with *muscle* [17]. Figure S5.2 shows the maximum likelhihood tree calculated with phyML on all Illumina chimpanzee and bonobo consensi, the panTro2 reference mitochondrial genome (Clint), the bonobo reference mitochondria (Ulindi) and four additional full mitochondrial genomes for eastern, central, nigerian-cameroonian and western chimpanzee [47] from GenBank. As expected, the sequences of Illumina-sequenced chimpanzee individuals group with the previously published chimpanzee sequences according to geographical assignment and bonobo individuals form a clade separate from all chimpanzees.

Sequence Read Archive Accessions and Data Availability

All sequences have been made available through the Sequence Read Archive under study id ERP000602 (http://www.ebi.ac.uk/ena/data/view/ERP000602).

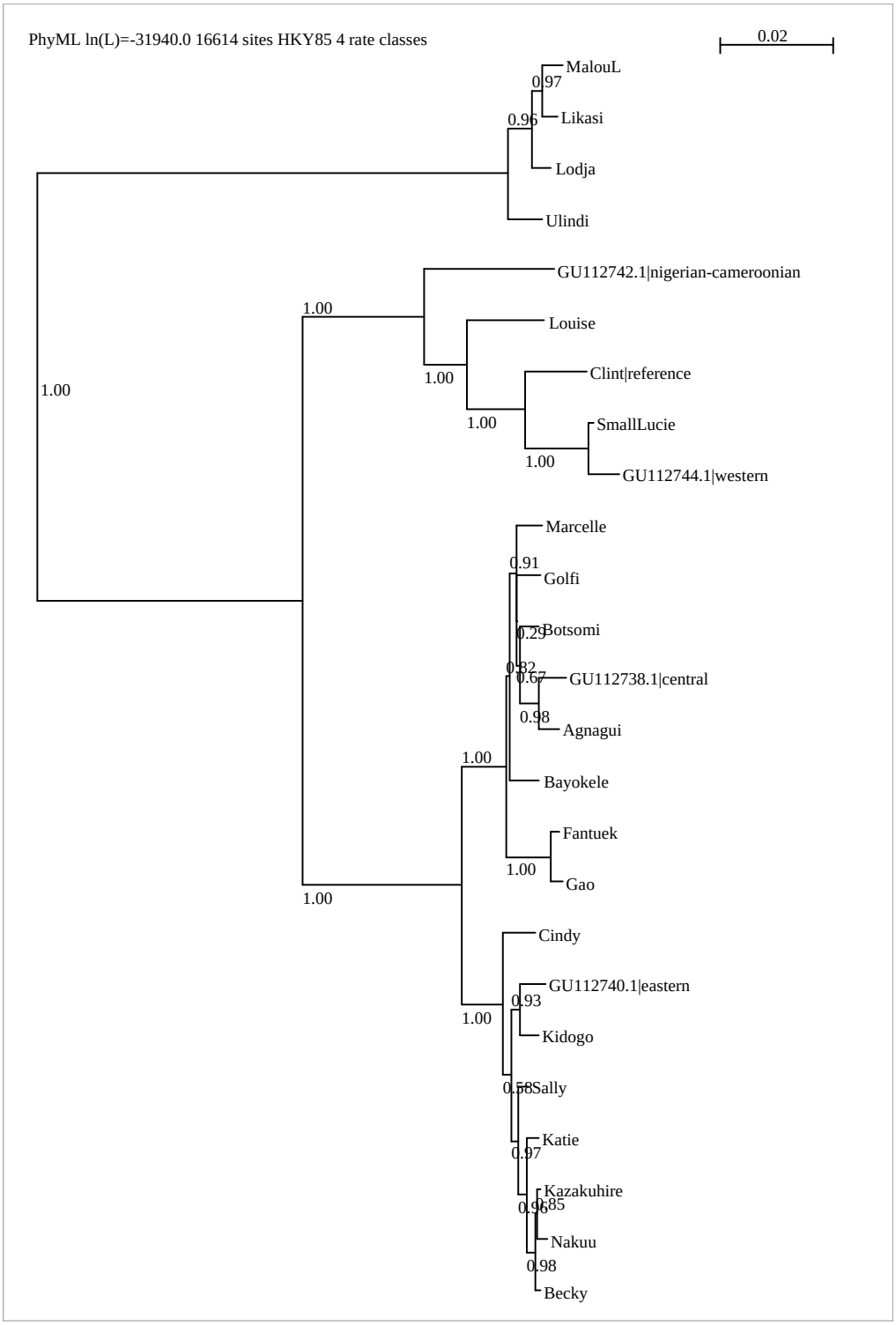


Figure S5.2: Tree of chimpanzee and bonobo mitochondrial sequences.

Supplementary Information 6

Retrotransposon Evolution in the Bonobo Genome

Keiko Akagi[1], Saneyuki Higashino[2], David E. Symer[1,3,4]*

[1]Human Cancer Genetics Program and Department of Molecular Virology, Immunology and Medical Genetics, [3]Department of Internal Medicine and [4]Department of Biomedical Informatics, The Ohio State University Comprehensive Cancer Center, Columbus, OH 43210; [2]Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, Kanagawa, Japan

* To whom correspondence should be addressed (david.symer@osumc.edu)

Transposable elements comprise almost half of the human and other mammalian genomes[48]. Ongoing mobilization of certain families of retrotransposons has caused extensive genomic variation within and between primate species[49], with many possible functional implications. Here we mapped and comparatively analyzed retrotransposon integrants present in the human, chimpanzee, and bonobo genomes. More than 2.5 million integrants occur at orthologous loci in the three species, consistent with the close, common ancestry of the genomes as they are identical by descent[50]. We also found over 1,500 integrants that are uniquely present in the bonobo genome, most of which are L1, *Alu* and SVA retrotransposons. In addition, we found and experimentally verified numerous transposon insertions that were best explained by incomplete lineage sorting[51]. The rate of *Alu* retrotransposition increased in all three primates after their speciation, with *Alu* retrotransposons accumulating twice as frequently in human as in either bonobo or chimpanzee. By contrast, the L1 retrotransposition rate stayed fairly constant throughout the ancestral and modern lineages, while SVA retrotransposons recently have begun to accumulate in the modern species. Using Alucode[52], we identified two previously unreported *Alu* subfamilies that have accumulated predominantly in chimpanzee and bonobo, indicating that the retrotransposons themselves continue to evolve. In assessing the functional impacts of transposition, we noted that recent L1 integrants are enriched in and around particular genes with ontology terms such as “neuronal activities”, but are relatively depleted from other essential genes with “nucleic acid metabolism” activities. By contrast, we observed no preference in ontology terms for *Alu* integrants. These findings emphasize both the close, common ancestry of bonobos, chimpanzees and humans, and the substantial, ongoing contribution by retrotransposons to primate evolution.

1 - Identification of Transposon Integrants in Primates

Detection of transposon integrants

We used trace sequence alignments to identify the present/absent status of transposable elements inserted in the bonobo, chimpanzee and human genomes[53]. We used GMAP[54] to align all available bonobo (198 million reads) and chimpanzee (46 million reads) sequence traces[55], both single-end and paired-end, to the reference human genome (UCSC genome browser hg18). Alignment outputs were captured in a relational database containing human reference coordinates, percent identity and percent coverage.

To detect candidate integrant variants, we binned aligned reads into two groups, i.e. well-aligned reads and non-well-aligned reads[56, 57]. By "well-aligned", we mean > 90% coverage and > 95% identity across the entire sequence trace. The "non-well-aligned" traces were further divided into two categories based on their alignment signatures. In the first case, we identified transposon integrants that are absent from the bonobo genome or the chimpanzee genome, but present in the human reference assembly. In these cases, the bonobo or chimpanzee sequence traces were aligned into two fragments skipping over the human repetitive element entirely, indicating their absence from the ape genome (Figure S6-1A). Aligned (anchoring) sequences were required to include > 50 nt of coverage on either side of such a human repetitive element, and the candidate integrant lengths were required to be > 100 nt. These candidate human-specific integrants were classified using RepeatMasker[58] (i.e. using option: -species=primates).

Transposon variants absent from the human genome but present in either chimpanzee or bonobo were identified from "non-well-aligned" sequence traces aligned only partially to the human reference genome due to inserted transposon sequences (anchored fragment size > 50 nt, unaligned fragment size > 50 nt) (Figure S6-1B). Such sequence traces were extracted and examined for their content of repetitive elements by RepeatMasker (option: -species = primates). For traces that contain transposon sequences in their fragmented regions, we removed such sequences and re-aligned the remainders against the human reference genome using BLAT[1] (90% coverage, 95% identity) to confirm their unique alignments. Finally, alignment coordinates of the "non-well-aligned" traces were clustered[53] into loci containing two or more supporting traces without contradicting traces. Resulting loci were categorized into lineage groups (e.g. bonobo-specific, etc.) based on the trace coverage of the coordinates defined by the human reference genome.

Comparison of transposon integrants in bonobo, chimp and human genomes

Total counts of transposon integrants of various families identified in each lineage are presented in Table S6-1. As expected, the vast majority (99.7%) of transposable elements present in the human genome are also present in the bonobo and chimpanzee genomes. Because transposon integrants have accumulated over evolutionary time, and given the extremely low likelihood that precise structures of integration could occur independently, these results strongly corroborate the close common ancestry of these primates; the shared transposons are identical by descent[50].

We also identified 6,641 human-specific, 1,590 bonobo-specific, 1,079 chimpanzee-specific, and 1,207 bonobo/chimpanzee-shared transposon integrants (Table S6-1), demonstrating that retrotransposition has continued in each lineage after speciation. The number of human-specific transposon integrants is more than two times more than the combined number of bonobo-specific and bonobo/chimpanzee-shared integrants. This increase in human-specific integrants can be attributed particularly to accumulation of *Alu* retrotransposons[59]. While the number of human-specific *Alu* integrants is four times more than the combined number of bonobo-specific and bonobo/chimpanzee-shared *Alu* integrants, the number of human-specific L1 integrants is very similar to the combined number of bonobo-specific and bonobo/chimpanzee-shared L1 integrants.

In addition to observing that the patterns of transposon integrants mostly match expected phylogenetic relationships, we identified a small number of integrants whose absence (A) or presence (P) status diverges from these expectations. We found 48 transposon integrants that are shared in human and bonobo, but not present at the corresponding target sites in chimpanzee, and 38 integrants that are present in human and chimpanzee, but not in bonobo. These integrants are strong candidates for incomplete lineage sorting (ILS), i.e. their genealogy does not match the expected overall phylogenetic relationship between genomes[51].

To corroborate these transposon integrants further as “ILS transposons”, we further examined their absence/presence status in the orangutan genome assembly (ponAbe2) as an outgroup[60]. This “filtering” helped remove particular cases where integrants may have been deleted from a single genome, as they would not be bona fide cases of ILS. Specifically, I used BLAST to align a human transposon integrant including flanking sequences to its corresponding

orangutan region to exclude the corresponding orangutan transposon ortholog. We required that the flanking genomic sequences on both sides of human transposon integrants must align with >95% identity to the orangutan genome, and there must be no transposon remnant in the corresponding orangutan genome region. Demanding such a high quality alignment on both sides reduced the number of BC cases. Upon this further filtering, 27 BH integrants were found to be present in bonobo and human and absent in chimpanzee and orangutan, 30 CH integrants are present in human and chimpanzee and absent in bonobo and orangutan, and 947 BC integrants are present in bonobo and chimpanzee and absent in human and orangutan. The former two categories are ILS transposons.

We compared the ILS transposon integrants to the ILS status of flanking genomic blocks called from nearby SNPs using the CoalHMM model (SOM8: Speciation times, ancestral population sizes and incomplete lineage sorting). We observed a significant enrichment of BH status calls around transposon integrants that are present in bonobo and human but absent in chimpanzee and orangutan (p=6.3e-11). Similarly, we observed a significant enrichment of CH status calls around transposon integrants that are present in chimpanzee and human but absent in bonobo and orangutan (p=1.73e-6) (Fig. S6-2).

We also examined the overlap of ILS status between ILS segmental duplications (see SOM4: Segmental Duplication Analysis of the Bonobo Genome) and ILS status calls made by SNPs. We compared 17 segmental duplications (10 CH segmental duplication loci and 7 BH segmental duplication loci) to the neighboring ILS blocks predicted by CoalHMM. Unlike the association between ILS transposons, the ILS status of segmental duplications generally did not overlap with the ILS status predicted by CoalHMM. Only one CH segmental duplication locus shared its ILS status with the CoalHMM prediction.

Experimental validation

Using PCR, we validated a collection of bonobo-specific transposon integrants that were predicted from sequence trace alignments (Figure S6-3). Genomic DNAs from bonobo ('Ulindi'), chimpanzee, and human individuals were assayed with oligonucleotide PCR primers that were designed to cross specific integrant target sites. If a predicted transposon integrant occupies a target site, a longer PCR product would be obtained than if the target site were empty. We performed PCR reactions for 32 bonobo specific transposon integrants. The results

demonstrate that 31 out of 32 predicted bonobo-specific transposon integrants are present in bonobo (97% true positive rate). At one particular integrant site, we observed two PCR amplification products: one contained the predicted *Alu* integrant, while another amplified the empty target site, suggesting heterozygosity at the site. Thus this element may not be fixed in the bonobo population.

To validate predicted ILS transposons, we assayed 25 arbitrarily chosen, predicted ILS integrants (with predicted status either BH or CH) using PCR. Out of 14 predicted BH ILS cases tested, we confirmed 13 as BH. The failed case was only present in human (H+, B-, C-). We tested 11 predicted CH ILS cases and confirmed 8 as CH. One of these confirmed cases had two bands in the chimpanzee gel lane, again suggesting heterozygosity. Thus this particular integrant may not be fixed in the chimpanzee population. Of the three failed CH cases, two are present in all three primates (H+, B+, C+), while another is present only in human (H+, B-, C-). Overall, we confirmed 21 out of 25 predicted ILS transposons (84% true positive rate).

2 - Rates of Transposon Accumulation

Using all available sequence trace datasets, we observed six times more human-specific *Alu* elements than bonobo-specific *Alu* elements (Table S6-1). We note that these counts could be affected by differences in trace coverage and sequencing platforms used, and by use of the human genome as the reference for trace mapping.

Therefore, to compare rates of transposon accumulation in the human, chimpanzee, and bonobo lineages more accurately, we first normalized the sequence trace datasets. To remove biases in coverage and sequencing platforms, we normalized sequence read lengths from conventional Sanger sequencing (starting average read length ~ 800 nt) to match read lengths from 454 Titanium sequencing as obtained from the bonobo sequencing project (average read length ~450 nt). To match sequence trace counts, we used 42 million traces (18.9 Gbp, 6.5x coverage) from one individual of each species: 454 Titanium reads for the bonobo individual ("Ulindi"), Sanger sequencing reads for the chimpanzee individual ("Clint")[55], and Sanger sequencing reads for the human individual ("Craig Venter")[61]. To identify integrants uniquely present in the three primates but not present in the orangutan genome, we mapped these size-adjusted traces against the orangutan genome assembly (UCSC genome browser ponAbe2). We obtained >20 million well-aligned traces (human=23.6 million traces, bonobo=27.5 million

traces, and chimpanzee=21.3 million traces) and ~30,000 transposon variant-supporting traces for each species.

As shown in Table S6-2, we identified 1,290 human-specific, 209 bonobo-specific, and 256 chimpanzee-specific *Alu* integrants that are not present in the normalized orangutan genome sequence traces. Using these normalized counts, we found that the number of human-specific *Alu* integrants is still approximately six times that of bonobo-specific and chimpanzee-specific integrants. We calculated the rate of *Alu* retrotransposon accumulation for each lineage by using the following average split times: 2.2 million years ago for bonobo-chimpanzee split, 6.5 million years ago for human-pan (bonobo and chimpanzee) split, and 15 million years ago for orangutan-human split. The calculated accumulation rate of *Alu* elements in human is 198 integrants per million years, and the accumulation rates for bonobo specific and chimpanzee specific lineages are 95 and 116 integrants per million years respectively. Thus the rate of *Alu* accumulation in human has been approximately twice as high as in the bonobo specific- and chimpanzee specific lineages. This is consistent with a previous analysis comparing human and chimpanzee retrotransposition[62]. Interestingly, the rates of *Alu* retrotransposon accumulation in the bonobo/chimpanzee shared- and bonobo/chimpanzee/human shared lineages were very low (approximately 20 and 12 integrants per million years) compared with the lineage-specific rates. Thus the rates of *Alu* retrotransposition have significantly increased in the three distinct primates compared to in their shared ancestors. This surprising, recent acceleration in the recent lineages is compatible with the nearly neutral model of evolution, where slightly deleterious mutations may rise to appreciable frequencies before their eventual erasure from the population by negative selection. By contrast, the rate of L1 retrotransposition has remained fairly constant in each lineage (30 -56 integrants/million years). Similar analyses also were performed using a simulated human genome lacking all repetitive sequences as a reference genome. Comparable results were obtained (data not shown).

3 - Diversity of *Alu* subfamilies

To study lineage-specific differences in *Alu* retrotransposon sequences, we categorized *Alu* subfamilies using the program Alucode[52, 59]. This algorithm is based on identification and clustering of similar *Alu* transposon sequences. Since sequence variants present in a parental transposon copy most likely will be inherited identically in progeny copies, all of the related

elements can be identified as members of a subfamily[63].

We extracted full-length *Alu* elements (size >280 nt) from the three primate genome assemblies. Subsequently we generated a multiple sequence alignment of all extracted elements and tabulated the nucleotide value of each position based on the *Alu*Sx consensus sequence. *Alu*S represents the major burst of *Alu* elements in primates about 36 million years ago, and Sx is one of the most common members of the *Alu*S subgroup. Based on our analysis of 65,036 *Alu* elements using Alucode[52, 59], we identified 53 *Alu* subfamilies. Of these, 32 belong to the young *Alu*Y subfamilies (Table S6-5). We constructed a minimum spanning tree for these *Alu*Y subfamilies to summarize their evolutionary relationships (Figure S6-6). We identified 24 subfamilies shared among the three primates, including six human-specific subfamilies (nodes 12, 19, 24, 25, 29, 30), and two *Pan*-lineage (bonobo and chimpanzee)-specific subfamilies (nodes 14 and 26). The two *Pan*-lineage specific *Alu* subfamilies identified here have not been reported previously; we call them *Alu*P1 and *Alu*P2. The subfamilies identified here range from 20 to 3,308 elements; the biggest subfamily is the generic *Alu*Y group. The p values for subfamily partition range from 1e-16 to <1e -100 as shown in Table S6-5. Based on a majority rule, *Alu* consensus sequences were derived from each group and are shown in Figure S6-7, including the new consensus sequences defining *Alu*P1 and *Alu*P2.

We also categorized L1 subfamilies using Alucode for L1 3' end sequences. We detected 19 L1 subfamilies diverged from L1PA3. We confirmed two subfamilies specific to the human lineage (L1Hs) and one subfamily specific to the bonobo/chimpanzee lineage (L1Pt), corroborating a recent report[64]. We did not detect any bonobo-specific L1 subfamilies.

4 - Genes near transposon integrants

Method of ontology analysis

To test various hypotheses about the genome-wide distributions of transposon integrant variants between the three primates, we generated lists of simulated integrants using a random number generator to assign chromosomal coordinates. To approximate genomic or intragenic distributions, we created 2 million simulated insertions based on the human genome. We counted numbers of observed vs. *in silico* transposon integrants inside or within +/-50 kb of RefSeq genes.

To investigate whether genes involved in various ontological categories (biological

processes or molecular functions) are affected by observed transposon integrants, we used gene IDs associated with each accession to query the PANTHER database (version 6) at http://www.pantherdb.org[65]. For comparison, simulated integrants were used to calculate the expected patterns of integrants and affected genes. Ontological categories (biological processes or molecular functions) were deemed significantly affected if their p values are <0.05 as determined by the binominal statistic. In this analysis we tested the 29 top node categories from biological processes and 28 top node categories from molecular functions. We applied Bonferroni's multiple testing correction method to obtain adjusted p values.

We searched for over-/under- representation of gene ontology classifications among RefSeq genes containing transposon integrant variants among three primates. RefSeq genes within +/-50 kb from transposon integrations were subjected to gene ontology analysis based on biological process and molecular function in the PANTHER database. There were no significantly associated ontology terms for *Alu* integrants. By contrast, we found enrichment of biological processes for L1 transposon integrants, such as "neuronal activities", "cell adhesion", "developmental processes", while genes involved in "nucleoside, nucleotide and nucleic acid metabolism" were rare (Figure S6-4A & Table S6-3B). For molecular functions, genes with nearby L1 transposon integrants were enriched in "extracellular matrix", "cell adhesion molecule", while genes involved with "transcription factor" were less frequently observed around L1 integrants (Figure S6-4B & Table S6-3D). We obtained similar ontology results using L1 integrants located inside of RefSeq genes (Table S6-3A and C).

We examined whether bonobo-specific L1 elements are depleted or enriched in particular categories of genes because of the genes' lengths or GC content (Fig. S6-5). All counts were compared with a genome-wide simulation of integrants mapped to the reference human genome assembly ("*in silico*") that normalizes for gene lengths. L1 integrants are preferentially localized in AT-rich regions of the human and mouse genomes[48, 66, 67]. We counted RefSeq genes in the bonobo, chimpanzee and human genomes, divided into categories according to flanking GC content. We then counted intronic L1 integrants for genes in each bin (Fig S6-4A). Again, as a control, we simulated integrants genome-wide according to the reference human genome assembly. Of these simulated integrants, 47% are inside of genes with <40% GC content. By contrast, >70% of L1 integrants in each species are located inside of genes with <40% GC content (Fig. S6-5A). The enrichment of L1 integration in AT-rich genes is significant in all

lineages compared to *in silico* control integrants (p<5e-10). We compared the distributions of GC contents for genes in ontology categories for "neuronal activities" and "nucleoside, nucleotide, and nucleic acid metabolism" to those for RefSeq genes overall (Fig S6-4B). We did not observe significant differences in GC content between them (p> 0.4).

As another control for possible bias in L1 integrants enriched or depleted from certain gene ontology categories, we re-analyzed gene ontologies using only genes with less than 40% GC content (i.e. counting only 5,437 out of 23,216 refSeq genes). Even within low GC content genes, we observed the trend of increased L1 integrants in "neuronal activities" genes and "developmental processes" in all species (data not shown). The enrichment of intronic L1 transposon integrants in these terms was statistically significant for combined lineage specific L1 integrants ("neuronal activities": p=6.2-e03, "developmental processes": p=9.8e-5).

Orientation of Alu and L1 integrants

L1 insertions in introns of genes may affect transcript levels[68]. This effect appears to be more substantial for insertions in the sense orientation than antisense orientation, since such integrants tend to be relatively depleted[48, 49]. In keeping with these results, we found a genome-wide reduction of sense orientation L1 insertions in introns (Table S6-4B). We also observed a smaller reduction of sense oriented *Alu* insertions in introns (Table S6-4A).

Surprisingly, both *Alu* and L1 retrotransposon integrants in introns that are shared in both the bonobo and chimpanzee lineages have no orientation bias. Their ratio of sense to antisense orientations is close to 50:50, as expected from random chance. By contrast, more recently inserted lineage-specific *Alu* and L1 retrotransposon integrants show a reduction of sense-oriented integrants.

We also examined the association between ontology terms and the orientation bias of L1 integrants. We compared sense-oriented L1 integrants located inside genes for ontology categories in the PANTHER database to the genome-wide sense orientation rate in each lineage. We did not observe a statistically significant bias for any gene ontology terms, relative to integrant orientation.

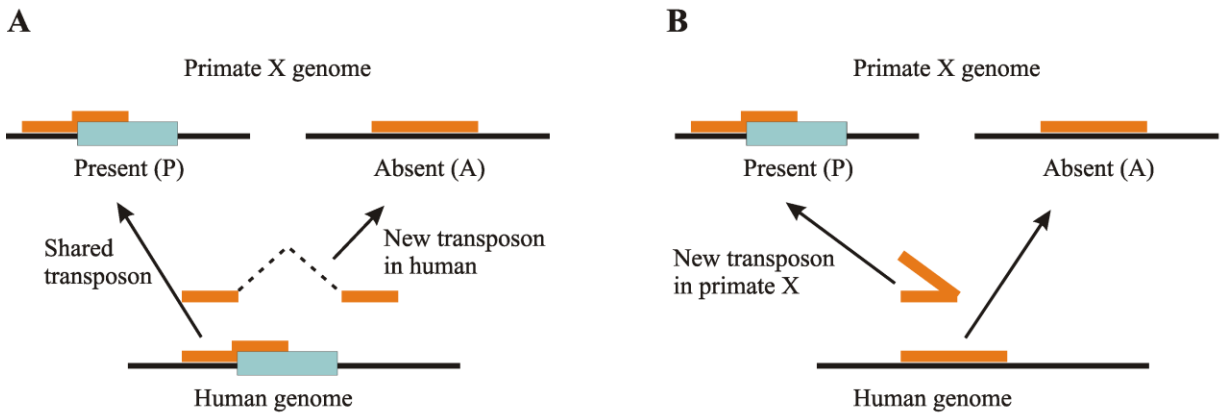


Figure S6-1: Discovery of transposon integrant variants from sequence trace alignments. (A) A schematic illustrating primate X traces (*orange rectangle*) aligned against the human genome (*horizontal black line, bottom*), identifying the presence or absence status of retrotransposon integrants (*light blue rectangle*). "Well-aligned" traces spanning across the transposon integrant junction support the presence of the integrant in primate X. If the transposon integrant is absent in primate X, the trace is aligned as two fragments skipping the human unique transposon integrant. (B) A schematic illustrating trace alignments for new transposon integrants in primate X. A trace (*orange*) from a new transposon integrant in primate X aligns well only partially to the human reference genome (*bottom*). The trace fragment that contains the transposon does not align to the reference genome.

transposon family	human specific human=P, bonobo=A, chimp=A	bonobo specific human=A, bonobo=P, chimp=A	chimpanzee specific human=A, bonobo=A, chimp=P	BC shared human=A, bonobo=P, chimp=P	BH shared human=P, bonobo=P, chimp=A	CH shared human=P, bonobo=A, chimp=P	HBC shared human=P, bonobo=P, chimp=P
Alu	5,203	859	664	403	31	16	725,765
ERV1	6	38	21	82	1	0	105,115
ERVK	57	14	3	13	1	0	4,347
ERVL	5	16	4	7	0	0	108,017
ERVL-MaLR	4	38	3	13	2	3	245,531
hAT-Charlie	0	6	0	10	1	0	169,749
L1	978	540	338	635	9	16	625,008
L2	2	11	1	9	1	1	302,002
MIR	7	10	5	15	1	0	393,175
SVA	379	46	37	16	1	1	359
TcMar-Tigger	0	12	3	4	0	1	69,105
Total	6,641	1,590	1,079	1,207	48	38	2,748,173

Table S6-1: Counts of transposons in human, bonobo and chimpanzee. All available shotgun-sequencing traces from the current bonobo and previous chimpanzee[55] genome projects were aligned to human reference genome (hg18), and the presence (P) or absence (A) of transposons was identified from trace alignment signatures. Only transposons confirmed by two or more traces were included. Column headers: BC shared, elements present in both bonobo and chimpanzee but not human; BH shared, elements present in both bonobo and human but not chimpanzee (i.e. ILS cases); CH shared, elements present in both chimpanzee and human but not bonobo (i.e. ILS cases); HBC shared, elements present in all three species, which include ancient integrants. We summed particular classes of retrotransposons that are known to be actively mobilized currently, i.e. *Alu*, L1 and SVA elements, to calculate lineage-specific retrotransposon integrants in bonobo and chimpanzee (Table 1).

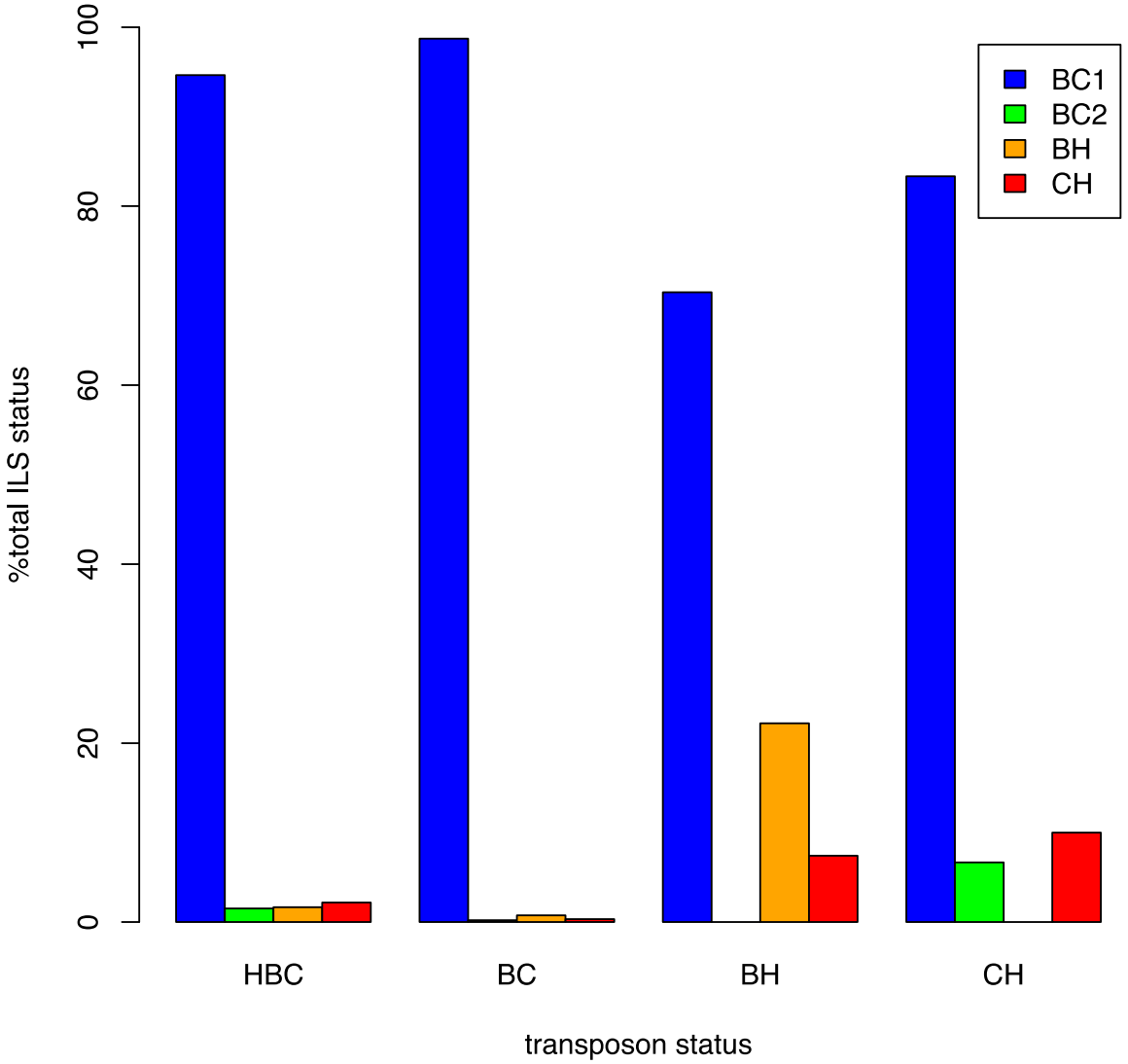


Figure S6-2: ILS transposons are associated with CoalHMM-predicted ILS blocks. Transposon integrants that were absent in orangutan were categorized according to their presence or absence from the human, bonobo and chimpanzee genomes as indicated (*x-axis*, HBC=human/bonobo/chimpanzee shared, BC=bonobo/chimpanzee shared, BH=bonobo/human shared, and CH=chimpanzee/human shared integrants). These integrants were assessed for predicted flanking Coal-HMM ILS status (*key*), based on SNPs as called by CoalHMM (see SOM 8). Fractions of integrants mapping within or adjacent to ILS blocks are indicated: BC1 (*blue*), BC2 (*green*), BH (*orange*), CH (*red*). In the CoalHMM ILS state BC1 (*key, blue*), bonobo and chimpanzee coalesced before speciation of bonobo, chimpanzee, and human (preferred state). In the other three ILS states, all three species coalesced in the common ancestor of bonobo, chimpanzee and human. In state BC2, bonobo and chimpanzee coalesced first. In ILS state BH, bonobo and human coalesced first. In state CH, chimpanzee and human coalesced first.

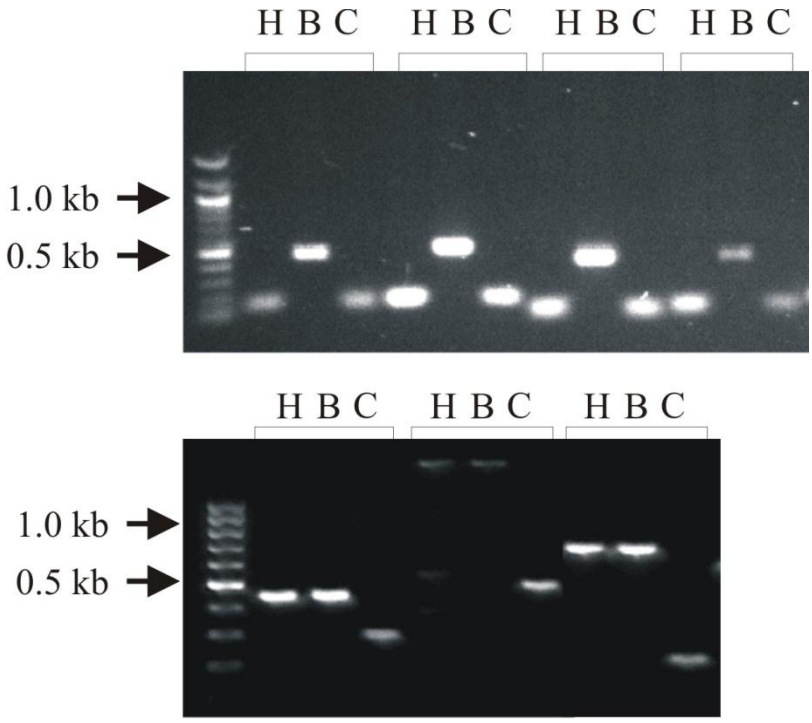


Figure S6-3: PCR verification of predicted transposon integrants. We confirmed the presence of transposon integrants by designing primers spanning integrant target sites that had been predicted by our computational method. If a transposon integrant is present in a sample, the longer product is amplified by PCR. The DNA source used in each lane is indicated (H=human, B=bonobo, C=chimpanzee). *Upper panel*: Four bonobo-lineage specific transposon integrant sites. Only bonobo lanes show the larger PCR products, reflecting the bonobo-specific nature of the integrants. *Lower panel*: Three loci containing incomplete lineage sorting (ILS)-transposons. These integrants were shared in bonobo and human, but not in chimpanzee, and the bands showed the appropriate size products as predicted only in bonobo and human lanes.

transposon family	unique to human		unique to bonobo		unique to chimpanzee		unique to bonobo/chimp		unique to human, bonobo and chimp	
	count	rate	count	rate	count	rate	count	rate	count	rate
Alu	1,290	198.5	209	95.0	256	116.4	87	20.2	106	12.5
ERV1	18	-	3	-	8	-	28	-	17	-
ERVK	10	-	5	-	2	-	1	-	7	-
ERVL	7	-	1	-	1	-	4	-	12	-
ERVL-MaLR	31	-	10	-	8	-	8	-	25	-
L1	313	48.2	124	56.4	119	54.1	131	30.5	322	37.9
L2	3	-	4	-	2	-	1	-	18	-
MIR	3	-	1	-	1	-	0	-	26	-
MuDR	0	-	0	-	0	-	0	-	0	-
SVA	55	8.5	7	3.2	22	10.0	3	0.7	6	0.7
Total	1,730	266.2	364	165.5	419	190.5	263	61.2	539	63.4

Table S6-2: Rates of transposon accumulation in human, bonobo and chimp. To normalize the overall coverage of sequencing data, we normalized the lengths and numbers of available sequence traces from each ape genome. To minimize the possibility that aligned traces exhibit a bias toward integrants in the human genome (because that reference assembly is more refined), we also aligned the traces to the orangutan genome (ponAbe2) as an outgroup. The rates of transposon integration events were estimated for each lineage upon normalization. New transposition (accumulation) events were counted from trace alignment signatures, and then rates of accumulation were calculated by dividing these numbers by the average speciation times; 2.2 million years ago for bonobo-chimpanzee split, 6.5 million years ago for human-pan (bonobo and chimpanzee) split, and 15 million years ago for orangutan-human split. Column headers: *unique to bonobo/ chimp*, elements present in the bonobo and chimpanzee species but not human or orangutan; *unique to human, bonobo and chimpanzee*, elements present in all three species but not orangutan.

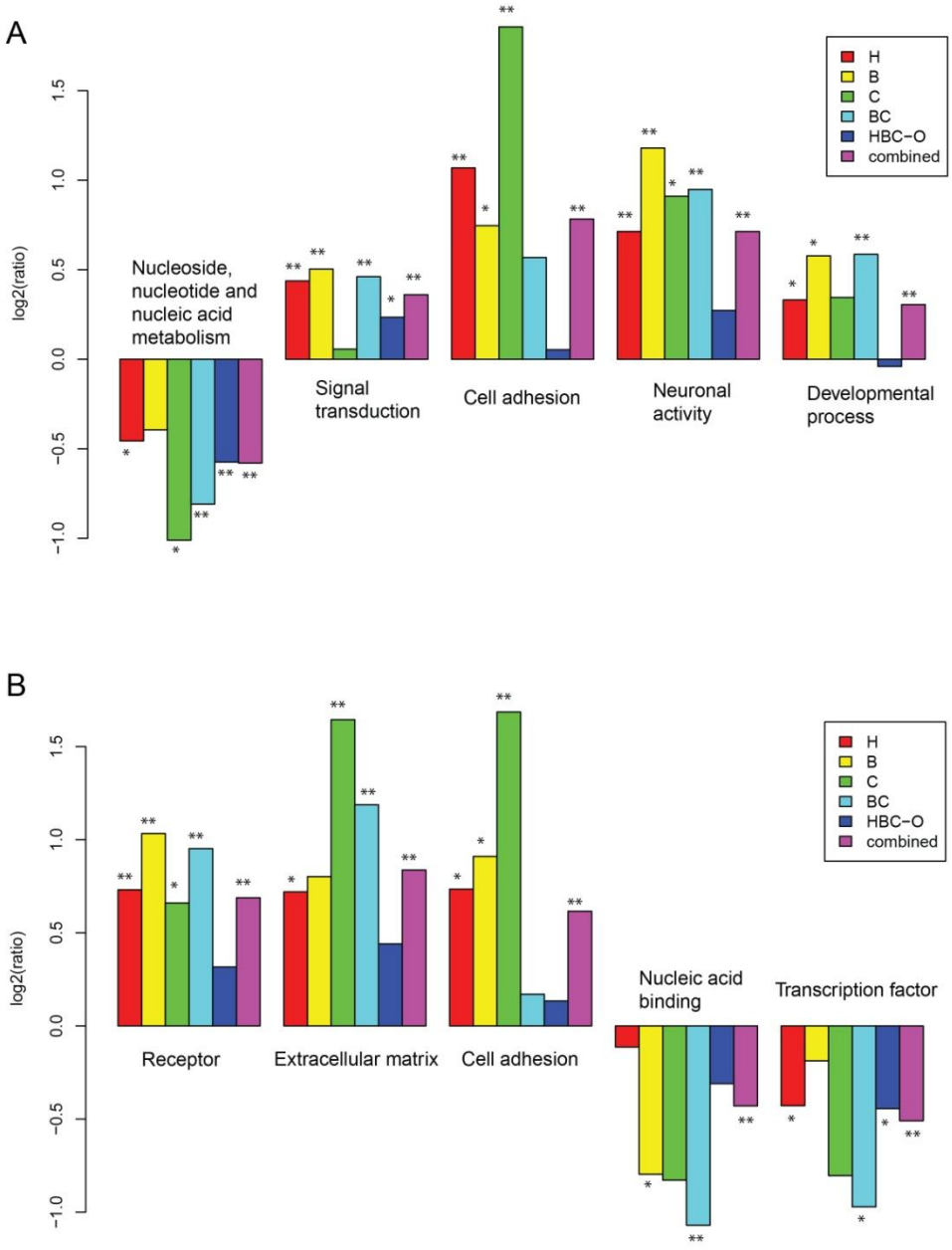
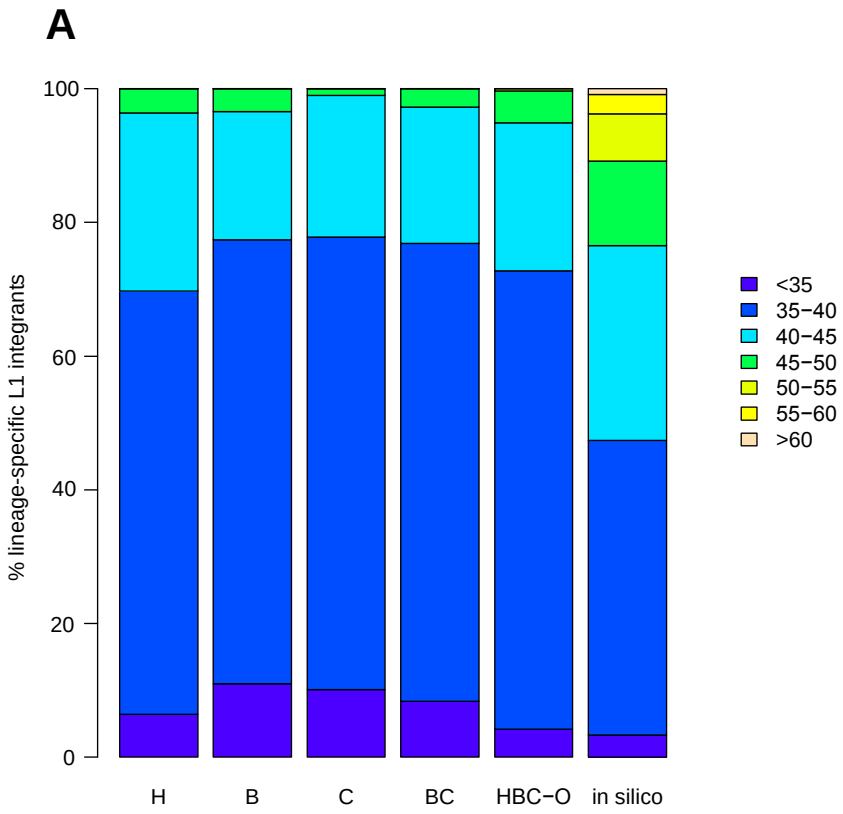


Figure S6-4: Enrichment or depletion of L1 integrants in various gene function categories. L1 integrants that mapped within 50 kb up- or downstream of various genes were identified in various primate lineages (*key*). Ontological categories of genes with observed integrants were assessed, including (*upper panel*) biological processes and (*lower panel*) molecular functions. Expected numbers of integrants were calculated by *in silico* simulations of 2 million integrants. The five most frequent ontological categories (based on binomial test), in which observed gene numbers are significantly different from expected numbers, are depicted. The log of the (observed/expected) ratio is shown in each case. *:p<0.05, **: adjusted p <0.05. *Key*: H, human-specific integrants; B, bonobo; C, chimp; BC, bonobo/chimp shared; HBC-O, human, bonobo and chimp shared integrants that are absent in orangutan; combined, all of the lineages combined with no integrants in orangutan.


A
% lineage-specific L1 integrants
100
80
60
40
20
0
H
B
C
BC
HBC−O
in silico
<35
35−40
40−45
45−50
50−55
55−60
>60

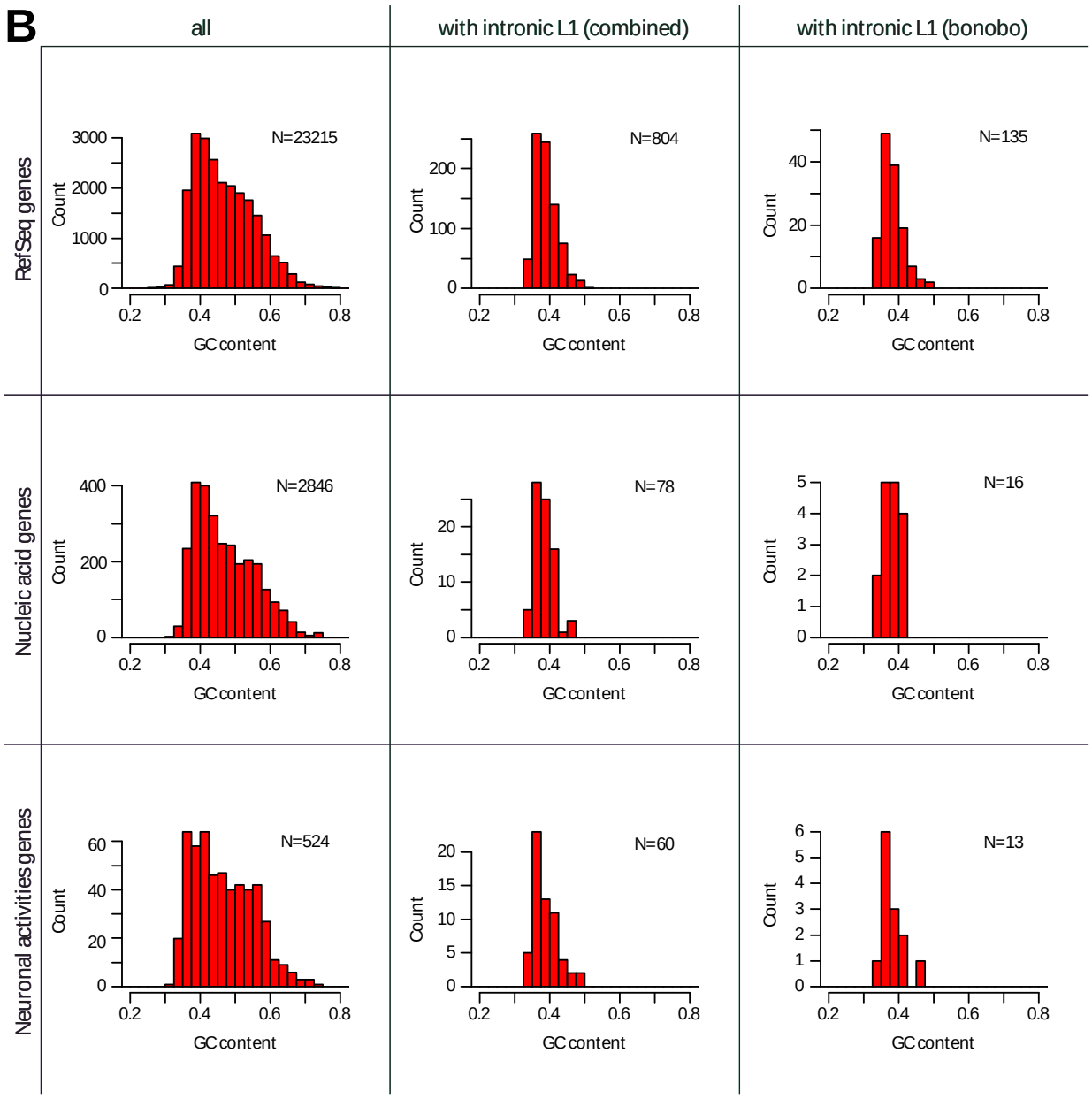

B
all
with intronic L1 (combined)
with intronic L1 (bonobo)
RefSeq genes
Nucleic acid genes
Neuronal activities genes
Count
GC content
N=23215
N=804
N=135
N=2846
N=78
N=16
N=524
N=60
N=13

Figure S6-5: GC content of genes and gene ontology terms. (A) **Lineage-specific L1 integrants in** RefSeq genes were categorized according to the genes' GC content as indicated (*key*). Corresponding fractions of all intronic L1 integrants are shown for each lineage (H: human-specific, B: bonobo-specific, C: chimpanzee-specific, BC: bonobo/chimp-shared, and HBC-O: human/bonobo/chimp-present and orangutan-absent). *Right*: expected numbers of integrants were calculated by *in silico* simulations of 2 million integrants. In all lineages, L1 integrants are enriched in genes with GC content < 40% ($p < 5e-10$). (B) Counts of Refseq genes (*top row*) and gene ontology categories "nucleic acid" (*middle row*) and "neuronal activities" (*bottom row*), categorized according to presence in the reference human genome (*left column*) or their inclusion of intronic L1 integrants in any of the three HBC-specific lineages after orangutan (*middle, "combined"*) or in the bonobo-specific lineage (*right*). The results display no significant difference in the GC contents of these ontology categories compared with the reference genome overall ($p > 0.4$).

term	in-silico	human (n=328)			bonobo (n=132)			chimpanzee (n=86)			BC (n=193)			HBC-O (n=283)			combined (n=1022)		
		hit	rate	pval	hit	rate	pval	hit	rate	pval	hit	rate	pval	hit	rate	pval	hit	rate	pval
Amino acid metabolism	0.9%	3	0.9%	7.7E-01	0	0.0%	6.4E-01	1	1.2%	5.4E-01	3	1.6%	2.6E-01	1	0.4%	5.3E-01	8	0.8%	8.7E-01
Apoptosis	2.5%	10	3.0%	4.7E-01	2	1.5%	7.8E-01	1	1.2%	7.3E-01	5	2.6%	8.2E-01	6	2.1%	8.5E-01	24	2.3%	9.2E-01
Biological process unclassified	30.6%	84	25.6%	5.5E-02	36	27.3%	4.5E-01	21	24.4%	2.4E-01	45	23.3%	2.9E-02	99	35.0%	1.2E-01	285	27.9%	6.2E-02
Blood circulation and gas exchange	0.2%	0	0.0%	1.0E+00	0	0.0%	1.0E+00	0	0.0%	1.0E+00	0	0.0%	1.0E+00	1	0.4%	4.9E-01	1	0.1%	7.4E-01
Carbohydrate metabolism	3.2%	17	5.2%	5.9E-02	3	2.3%	8.0E-01	3	3.5%	7.6E-01	7	3.6%	6.8E-01	0	0.0%	**1.5E-04**	30	2.9%	6.6E-01
Cell adhesion	6.9%	36	11.0%	6.1E-03	13	9.8%	1.7E-01	17	19.8%	**7.2E-05**	19	9.8%	1.2E-01	16	5.7%	4.8E-01	101	9.9%	**3.3E-04**
Cell cycle	5.3%	15	4.6%	6.2E-01	5	3.8%	5.6E-01	6	7.0%	4.7E-01	9	4.7%	8.7E-01	13	4.6%	6.9E-01	48	4.7%	4.0E-01
Cell proliferation and differentiation	5.1%	15	4.6%	8.0E-01	5	3.8%	6.9E-01	5	5.8%	6.3E-01	13	6.7%	3.2E-01	14	4.9%	1.0E+00	52	5.1%	9.4E-01
Cell structure and motility	8.2%	22	6.7%	3.7E-01	17	12.9%	5.8E-02	16	18.6%	2.3E-03	22	11.4%	1.2E-01	21	7.4%	7.5E-01	98	9.6%	1.2E-01
Coenzyme and prosthetic group metabolism	0.7%	2	0.6%	1.0E+00	1	0.8%	5.9E-01	0	0.0%	1.0E+00	0	0.0%	6.5E-01	0	0.0%	2.7E-01	3	0.3%	1.8E-01
Developmental processes	18.2%	78	23.8%	1.2E-02	41	31.1%	**4.0E-04**	20	23.3%	2.6E-01	57	29.5%	**1.2E-04**	51	18.0%	1.0E+00	247	24.2%	**2.0E-06**
Electron transport	0.7%	3	0.9%	7.4E-01	0	0.0%	1.0E+00	0	0.0%	1.0E+00	0	0.0%	4.1E-01	3	1.1%	4.7E-01	6	0.6%	7.2E-01
Homeostasis	1.1%	3	0.9%	1.0E+00	1	0.8%	1.0E+00	0	0.0%	1.0E+00	3	1.6%	4.8E-01	3	1.1%	1.0E+00	10	1.0%	8.8E-01
Immunity and defense	4.7%	14	4.3%	9.0E-01	8	6.1%	4.1E-01	3	3.5%	8.0E-01	10	5.2%	7.3E-01	11	3.9%	6.7E-01	46	4.5%	8.8E-01
Intracellular protein traffic	6.1%	16	4.9%	4.2E-01	3	2.3%	6.8E-02	9	10.5%	1.1E-01	9	4.7%	5.5E-01	14	4.9%	4.6E-01	51	5.0%	1.3E-01
Lipid, fatty acid and steroid metabolism	4.2%	9	2.7%	2.2E-01	6	4.5%	8.3E-01	1	1.2%	2.7E-01	4	2.1%	1.5E-01	11	3.9%	8.8E-01	31	3.0%	5.2E-02
Muscle contraction	1.5%	6	1.8%	6.5E-01	4	3.0%	1.4E-01	2	2.3%	3.7E-01	4	2.1%	5.4E-01	5	1.8%	6.2E-01	21	2.1%	1.6E-01
Neuronal activities	7.2%	36	11.0%	1.3E-02	18	13.6%	9.8E-03	12	14.0%	3.2E-02	25	13.0%	4.6E-03	29	10.2%	5.0E-02	120	11.7%	**1.5E-07**
Nitrogen metabolism	0.1%	0	0.0%	1.0E+00	0	0.0%	1.0E+00	0	0.0%	1.0E+00	0	0.0%	1.0E+00	1	0.4%	2.1E-01	1	0.1%	5.7E-01
Nucleoside, nucleotide metabolism	14.6%	38	11.6%	1.4E-01	16	12.1%	5.4E-01	3	3.5%	1.9E-03	20	10.4%	1.0E-01	26	9.2%	8.8E-03	103	10.1%	**2.0E-05**
Oncogenesis	2.2%	8	2.4%	7.1E-01	2	1.5%	1.0E+00	2	2.3%	7.2E-01	8	4.1%	8.0E-02	8	2.8%	4.2E-01	28	2.7%	2.4E-01
Other metabolism	2.2%	10	3.0%	3.4E-01	2	1.5%	1.0E+00	2	2.3%	7.2E-01	6	3.1%	3.3E-01	5	1.8%	8.4E-01	25	2.4%	6.0E-01
Phosphate metabolism	0.7%	2	0.6%	1.0E+00	1	0.8%	6.3E-01	1	1.2%	4.7E-01	1	0.5%	1.0E+00	4	1.4%	1.6E-01	9	0.9%	5.8E-01
Protein metabolism and modification	14.9%	50	15.2%	8.8E-01	23	17.4%	3.9E-01	14	16.3%	7.6E-01	28	14.5%	1.0E+00	45	15.9%	6.2E-01	160	15.7%	5.1E-01
Protein targeting and localization	1.7%	5	1.5%	1.0E+00	1	0.8%	7.3E-01	3	3.5%	1.8E-01	0	0.0%	8.5E-02	3	1.1%	6.4E-01	12	1.2%	2.3E-01
Sensory perception	1.8%	3	0.9%	3.0E-01	1	0.8%	7.4E-01	0	0.0%	4.1E-01	5	2.6%	4.0E-01	3	1.1%	5.0E-01	12	1.2%	1.6E-01
Signal transduction	28.0%	108	32.9%	5.6E-02	48	36.4%	4.1E-02	29	33.7%	2.3E-01	74	38.3%	2.2E-03	82	29.0%	7.4E-01	341	33.4%	**1.9E-04**
Sulfur metabolism	0.5%	2	0.6%	6.7E-01	1	0.8%	4.7E-01	0	0.0%	1.0E+00	3	1.6%	6.7E-02	2	0.7%	3.9E-01	8	0.8%	1.7E-01
Transport	8.3%	29	8.8%	6.9E-01	11	8.3%	1.0E+00	5	5.8%	5.6E-01	16	8.3%	1.0E+00	28	9.9%	3.3E-01	89	8.7%	6.5E-01

Table S6-3(A). "Biological process" gene ontological categories are enriched/depleted in intronic L1 integrants. P values were calculated by binomial statistics. *Red highlights:* p<0.05; *boxed, bold red*: significant after Bonferroni correction.

term	in-silico	human (n=638)			bonobo (n=226)			chimpanzee (n=154)			BC (n=346)			HBC-O (n=645)			combined (n=2009)		
		hit	rate	pval	hit	rate	pval	hit	rate	pval	hit	rate	pval	hit	rate	pval	hit	rate	pval
Amino acid metabolism	1.0%	6	0.9%	1.0E+00	0	0.0%	1.8E-01	2	1.3%	6.7E-01	6	1.7%	1.8E-01	2	0.3%	7.6E-02	16	0.8%	3.8E-01
Apoptosis	2.6%	21	3.3%	2.6E-01	2	0.9%	1.4E-01	3	1.9%	1.0E+00	11	3.2%	4.0E-01	26	4.0%	2.4E-02	63	3.1%	1.0E-01
Biological process unclassified	33.6%	185	29.0%	1.5E-02	63	27.9%	7.8E-02	40	26.0%	4.9E-02	105	30.3%	2.1E-01	240	37.2%	5.5E-02	633	31.5%	5.0E-02
Blood circulation and gas exchange	0.4%	0	0.0%	1.8E-01	0	0.0%	1.0E+00	0	0.0%	1.0E+00	1	0.3%	1.0E+00	3	0.5%	5.2E-01	4	0.2%	2.7E-01
Carbohydrate metabolism	3.1%	22	3.4%	5.7E-01	7	3.1%	1.0E+00	5	3.2%	8.1E-01	9	2.6%	7.6E-01	13	2.0%	1.4E-01	56	2.8%	4.8E-01
Cell adhesion	4.5%	60	9.4%	**1.4E-07**	17	7.5%	3.5E-02	25	16.2%	**2.7E-08**	23	6.6%	6.7E-02	30	4.7%	7.8E-01	155	7.7%	**1.5E-10**
Cell cycle	5.1%	20	3.1%	2.4E-02	6	2.7%	1.3E-01	9	5.8%	5.8E-01	14	4.0%	4.6E-01	27	4.2%	3.7E-01	76	3.8%	8.1E-03
Cell proliferation and differentiation	4.8%	26	4.1%	4.1E-01	8	3.5%	4.4E-01	7	4.5%	1.0E+00	17	4.9%	9.0E-01	27	4.2%	5.2E-01	85	4.2%	2.1E-01
Cell structure and motility	6.7%	36	5.6%	3.0E-01	23	10.2%	4.5E-02	18	11.7%	2.2E-02	30	8.7%	1.6E-01	30	4.7%	3.4E-02	137	6.8%	8.6E-01
Coenzyme and prosthetic group metabolism	0.8%	3	0.5%	5.1E-01	2	0.9%	7.1E-01	0	0.0%	6.4E-01	2	0.6%	1.0E+00	6	0.9%	6.6E-01	13	0.6%	4.6E-01
Developmental processes	14.8%	119	18.7%	8.7E-03	50	22.1%	3.5E-03	29	18.8%	1.7E-01	77	22.3%	**2.6E-04**	93	14.4%	8.2E-01	368	18.3%	**1.9E-05**
Electron transport	1.0%	8	1.3%	5.5E-01	0	0.0%	1.8E-01	0	0.0%	4.2E-01	0	0.0%	5.6E-02	8	1.2%	5.5E-01	16	0.8%	3.7E-01
Homeostasis	1.0%	5	0.8%	6.9E-01	1	0.4%	7.3E-01	0	0.0%	4.2E-01	5	1.4%	4.1E-01	7	1.1%	8.4E-01	18	0.9%	6.6E-01
Immunity and defense	6.6%	45	7.1%	6.3E-01	14	6.2%	1.0E+00	20	13.0%	3.2E-03	29	8.4%	1.9E-01	59	9.1%	1.1E-02	167	8.3%	2.5E-03
Intracellular protein traffic	5.6%	34	5.3%	8.6E-01	6	2.7%	5.8E-02	12	7.8%	2.2E-01	16	4.6%	5.6E-01	33	5.1%	6.7E-01	101	5.0%	3.1E-01
Lipid, fatty acid and steroid metabolism	4.4%	19	3.0%	9.9E-02	7	3.1%	4.2E-01	2	1.3%	7.2E-02	9	2.6%	1.1E-01	22	3.4%	2.5E-01	59	2.9%	**1.1E-03**
Muscle contraction	1.1%	8	1.3%	7.1E-01	7	3.1%	1.6E-02	2	1.3%	7.0E-01	7	2.0%	1.2E-01	9	1.4%	4.6E-01	33	1.6%	4.5E-02
Neuronal activities	4.5%	47	7.4%	**1.1E-03**	23	10.2%	**2.8E-04**	13	8.4%	2.9E-02	30	8.7%	**6.4E-04**	35	5.4%	2.5E-01	148	7.4%	**9.3E-09**
Nitrogen metabolism	0.1%	0	0.0%	1.0E+00	0	0.0%	1.0E+00	0	0.0%	1.0E+00	0	0.0%	1.0E+00	1	0.2%	4.1E-01	1	0.0%	1.0E+00
Nucleoside, nucleotide and nucleic acid metabolism	15.7%	73	11.4%	2.7E-03	27	11.9%	1.4E-01	12	7.8%	5.3E-03	31	9.0%	**3.6E-04**	68	10.5%	**1.8E-04**	211	10.5%	**2.0E-11**
Oncogenesis	2.1%	11	1.7%	5.8E-01	3	1.3%	6.4E-01	5	3.2%	2.7E-01	13	3.8%	5.8E-02	19	2.9%	1.7E-01	51	2.5%	1.9E-01
Other metabolism	2.5%	14	2.2%	7.0E-01	3	1.3%	3.9E-01	2	1.3%	6.0E-01	9	2.6%	8.6E-01	13	2.0%	5.3E-01	41	2.0%	2.0E-01
Phosphate metabolism	0.6%	3	0.5%	8.0E-01	1	0.4%	1.0E+00	1	0.6%	6.2E-01	2	0.6%	1.0E+00	6	0.9%	3.1E-01	13	0.6%	8.9E-01
Protein metabolism and modification	14.1%	88	13.8%	8.6E-01	33	14.6%	8.5E-01	21	13.6%	1.0E+00	45	13.0%	5.9E-01	83	12.9%	4.0E-01	270	13.4%	3.9E-01
Protein targeting and localization	1.3%	8	1.3%	1.0E+00	4	1.8%	5.4E-01	4	2.6%	1.4E-01	0	0.0%	2.7E-02	5	0.8%	3.0E-01	21	1.0%	3.7E-01
Sensory perception	2.3%	25	3.9%	1.1E-02	7	3.1%	3.7E-01	1	0.6%	2.7E-01	18	5.2%	**1.6E-03**	14	2.2%	1.0E+00	65	3.2%	7.1E-03
Signal transduction	23.7%	205	32.1%	**1.5E-06**	76	33.6%	**7.4E-04**	38	24.7%	7.8E-01	113	32.7%	**1.8E-04**	180	27.9%	1.4E-02	612	30.5%	**5.7E-12**
Sulfur metabolism	0.5%	4	0.6%	5.6E-01	1	0.4%	1.0E+00	0	0.0%	1.0E+00	3	0.9%	2.3E-01	7	1.1%	3.9E-02	15	0.7%	1.0E-01
Transport	7.6%	58	9.1%	1.6E-01	19	8.4%	6.2E-01	15	9.7%	2.9E-01	24	6.9%	7.6E-01	58	9.0%	1.8E-01	174	8.7%	8.4E-02

Table S6-3(B): "Biological process" gene ontological categories are enriched/depleted in nearby (+/-50 kb) L1 integrants.
P-values were calculated by binomial statistics. *Red highlights:* p<0.05; *boxed, bold red*: significant after Bonferroni correction.

term	in-silico	human (n=328)			bonobo (n=132)			chimpanzee (n=86)			BC (n=193)			HBC-O (n=283)			combined (n=1022)		
		hit	rate	pval	hit	rate	pval	hit	rate	pval	hit	rate	pval	hit	rate	pval	hit	rate	pval
Cell adhesion molecule	4.2%	14	4.3%	8.9E-01	8	6.1%	2.7E-01	8	9.3%	2.7E-02	10	5.2%	4.7E-01	10	3.5%	7.6E-01	50	4.9%	2.4E-01
Cell junction protein	1.0%	6	1.8%	1.5E-01	2	1.5%	3.8E-01	5	5.8%	1.8E-03	2	1.0%	7.2E-01	1	0.4%	5.4E-01	16	1.6%	8.2E-02
Chaperone	0.5%	1	0.3%	1.0E+00	0	0.0%	1.0E+00	0	0.0%	1.0E+00	1	0.5%	6.0E-01	2	0.7%	3.9E-01	4	0.4%	1.0E+00
Cytoskeletal protein	5.7%	20	6.1%	7.2E-01	11	8.3%	1.8E-01	7	8.1%	3.4E-01	18	9.3%	4.0E-02	19	6.7%	4.4E-01	75	7.3%	2.5E-02
Defense/immunity protein	0.8%	3	0.9%	7.5E-01	0	0.0%	6.3E-01	1	1.2%	5.0E-01	1	0.5%	1.0E+00	3	1.1%	5.0E-01	8	0.8%	1.0E+00
Extracellular matrix	2.9%	16	4.9%	4.5E-02	7	5.3%	1.1E-01	7	8.1%	1.2E-02	13	6.7%	4.3E-03	14	4.9%	4.8E-02	57	5.6%	**3.7E-06**
Hydrolase	3.9%	20	6.1%	6.2E-02	3	2.3%	5.0E-01	0	0.0%	5.2E-02	8	4.1%	8.5E-01	11	3.9%	1.0E+00	42	4.1%	7.5E-01
Ion channel	3.4%	14	4.3%	3.6E-01	8	6.1%	8.9E-02	5	5.8%	2.2E-01	8	4.1%	5.4E-01	7	2.5%	5.1E-01	42	4.1%	1.9E-01
Isomerase	0.6%	3	0.9%	4.4E-01	0	0.0%	1.0E+00	1	1.2%	3.9E-01	1	0.5%	1.0E+00	0	0.0%	4.2E-01	5	0.5%	1.0E+00
Kinase	5.6%	28	8.5%	2.9E-02	12	9.1%	8.5E-02	6	7.0%	4.8E-01	11	5.7%	8.7E-01	26	9.2%	1.3E-02	83	8.1%	**7.8E-04**
Ligase	2.3%	7	2.1%	1.0E+00	2	1.5%	7.7E-01	3	3.5%	4.6E-01	4	2.1%	1.0E+00	7	2.5%	8.4E-01	23	2.3%	1.0E+00
Lyase	0.7%	4	1.2%	3.1E-01	0	0.0%	1.0E+00	1	1.2%	4.7E-01	1	0.5%	1.0E+00	1	0.4%	7.3E-01	7	0.7%	1.0E+00
Membrane traffic protein	2.1%	6	1.8%	1.0E+00	0	0.0%	1.2E-01	1	1.2%	1.0E+00	5	2.6%	6.1E-01	7	2.5%	5.4E-01	19	1.9%	7.4E-01
Miscellaneous function	3.5%	13	4.0%	6.5E-01	5	3.8%	8.1E-01	0	0.0%	7.8E-02	6	3.1%	1.0E+00	9	3.2%	8.7E-01	33	3.2%	6.7E-01
Molecular function unclassified	28.6%	87	26.5%	4.3E-01	37	28.0%	9.2E-01	19	22.1%	2.3E-01	39	20.2%	1.1E-02	87	30.7%	4.3E-01	269	26.3%	1.1E-01
Nucleic acid binding	10.2%	28	8.5%	3.6E-01	9	6.8%	2.5E-01	2	2.3%	1.1E-02	15	7.8%	2.9E-01	22	7.8%	2.0E-01	76	7.4%	2.7E-03
Oxidoreductase	2.5%	10	3.0%	4.8E-01	1	0.8%	2.7E-01	2	2.3%	1.0E+00	6	3.1%	4.9E-01	5	1.8%	5.7E-01	24	2.3%	8.4E-01
Phosphatase	2.2%	5	1.5%	5.7E-01	6	4.5%	6.9E-02	4	4.7%	1.2E-01	1	0.5%	1.4E-01	4	1.4%	5.4E-01	20	2.0%	7.5E-01
Protease	2.3%	5	1.5%	4.6E-01	4	3.0%	5.5E-01	1	1.2%	1.0E+00	8	4.1%	8.6E-02	10	3.5%	1.6E-01	28	2.7%	2.9E-01
Receptor	9.4%	44	13.4%	1.7E-02	24	18.2%	**1.5E-03**	15	17.4%	1.6E-02	29	15.0%	1.3E-02	33	11.7%	1.8E-01	145	14.2%	**6.9E-07**
Select calcium binding protein	2.4%	12	3.7%	1.4E-01	4	3.0%	5.6E-01	2	2.3%	1.0E+00	7	3.6%	2.3E-01	8	2.8%	5.6E-01	33	3.2%	7.9E-02
Select regulatory molecule	7.6%	17	5.2%	1.2E-01	4	3.0%	4.7E-02	9	10.5%	3.1E-01	14	7.3%	1.0E+00	17	6.0%	3.7E-01	61	6.0%	4.5E-02
Signaling molecule	4.6%	17	5.2%	6.0E-01	6	4.5%	1.0E+00	4	4.7%	1.0E+00	12	6.2%	3.0E-01	20	7.1%	6.4E-02	59	5.8%	8.7E-02
Synthase and synthetase	0.7%	1	0.3%	7.3E-01	1	0.8%	6.0E-01	0	0.0%	1.0E+00	1	0.5%	1.0E+00	0	0.0%	2.7E-01	3	0.3%	1.8E-01
Transcription factor	8.3%	18	5.5%	7.1E-02	9	6.8%	6.4E-01	3	3.5%	1.2E-01	9	4.7%	6.8E-02	17	6.0%	1.9E-01	56	5.5%	**6.6E-04**
Transfer/carrier protein	1.3%	2	0.6%	4.6E-01	2	1.5%	6.9E-01	1	1.2%	1.0E+00	1	0.5%	5.3E-01	2	0.7%	6.0E-01	8	0.8%	1.7E-01
Transferase	4.7%	12	3.7%	4.3E-01	5	3.8%	8.4E-01	4	4.7%	1.0E+00	10	5.2%	7.3E-01	11	3.9%	6.7E-01	42	4.1%	3.8E-01
Transporter	3.3%	13	4.0%	5.4E-01	5	3.8%	6.3E-01	0	0.0%	1.2E-01	4	2.1%	4.2E-01	11	3.9%	6.2E-01	33	3.2%	9.3E-01

Table S6-3(C). "Molecular functions" gene ontological categories are enriched/depleted in intronic L1 integrants.
P-values were calculated by binomial statistics. *Red highlights:* p<0.05; *boxed, bold red*: significant after Bonferroni correction.

term	in-silico	human (n=638)			bonobo (n=226)			chimpanzee (n=154)			BC (n=346)			HBC-O (n=645)			combined (n=2009)		
		hit	rate	pval	hit	rate	pval	hit	rate	pval	hit	rate	pval	hit	rate	pval	hit	rate	pval
Cell adhesion molecule	2.8%	30	4.7%	8.0E-03	12	5.3%	4.0E-02	14	9.1%	**1.3E-04**	11	3.2%	6.3E-01	20	3.1%	6.3E-01	87	4.3%	**1.5E-04**
Cell junction protein	0.7%	8	1.3%	8.8E-02	3	1.3%	2.0E-01	5	3.2%	4.4E-03	3	0.9%	5.1E-01	1	0.2%	1.4E-01	20	1.0%	1.0E-01
Chaperone	0.7%	2	0.3%	3.4E-01	2	0.9%	6.7E-01	1	0.6%	1.0E+00	2	0.6%	1.0E+00	4	0.6%	1.0E+00	11	0.5%	5.9E-01
Cytoskeletal protein	4.6%	25	3.9%	5.1E-01	16	7.1%	7.9E-02	7	4.5%	1.0E+00	22	6.4%	1.2E-01	27	4.2%	7.1E-01	97	4.8%	5.9E-01
Defense/immunity protein	1.3%	13	2.0%	1.2E-01	1	0.4%	3.8E-01	2	1.3%	1.0E+00	9	2.6%	5.5E-02	10	1.6%	6.1E-01	35	1.7%	1.2E-01
Extracellular matrix	2.3%	24	3.8%	2.3E-02	9	4.0%	1.1E-01	11	7.1%	**8.9E-04**	18	5.2%	**1.5E-03**	20	3.1%	1.8E-01	82	4.1%	**9.7E-07**
Hydrolase	3.7%	31	4.9%	1.2E-01	4	1.8%	1.6E-01	3	1.9%	3.9E-01	12	3.5%	1.0E+00	19	2.9%	4.0E-01	69	3.4%	5.9E-01
Ion channel	2.4%	19	3.0%	3.0E-01	9	4.0%	1.2E-01	6	3.9%	1.9E-01	13	3.8%	1.1E-01	13	2.0%	7.0E-01	60	3.0%	7.9E-02
Isomerase	0.7%	4	0.6%	1.0E+00	3	1.3%	2.0E-01	1	0.6%	1.0E+00	2	0.6%	1.0E+00	1	0.2%	1.4E-01	11	0.5%	5.9E-01
Kinase	4.2%	34	5.3%	1.6E-01	12	5.3%	4.0E-01	7	4.5%	6.9E-01	14	4.0%	1.0E+00	34	5.3%	1.7E-01	101	5.0%	5.8E-02
Ligase	2.2%	10	1.6%	3.4E-01	3	1.3%	6.4E-01	3	1.9%	1.0E+00	9	2.6%	5.8E-01	12	1.9%	6.9E-01	37	1.8%	3.6E-01
Lyase	0.8%	4	0.6%	1.0E+00	0	0.0%	4.2E-01	1	0.6%	1.0E+00	2	0.6%	1.0E+00	2	0.3%	2.5E-01	9	0.4%	1.5E-01
Membrane traffic protein	1.8%	17	2.7%	1.4E-01	1	0.4%	1.4E-01	2	1.3%	1.0E+00	9	2.6%	3.1E-01	12	1.9%	8.8E-01	41	2.0%	5.1E-01
Miscellaneous function	3.9%	29	4.5%	4.1E-01	11	4.9%	3.9E-01	9	5.8%	2.1E-01	18	5.2%	2.1E-01	21	3.3%	4.8E-01	88	4.4%	2.7E-01
Molecular function unclassified	31.6%	177	27.7%	3.7E-02	63	27.9%	2.5E-01	37	24.0%	4.6E-02	88	25.4%	1.3E-02	227	35.2%	5.1E-02	592	29.5%	4.1E-02
Nucleic acid binding	11.5%	68	10.7%	5.4E-01	15	6.6%	2.1E-02	10	6.5%	5.7E-02	19	5.5%	**1.9E-04**	60	9.3%	8.4E-02	172	8.6%	**1.7E-05**
Oxidoreductase	3.1%	16	2.5%	4.9E-01	4	1.8%	3.3E-01	2	1.3%	3.4E-01	10	2.9%	1.0E+00	17	2.6%	6.5E-01	49	2.4%	1.2E-01
Phosphatase	1.6%	8	1.3%	6.4E-01	7	3.1%	1.0E-01	4	2.6%	3.2E-01	3	0.9%	3.9E-01	12	1.9%	6.4E-01	34	1.7%	7.9E-01
Protease	2.5%	12	1.9%	3.7E-01	6	2.7%	8.3E-01	4	2.6%	8.0E-01	11	3.2%	3.9E-01	22	3.4%	1.3E-01	55	2.7%	4.7E-01
Receptor	8.2%	87	13.6%	**4.7E-06**	38	16.8%	**2.6E-05**	20	13.0%	3.9E-02	55	15.9%	**2.9E-06**	66	10.2%	7.2E-02	266	13.2%	**2.9E-14**
Select calcium binding protein	1.9%	16	2.5%	2.4E-01	7	3.1%	2.1E-01	6	3.9%	7.0E-02	8	2.3%	5.5E-01	13	2.0%	7.7E-01	50	2.5%	4.7E-02
Select regulatory molecule	6.3%	34	5.3%	3.3E-01	6	2.7%	1.9E-02	10	6.5%	8.7E-01	26	7.5%	3.8E-01	33	5.1%	2.3E-01	109	5.4%	9.9E-02
Signaling molecule	4.3%	27	4.2%	1.0E+00	11	4.9%	6.2E-01	8	5.2%	5.5E-01	21	6.1%	1.1E-01	44	6.8%	2.4E-03	111	5.5%	6.7E-03
Synthase and synthetase	0.9%	1	0.2%	3.6E-02	2	0.9%	1.0E+00	0	0.0%	6.5E-01	2	0.6%	7.8E-01	3	0.5%	3.0E-01	8	0.4%	1.3E-02
Transcription factor	9.1%	43	6.7%	3.9E-02	18	8.0%	6.4E-01	8	5.2%	1.2E-01	16	4.6%	2.6E-03	43	6.7%	3.3E-02	128	6.4%	**1.3E-05**
Transfer/carrier protein	1.5%	6	0.9%	3.2E-01	4	1.8%	5.8E-01	5	3.2%	8.0E-02	1	0.3%	7.1E-02	5	0.8%	1.9E-01	21	1.0%	1.2E-01
Transferase	4.6%	21	3.3%	1.3E-01	11	4.9%	7.5E-01	8	5.2%	7.0E-01	15	4.3%	1.0E+00	28	4.3%	8.5E-01	83	4.1%	3.6E-01
Transporter	3.4%	28	4.4%	1.5E-01	8	3.5%	8.5E-01	4	2.6%	8.2E-01	7	2.0%	2.3E-01	22	3.4%	9.1E-01	69	3.4%	8.5E-01

Table S6-3(D). "Molecular functions" gene ontological categories are enriched/depleted in nearby (+/-50 kb) L1 integrants.
P-values are calculated by binomial statistics. *Red highlights:* p<0.05; *boxed, bold red*: significant after Bonferroni correction.

a) *Alu* intronic inserts

lineage	sense	anti-sense	total	%sense	p-value	
human-specific	1,070	1,203	2,273	47.07%	5.6E-03	**
bonobo-specific	140	193	333	42.04%	4.3E-03	**
chimpanzee-specific	107	143	250	42.80%	2.7E-02	*
BC shared	80	78	158	50.63%	9.4E-01	
HBC-O	106	116	222	47.75%	5.5E-01	
total	1,503	1,733	3,236	46.45%	5.6E-05	**

b) L1 intronic inserts

lineage	sense	anti-sense	total	%sense	p-value	
human-specific	130	198	328	39.63%	2.1E-04	**
bonobo-specific	53	79	132	40.15%	2.9E-02	*
chimpanzee-specific	36	50	86	41.86%	1.6E-01	
BC shared	90	103	193	46.63%	3.9E-01	
HBC-O	106	177	283	37.46%	2.9E-05	**
total	415	607	1,022	40.61%	2.1E-09	**

Table S6-4: Relative counts and orientations of *Alu* and L1 retrotransposons in genes. (a) Relative orientation of *Alu* retrotransposons integrated inside of NCBI RefSeq genes; and (b) L1 retrotransposons integrated inside of genes. The statistical significance of departures from expected orientations that would occur by chance (i.e. 50% sense: 50% antisense orientation) was calculated using conventional binomial statistics. *BC*: present in bonobo and chimpanzee, but not present in human and orangutan; *HBC-O*: shared in human, bonobo, and chimpanzee, but not present in orangutan.

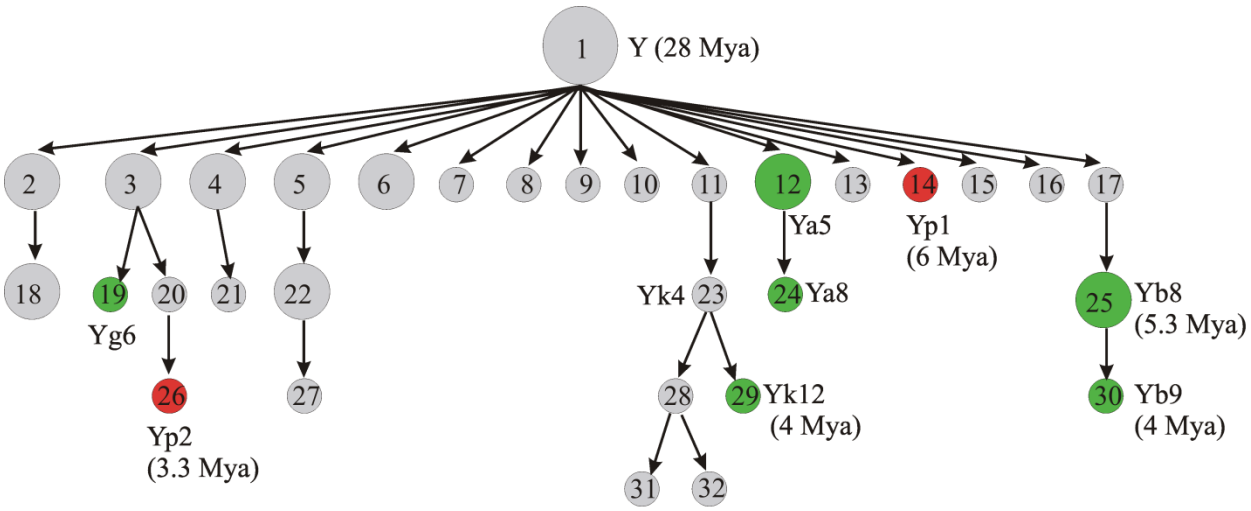


Figure S6-6: Identification of *Alu*Y subfamilies. The tree is based on an Alucode analysis of 65,036 *Alu* elements extracted from three primate genome assemblies (large nodes=subfamilies with more than 5000 elements; medium nodes=500 to 5000 elements; small nodes=less than 500 elements). Fifty-three *Alu* families were identified by Alucode (MINCOUNT=50), of which 45 subfamilies are shared between the three primates. From these, *Alu*Y subfamilies (i.e. diverged from *Alu*Y) are shown here. Twenty-four subfamilies are shared in three primates, human, bonobo and/or chimpanzee (*gray* nodes). Six subfamilies are specific in human (*green*). Two previously unreported subfamilies, which we designate as Yp1 and Yp2, are unique to the bonobo/chimpanzee shared lineage (*red*). Approximate ages were calculated by applying a constant scaling factor of 0.15% divergence from consensus per million years[64]. See Table S6-5 and Figure S6-7.

ID	P value	count	mutation rate	age (Mya)	repBase	bonobo	chimpanzee	human	lineage
1	0	36,143	0.042	28	**AluY**	8,763	13,103	14,277	BCH
2	5.00E-005	1,133	0.041	27.3	AluY (*)	292	382	459	BCH
3	6.00E-012	2,045	0.042	28	AluY (*)	497	708	840	BCH
4	9.00E-005	2,598	0.043	28.7	AluY (*)	636	902	1,060	BCH
5	9.00E-005	656	0.041	27.3	AluY (*)	163	243	250	BCH
6	1.00E-004	2,386	0.043	28.7	AluY (*)	577	815	994	BCH
7	4.00E-096	140	0.036	24	AluY (*)	33	55	52	BCH
8	2.00E-053	112	0.045	30	AluY (*)	29	42	41	BCH
9	1.00E-017	176	0.046	30.7	AluY (*)	43	56	77	BCH
10	4.00E-001	119	0.044	29.3	AluY (*)	31	48	40	BCH
11	3.00E-012	302	0.036	24	AluYk4 (*)	71	102	129	BCH
12	0	2,893	0.009	6	**AluYa5**	24	26	2,843	H
13	2.00E-291	277	0.014	9.3	AluY (*)	31	63	183	BCH
14	3.00E-159	173	0.009	6	**AluY_p1**	54	114	5	BC
15	5.00E-059	116	0.04	26.7	AluY (*)	29	35	52	BCH
16	3.00E-012	120	0.039	26	AluY (*)	31	40	49	BCH
17	1.00E-163	451	0.032	21.3	AluYh9 (*)	86	139	226	BCH
18	3.00E-007	973	0.045	30	AluY (*)	217	345	411	BCH
19	0	358	0.011	7.3	**AluYg6**	5	4	349	H
20	0	308	0.013	8.7	AluY (*)	75	152	81	BCH
21	1.00E-005	198	0.039	26	AluY (*)	36	66	96	BCH
22	0	735	0.028	18.7	AluYf4 (*)	125	180	430	BCH
23	0	364	0.036	24	**AluYk4**	105	114	145	BCH
24	2.00E-020	63	0.012	8	**AluYa8**	0	0	63	H
25	0	1,814	0.008	5.3	**AluYb8**	6	5	1,803	H
26	0	61	0.005	3.3	**AluY_p2**	24	37	0	BC
27	3.00E-252	74	0.039	26	AluYf5 (*)	16	29	29	BCH
28	2.00E-122	71	0.035	23.3	AluYk4 (*)	19	23	29	BCH
29	3.00E-070	64	0.006	4	**AluYk12**	0	1	63	H
30	0	174	0.006	4	**AluYb9**	0	0	174	H
31	1.00E-064	132	0.036	24	AluYk4 (*)	40	39	53	BCH
32	2.00E-147	110	0.014	9.3	AluY (*)	16	23	71	BCH

Table S6-5: *Alu*Y subfamilies identified in human, bonobo and/or chimpanzee assemblies. Listed here are the total count, the individual counts in each of the three primates, and p values for each of the 32 *Alu*Y subfamilies identified by Alucode. Eight subfamilies are present in Repbase. Two *Alu*Y subfamilies are new subfamilies specific to the bonobo/chimp shared lineage, named here as *Alu*Yp1 and *Alu*Yp2. The mutation rate is calculated as the average divergence of each subfamily from its consensus sequence. The approximate age was obtained by applying a constant scaling factor of 0.15% divergence from consensus per million years[64].

```
AluSx     GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACCTGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACATGGTGAAACCCCGTCTCTACT
AluY      ..........................................................--..........A.........T......T....C.....................
AluYa5    ..........................................................--..........A.........T..C...T..A.C.....................
AluYa8    ..........................................................--..........A.........T..C...T..A.C.....................
AluYg6    ..................................................A.......--..........A.........T......T....C.....................
AluYb8    ..........................................................T......T--..........A.........T......T....A.....................
AluYb9    ..........................................................T......T--..........A.........T......T....A.....................
AluYk4    ..........................................................--..........A.........T......T....C.....................
AluYk12   ..........................................................--..........A.........T......T....C.....................
AluY_p1   ..........................................................--..........A.........T...A..T....C.....................
AluY_p2   ..............................T.....................A.....--..........A.........T.T....T....C.............T......

AluSx     AAAAATACAAAAA-TTAGCCGGGCGTGGTGGCGCGCGCCTGTAATCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATCGCTTGAACCCGGGAGGCGGAGGTTGCAGTGAGCCGAGATCGC
AluY      .............A.....................G.........G.....................................G..G...................C.....................
AluYa5    .............A.............A......G.........G...........T.......................G..G...................C...................C.
AluYa8    ....C........A.-...........A......G.........G...T.......T.......................G..G...................C...................C.
AluYg6    .............A...........A...................G..........A.......................G..G...................C..........T..........
AluYb8    .............A...........CA......G.........G.....................................G..G............A......C..................T..
AluYb9    .............A............C.......G.........G...........G.......................G..G............A......C..................T..
AluYk4    .............A............C.......G.........G.....................................G..G...................C..................A..
AluYk12   .............A..................T.G.....................................G..A.......AA.........C..............G...A..
AluY_p1   .......G.....A.....................G.........G.....................................G..G.......A..........C...........T.......G
AluY_p2   ....T........A.............T.....G.........G...........T.......................G..G...................C.....................

AluSx     GCCACTGCACTCCA-------GCCTGGGCGACAGAGCGAGACTCCGTCTC
AluY      ...........................................................
AluYa5    ...........................................................
AluYa8    ...........................................................
AluYg6    ....................................A......................
AluYb8    .........G...GCAGTCCG.....................................
AluYb9    .........G...GCAGTCCG.....................................
AluYk4    .........G...G..................A.........................
AluYk12   .........G.............T.......A....T.....................
AluY_p1   ..............................A.....A.....................
AluY_p2   ......................A......A....C.......................
```

Figure S6-7: Definition and alignments of *Alu* subfamily consensus sequences. The consensus sequence for the *Alu*Sx subfamily is shown at the top. The dots represent the same nucleotides as the AluSx consensus sequence. Deletions are shown as dashes. The newer subfamilies of *Alu*Y contain all the mutations of the ancestral *Alu*Y as well as extra diagnostic sequence variants. *Alu* subfamilies classified as bonobo/chimp lineage-specific are named here as *Alu*Yp1 and *Alu*Yp2.

Supplementary Information 7

Positive Selection in the Chimpanzee Genome

Michael Lachmann* and Kay Prüfer*

Max-Planck Institute for Evolutionary Anthropology, Leipzig, Germany

* To whom correspondence should be addressed (lachmann@eva.mpg.de, pruefer@eva.mpg.de)

Positive selection on a new mutation will cause a classical selective sweep. That is, the population will experience a bottleneck of size one at this site or in other words: all lineages from this population will coalesce at the time of the selective sweep. Therefore, any outgroup to the population under investigation, will always fall outside of the variation of the ingroup (external). When bonobo is used as outgroup to chimpanzees lineages coalesce often in the common ancestor and therefore frequently coalesce first with the outgroup (internal). This difference in prevalence of internal and external regions in dependence of selection can be exploited: The stronger the footprint of a selective sweep, the longer we expect the stretches of purely external regions to be, i.e. the length, and therefore our power to detect selection, of the external regions depends on the strength of selection and the recombination rate. With this method we aim to detect selection on the chimpanzee lineage since the split from the most recent common ancestor with bonobos. However, other types of local diversity reducing selection, such as background selection, or selection in the ancestor shortly before the split, would leave a similar pattern.

We use the sequence data from 17 chimpanzees and one bonobo (Ulindi) to scan the genome for regions where bonobos fall outside the variation seen in chimpanzee. The test is implemented as a hidden Markov model assigning the most likely state (bonobo falling inside or outside chimpanzee diversity) based on whether Ulindi shows the derived or ancestral variant at chimpanzee polymorphic positions. We find several long regions where bonobo falls outside of the chimpanzee variation. We compile a list of these regions and test the genes overlapping these regions for enrichment of specific Gene Ontology categories.

Data and Pre-processing

For our test of selection on the chimpanzee lineage, we need (i) chimpanzee SNPs, (ii) one bonobo allele at the chimpanzee SNP sites, (ii) the ancestral state relative to chimpanzee and bonobo and (iv) an estimate of the local recombination rate.

We use the Illumina resequencing data of 16 chimpanzees and the sequence reads of the individual sequenced for the chimpanzee reference genome (Clint) to detect polymorphic positions in chimpanzees. All reads were mapped to the human genome (hg18) and processed as described in SI 5.

On top of the chimpanzee SNPs, we map the reads of the bonobo individual Ulindi and retain only those chimpanzee SNPs on autosomes that are covered by a high quality base from Ulindi. For each such position we took one randomly sampled read of sufficient quality from Ulindi. A total of 14.3 million SNPs remain after filtering.

To infer the ancestral state of the chimpanzee SNPs, we use the whole genome alignments of the macaque, orangutan, and the human genomes as outgroups. We require at least two out of the three outgroups to be informative, and for all genomes to agree on the base at the interrogated position. This base then determines the ancestral state. We retain 12 million chimpanzee SNPs for further analysis after this step.

In addition, our test needs to use a recombination map. Since no chimpanzee recombination map is available, we use the recently published recombination map of the human genome from Kong et al. [69]. We average over a length of one megabase on the human genome between any adjacent chimpanzee SNPs to approximate the recombination rate between two sites. (Recombination hotspots between human and chimpanzee are known to differ in location [70, 71]; however large scale recombination rates seem to be conserved between both species (Peter Donnelly, personal communication and [72])). Sites without information in the recombination map are excluded from the study, leaving us with a total of 10.3 million SNP positions to analyze.

SNP Quality and Frequency Influence How Often Bonobo Shows the Derived Variant

Some of the sites called as SNPs in chimpanzee are incorrect due to erroneous base calls in reads or due to the mismapping of reads to non-orthologous locations in the human genome. In order to test the influence of these factors, we divided our set into three classes: sites where the chimpanzee reference from Clint supports the derived variant, sites where Clint supports the ancestral variant and sites where no read from Clint aligns. If Illumina-sequenced chimpanzee individuals would match the quality of the Sanger sequenced Clint-data, we would expect no difference between the three classes. However, the bonobo shows the derived alleles more often at positions where Clint also carries the derived allele. The effect is particularly pronounced when the derived variant is only detected in one other chimpanzee individual and only one further individual covers the site: Bonobo is derived in 7.3% of these cases where Clint is ancestral, while it is derived in 13.4% of these cases where Clint is derived. This difference is brought about by a difference in power for alignment and different error rates of the sequences (see SI 10 for further tests). We use the state of Clint as a separate parameter in our model to adjust the fraction of observed derived variants in bonobo (ie: when Clint is derived, a higher fraction of Ulindi derived is expected).

The frequency of the derived variant in chimpanzees is proportional to the age of the variant. Therefore, high frequency derived variants are more likely to be shared with the bonobo. Figure S7.1 shows the fraction of sites where bonobo carries the derived variant by the number of chimpanzee individuals carrying the derived variant. As expected, we observe a positive dependence between the frequencies. We therefore use the frequency of the derived variant together with the number of chimpanzee individuals covering the site to refine our model.

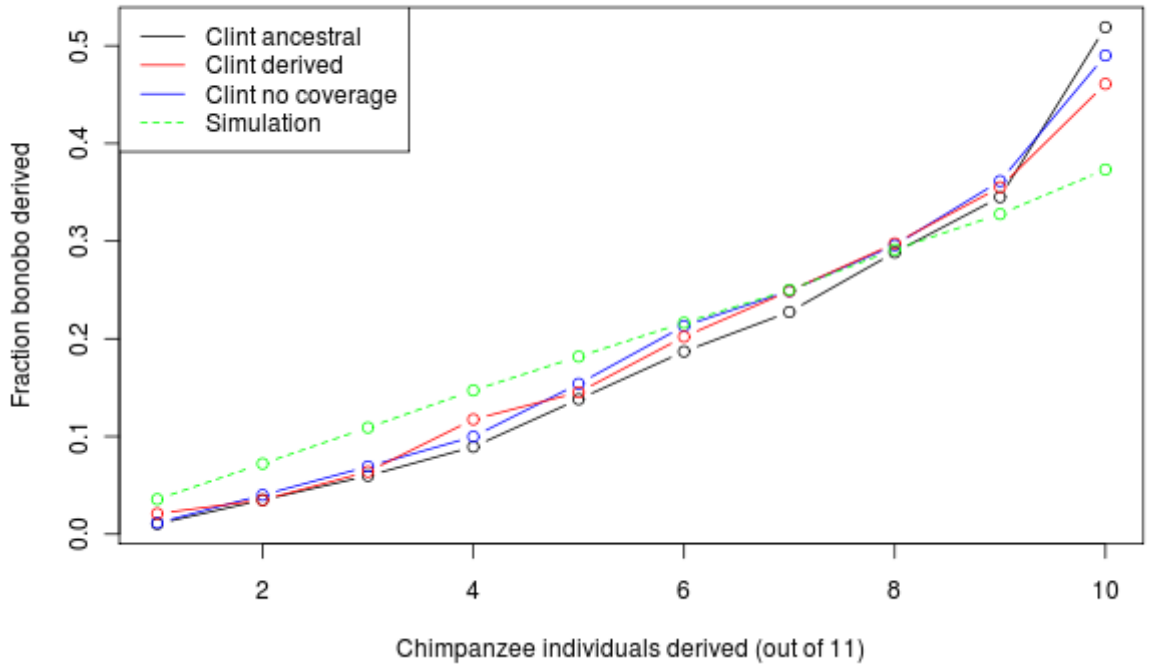


Figure S7.1: Fraction of sites where Ulindi is derived in dependence of number of chimpanzee individuals derived. Only sites with coverage from exactly 11 chimpanzee individuals are shown.

Coalescent Simulation to Infer the Ratio of Bonobo Internal/External and Linkage

We carry out coalescent simulations with the program *ms* [73] to approximate the expected fraction of the genome in which bonobo falls inside the variation of chimpanzees (internal) versus outside the variation of chimpanzee (external). The parameters of the simulation use the results from SI 8 for the ancestral population sizes (chimpanzee-bonobo: 27,000; chimpanzee-bonobo-human: 45,000) and split times (chimpanzee-bonobo: 1myr; chimpanzee-bonobo-human: 4.5myrs). In addition, we simulate with an uniform recombination rate and assume an effective population size of 10,000 for human and bonobo and 30,000 for chimpanzee. We do not model further chimpanzee substructure in our simulation. Despite this simplification, the simulation results closely match the observed fraction of derived sites in bonobo in dependence of frequency in chimpanzee (see Figure S7.1). By inspecting the coalescent trees along the simulated sequence, we observe that bonobo is internal for about 70% of the genome and external for about 30%. These values match the observed fraction of external and internal in a recently published study based on 15 ten kilobase regions[44].

We also analyzed the average length of regions that are internal versus external in our simulations. We observe that internal regions tend to be larger (average size 2.5 kilobases at average genome wide recombination rate) than external regions (average size 1 kilobase at average recombination rate).

Hidden Markov Model

We implement a hidden Markov model (HMM) to assign the hidden states *internal* and *external* to all SNP positions. For the HMM we need to specify emission probabilities and transition probabilities:

Emission probabilities

At each SNP position the model emits "*bonobo ancestral*" or "*bonobo derived*", with a probability that depends on the hidden state ("*internal*" or "*external*") and on the derived frequency at the site for Clint and for the other chimpanzees. Figure S7.3 shows a schematic description of the hidden Markov model. These emission probabilities were calculated as follows:

- Whenever bonobo is external to chimpanzee at a chimpanzee polymorphic site bonobo is expected to carry the ancestral variant. We might still see the derived variant in bonobo because of errors, mismappings or double mutations. Here, we assume a uniform background error rate of 1% for which bonobo may show the derived variant due to such problems.
- At internal sites, bonobo may show either the ancestral or the derived variant. In the model, the chance that bonobo has the derived allele is conditioned on the state of Clint, the frequency of the derived variant in chimpanzee, and the coverage by chimpanzee individuals. These conditional emission probabilities were measured from our data: We use the counts of observed bonobo ancestral and derived sites under these three parameters to calculate the emission probabilities at internal sites. We used the following formula:
  P(Derived)=P(Derived|external)P(external)+P(Derived|internal)P(internal), where in all cases we condition on a certain state of Clint and the other chimpanzees.
  If we assume P(Derived|external)=0, we have P(Derived|internal)=P(Derived)/P(internal). We took P(internal) from simulation, calculating the fraction internal conditioned on the number of individuals covering a site and the frequency of derived variants.

Transition probabilities

The transition probability is the chance that the next SNP has the same state (internal or external) as the previous one on the chromosome. This depends on the genetic distance between the sites, which we estimate from the smoothed recombination rates of a human map. We then approximate the decay of linkage, i.e the probability of staying internal or external near a site known to be internal or external (see Fig. S7.2), using an exponential distribution. The parameter λ of the exponential distribution is estimated from our simulation as 2500 for internal and 1000 for external regions. These values are scaled by the genetic distance between adjacent sites.

The model is implemented in C++ and uses the forward-backward algorithm and posterior decoding to assign states to the sites.

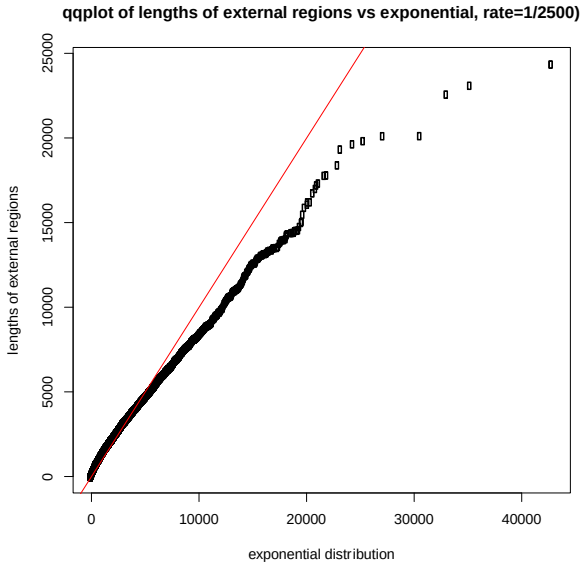


Figure S7.2: Match between exponential approximation and actual distribution of lengths of external regions.

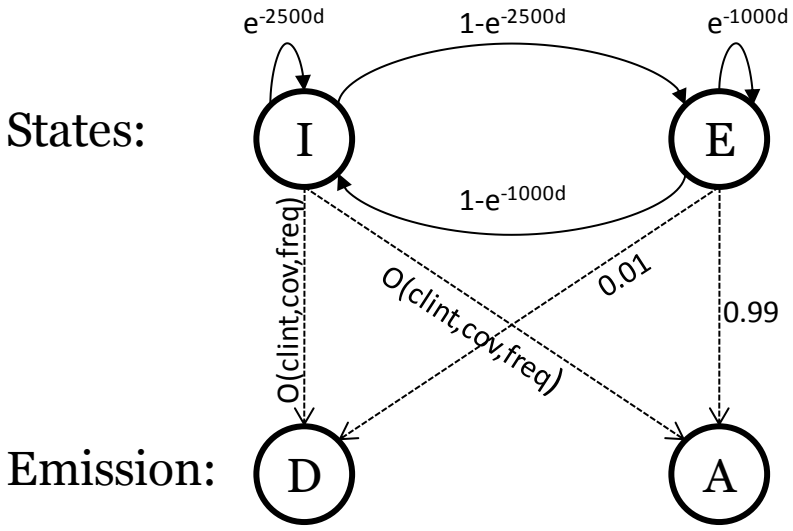


Figure S7.3: Schematic description of the hidden Markov model. Transitions (solid lines) between hidden states (top two circles) are dependent on the genetic distance *d* between adjacent SNPs. Emission probabilities are fixed for observing ancestral and derived from the external state. The emission probabilities for the internal state depend on the state of Clint, the coverage of individuals and the derived variants frequency in chimpanzee.

Scoring of Results

We run the HMM on the 10.6 million chimpanzee SNPs on all autosomes with parameters 2500 (1000) for the average internal (external) length, respectively. We retain only those SNPs with posterior probability of greater than 0.8 for external and internal. A total of 2.4 million SNPs were classified as internal, and 168000SNPs as external using this cutoff. From these SNPs we create a list of regions that contain solely

external calls (at least 2 adjacent external SNPs). The size of these regions is an indicator for the strength of selection acting on the regions. However, the size may in some instances be an artifact of a low number of SNPs in the region. To correct for this we limit the score of the region in dependence on the number SNPs and their distance to each other. We rescore each region in the following way: we calculate the 1 megabase average human recombination rate (*rr*) for the midpoint between any two adjacent SNPs in the region. Each pair of adjacent SNPs is assigned a value of 1000/*rr* if their physical distance exceeds this value, and the physical distance in basepairs otherwise. These values are summed over all pairs of SNPs and multiplied by the average recombination rate for the entire region. The scoring thus assigns a SNP-corrected value of genetic distance per region.

Because SNPs in lower recombining regions are truly informative for larger physical distance, these regions may tend to have higher scores. In agreement with this prediction,. we observe a negative correlation between scores and recombination rate (Pearson's r = -0.29; p-value < 2.2e-16).

The scoring is carried out with a smoothed human recombination rate since there is currently no recombination rate map for chimpanzee available. The smoothed human rate serves as an approximation and may introduce substantial error. An improved list can be generated once a recombination map for chimpanzee becomes available.

Dependency on Model Parameters

The ranking of regions according to score may depend on the choice of the parameter for the average genetic distance external and internal (2500 and 1000 according to simulation). In order to test the stability of the ranking of regions, we rerun our HMM with the parameters (2000, 1000) and (3000, 1000) for the average (internal, external) length. We then extract and score external regions as explained earlier. The resulting regions and their relative rank are then compared to the regions with parameter (2500,1000). We then compare the ranks for any overlapping regions from both runs of the HMM(Figures S7.4 and S7.5). High ranking regions are generally also high ranking when using a different set of parameters. Out of the top 100 regions found using parameter (2500,1000), 96 are also in the top 100 for (2000,1000) and 98 for (3000,1000). A run with parameter (1000,1000) identified 71 of the original regions. The scoring of regions is thus relatively stable and independent of the chosen parameters for the average length of internal and external regions.

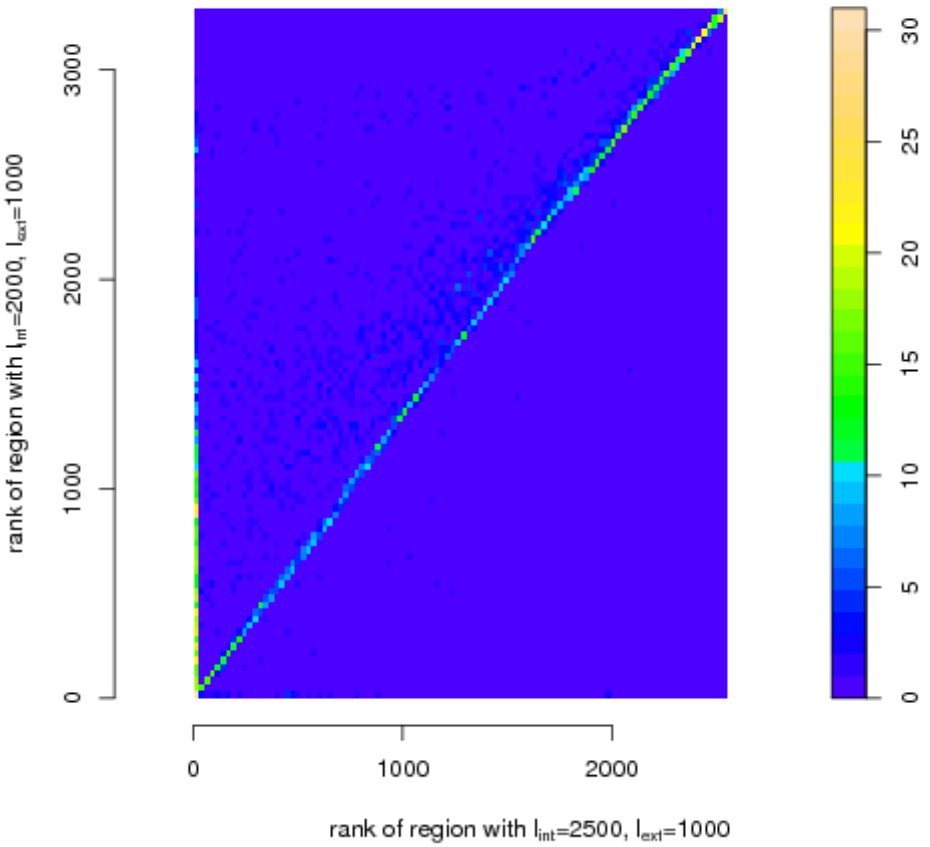


Figure S7.4: Overlap between ranks after scoring external regions. Shown are runs of the HMM with parameters (2500,1000) and (2000,1000) for the average length of (internal, external).

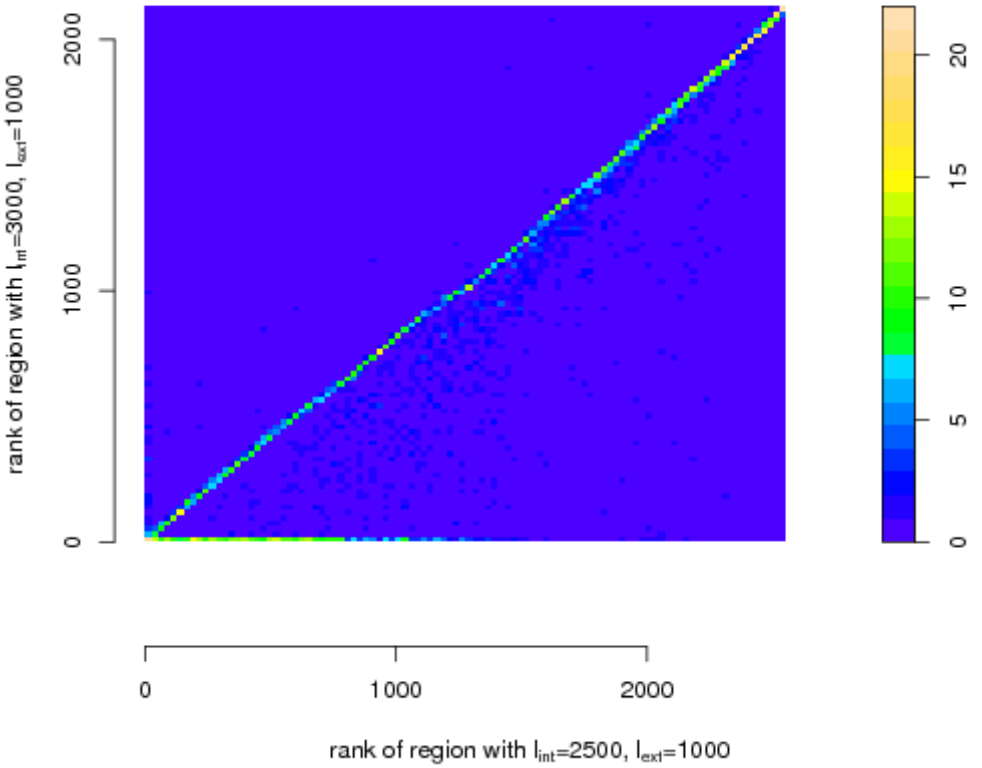


Figure S7.5: Overlap between ranks after scoring external regions. Shown are runs of the HMM with parameters (2500,1000) and (3000,1000) for the average length of (internal, external).

The Effect of Background Selection

Two different selective forces may contribute to the long external regions observed in our analysis: background selection and positive selection. Positive selection on the chimpanzee lineage, in the form of a selective sweep, coalesces all lineages. The recovery of diversity after this event will lead to regions that can be identified as external. Background selection, on the other hand, is a constant reduction in population size leading to a higher chance of earlier coalescence times.

In order to test for the contribution of background selection to our list of external regions, we annotate all regions with the average of the background selection value, *B*, corresponding to the amount of remaining diversity [74] in humans. Low *B* values indicate a low diversity due to the strong effects of negative selection. When we correlate the score for all regions with the *B* value, we find a negative correlation (Pearson's r = -0.26, p-value < 2.2e-16). In other words, the score increases with the expected amount of background selection.

Background selection is expected to have this effect. In places with much background selection, the effective population size is locally reduced, and therefore the ancestral tree of chimpanzees in the region coalesces earlier, so that Bonobo has a higher chance to be external. To find candidates for positive selection

we try to fit a distribution to the lengths of all non-positively selected regions for every B value. Outliers from this fit might have been positively selected, because they are even longer then their B value would predict. The distribution of external track lengths for a constant B, and thus constant effective population size, is expected to be approximately exponential. Since we are only interested in regions that are longer than expected, we are only interested in the right hand tail of the distribution, but for the fitting we leave out the extreme right-hand tail. Thus, to correct for the effect of background selection, we fit the 95%-99% tail of the distribution of lengths to the tail of an exponential distribution, so that for each region with length L, we assume that L*slope(B)+constant is from an exponential distribution. The reason for only using the 95%-99% of the tail, and not the full right-hand tail is that we only want to normalize out the effect of background selection, but not the effect of positive selection, and assume that less than 1% of the regions were under positive selection. The normalization will reorder the list of candidates, but is not meant to provide a real p-value against the hypothesis of only having undergone neutral evolution plus background selection.

In total, four parameters were used to specify this transformation (in the fit L*slope(B)+constant one parameter is the constant, and three parameters were used to fit the slope as a function of B). As can be seen in Figure S7.6, the resulting p-values show a good fit to an exponential distribution. The top 100 regions according to this corrected p-value are listed in Table S7.1.

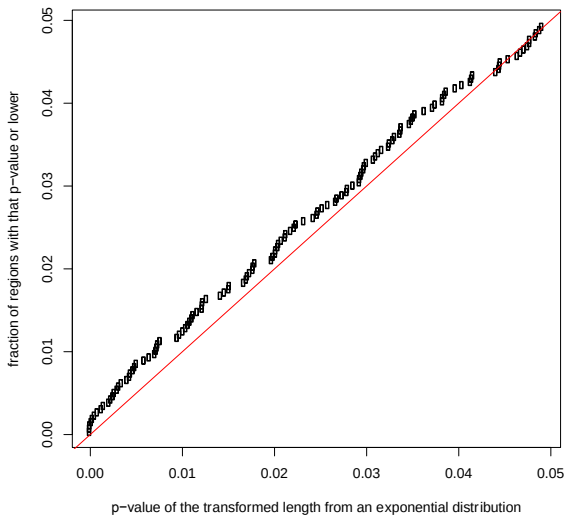


Figure S7.6: Cumulative distribution of p-values after B-dependent transformation of region lengths. The new p-value is taken from an exponential distribution.

Genes and Gene Ontology Analysis

We use the *Ensembl* [75] genome annotation (version 57) for the human genome to identify overlaps between external regions and genes (including exons and introns). We find a total of 759 genes that overlap external regions. If a gene overlaps with multiple external regions, we only associate it with the highest scoring one.

To address whether specific functional classes of genes have been particularly influenced by positive selection we test whether there are gene ontology categories that are enriched for high scoring genes using

FUNC [76] (wilcoxon rank test, minimum genes per category = 20). We find no significantly enriched categories. To demonstrate that a significant enrichment would be detectable with the given number of genes we also test for categories with low B values using the average B values for each external region. We find 31 significant categories (FWER < 0.05). The low number of genes tested is thus not the reason for the absence of a signal in the test for enrichment of positive selection.

Human chr.	Start (on hg18)	End (on hg18)	SRR	#SNPs	score	B	p
chr6	141009478	141136297	0.14	245	12855.9	894.3	0.00001
chr8	48724597	49109355	0.14	382	28492.5	244.1	0.00001
chr6	27571553	27807581	0.13	357	17113.2	583.7	0.00003
chr1	188865238	188954434	0.09	388	8364.0	927.6	0.00019
chr6	27253607	27390038	0.13	274	11762.4	646.8	0.00031
chr4	98715803	98800948	0.10	288	8101.7	841.9	0.00053
chr6	62264605	62423691	0.07	460	9230.2	705.0	0.00084
chr8	78899916	78971971	0.15	212	6727.4	874.4	0.00127
chr13	62776788	62804717	0.24	83	5688.6	986.8	0.00150
chr3	83595527	83663944	0.09	303	5882.9	908.1	0.00209
chr11	38950963	38983957	0.16	105	5416.0	960.5	0.00230
chr2	194512102	194580389	0.15	100	5538.3	925.5	0.00255
chr3	34602582	34629145	0.36	49	5574.7	912.6	0.00268
chr10	57997171	58060911	0.21	84	5015.0	981.1	0.00301
chr9	105440264	105463191	0.42	97	5534.6	891.2	0.00317
chr5	44407499	44474331	0.20	130	7239.6	693.7	0.00345
chr2	57904383	58003276	0.27	70	6838.5	707.3	0.00406
chr11	70589734	70594702	1.27	87	4932.1	929.0	0.00435
chr7	118907456	118980314	0.10	183	5424.5	851.3	0.00445
chr16	66027076	66175314	0.09	306	12370.3	35.7	0.00472
chr7	46502354	46507468	0.90	53	4538.8	976.0	0.00490
chr8	86377788	86460152	0.17	123	8663.6	531.6	0.00507
chr5	130092634	130180322	0.12	255	8553.9	516.4	0.00592
chr5	21119466	21142658	0.18	121	4282.3	969.7	0.00648
chr3	137759505	137901871	0.07	544	9568.1	394.4	0.00708
chr5	45619634	45729722	0.10	283	9040.9	440.6	0.00722
chr5	102123032	102185387	0.11	235	6765.9	629.5	0.00732
chr1	189524108	189584649	0.17	68	4301.5	938.1	0.00741
chr3	83512230	83576827	0.08	311	4984.7	819.0	0.00766
chr5	44790355	44880517	0.15	99	6953.6	572.0	0.00951
chr8	86688298	86723394	0.25	75	5303.4	733.6	0.00978
chr9	11901868	11907806	0.65	73	3831.2	966.5	0.01013
chr6	140939514	140982783	0.10	146	4150.9	894.3	0.01045
chr4	48624289	48764333	0.06	410	8867.4	360.3	0.01070
chr3	79726983	79774026	0.16	81	3949.5	924.5	0.01090
chr4	93527958	93549381	0.18	121	3810.2	949.6	0.01111
chr6	27944073	28079951	0.06	221	7726.7	471.7	0.01124
chr6	62777892	62824712	0.11	164	4916.4	751.0	0.01168
chr3	96471583	96502885	0.19	74	3956.6	897.2	0.01223
chr3	80109173	80170433	0.18	80	4073.3	874.4	0.01225
chr3	104037614	104051149	0.37	66	3865.9	914.5	0.01229
chr6	27419398	27476766	0.13	131	6206.3	592.1	0.01270
chr11	38861639	38885353	0.15	55	3484.0	965.0	0.01421
chr3	163931501	163970902	0.12	116	3582.9	934.1	0.01464
chr3	137557828	137722004	0.06	444	8290.1	316.6	0.01512
chr6	78591272	78615792	0.20	73	3820.5	874.8	0.01517
chr6	126847065	126881777	0.22	122	6458.6	514.3	0.01674
chr8	112597936	112629595	0.12	146	3783.5	855.0	0.01703
chr6	28241107	28337742	0.08	244	6827.5	469.9	0.01714
chr3	79925375	79946439	0.17	93	3530.2	903.1	0.01738

chr2	186518833	186623622	0.06	293	6539.9	493.1	0.01773
chr11	31243158	31342362	0.08	198	7441.9	386.2	0.01777
chr2	156489874	156500093	0.41	56	4186.1	771.6	0.01795
chr5	108008987	108021732	0.26	56	3288.7	926.3	0.01976
chr8	100056578	100320276	0.03	840	8281.7	69.9	0.01992
chr16	70790331	70868962	0.10	242	7897.1	228.8	0.02018
chr20	29394135	29469375	0.07	212	5025.6	631.7	0.02022
chr3	138078903	138155331	0.09	257	6952.6	407.4	0.02045
chr12	83057853	83080610	0.15	77	3181.0	943.0	0.02053
chr2	203856800	203995880	0.18	97	7989.4	154.2	0.02082
chr2	187592378	187688040	0.16	54	4696.1	662.5	0.02125
chr5	50460741	50484144	0.20	66	3919.4	777.4	0.02131
chr13	54613101	54632855	0.18	88	3628.5	825.4	0.02184
chr3	98365530	98384988	0.19	100	3681.5	810.7	0.02223
chr5	130336877	130450444	0.05	360	6183.9	476.9	0.02243
chr1	49828699	49909867	0.07	177	5238.8	577.5	0.02326
chr3	98120001	98168190	0.08	130	3654.2	794.1	0.02429
chr1	172159796	172417876	0.03	502	7399.9	213.1	0.02471
chr3	95319692	95368654	0.13	135	5080.8	582.3	0.02480
chr4	135473558	135490980	0.18	91	3091.9	907.2	0.02526
chr3	83434938	83483484	0.07	179	3176.4	879.5	0.02586
chr2	58426672	58442600	0.24	46	3387.4	822.9	0.02673
chr4	48457161	48520729	0.11	240	7149.2	224.5	0.02687
chr14	45386296	45404136	0.40	44	2998.4	907.6	0.02742
chr11	85005503	85091858	0.10	194	6879.0	279.9	0.02795
chr4	62318747	62343826	0.52	19	3268.3	836.1	0.02800
chr6	49370986	49387670	0.22	131	3628.6	758.0	0.02860
chr15	42073456	42197888	0.06	319	7038.0	167.9	0.02930
chr7	63355579	63370842	0.19	72	2888.6	917.0	0.02933
chr22	27132514	27219912	0.07	300	5986.5	423.5	0.02947
chr6	69748819	69761871	0.27	65	3551.4	763.8	0.02957
chr19	42130285	42203131	0.06	244	4311.3	640.1	0.02978
chr14	59908674	59983872	0.13	122	6029.4	412.9	0.02984
chr6	81442192	81470729	0.15	64	3036.2	870.7	0.03006
chr5	100079401	100106957	0.13	82	3308.9	801.1	0.03082
chr1	72551323	72560598	0.36	30	2980.2	875.1	0.03106
chr12	49254394	49322633	0.10	183	6762.1	214.4	0.03136
chr3	161941734	161997448	0.18	54	3592.3	738.2	0.03172
chr19	47722571	47748013	0.15	105	3697.0	713.8	0.03248
chr8	63564224	63576907	0.23	45	2870.9	889.5	0.03256
chr1	195103064	195144872	0.23	87	5351.3	477.6	0.03294
chr4	123944410	123963151	0.23	51	3532.6	738.0	0.03311
chr11	70576024	70581148	1.29	46	2694.2	929.0	0.03370
chr13	55739310	55762590	0.11	83	2638.8	945.0	0.03381
chr2	194136881	194194071	0.10	92	2837.3	886.6	0.03383
chr6	87905947	88042071	0.24	73	6366.3	263.5	0.03473
chr2	58102559	58126829	0.24	32	3572.5	716.7	0.03494
chr8	79027512	79049017	0.17	80	3009.2	831.2	0.03513
chr16	65724398	65905138	0.11	132	6764.4	28.1	0.03532
chr1	190094188	190102663	0.31	60	2590.1	935.8	0.03633

Table S7.1: Table of top 100 external regions according to background selection corrected p-value. Coordinates are given on the human genome (**hg18**). Column *SRR* gives the average human recombination rate [69] in a 1 megabase window centered at the midpoint of each region. Column *B* gives the average background selection value in the region. Column *p* gives the background selection corrected p-value.

Evidence for Positive Selection nearby the MHC

We find that 5 regions among the 50 highest scoring regions fall upstream of the MHC on chromosome 6 (Figure S7.7). The genes overlapping those regions are given in Table S7.2.

Region Rank	Genes (RefSeq)
3	LOC100507173
5	PRSS16
37	HIST1H3I, HIST1H4L, HIST1H3J, HIST1H2AM, HIST1H2BO, OR2B2, OR2B6
42	ZNF204P, ZNF391
49	TOP2P1, ZNF193, ZKSCAN4, NKAPL

Table S7.2: Genes overlapping the five high-scoring external regions on chromosome 6.

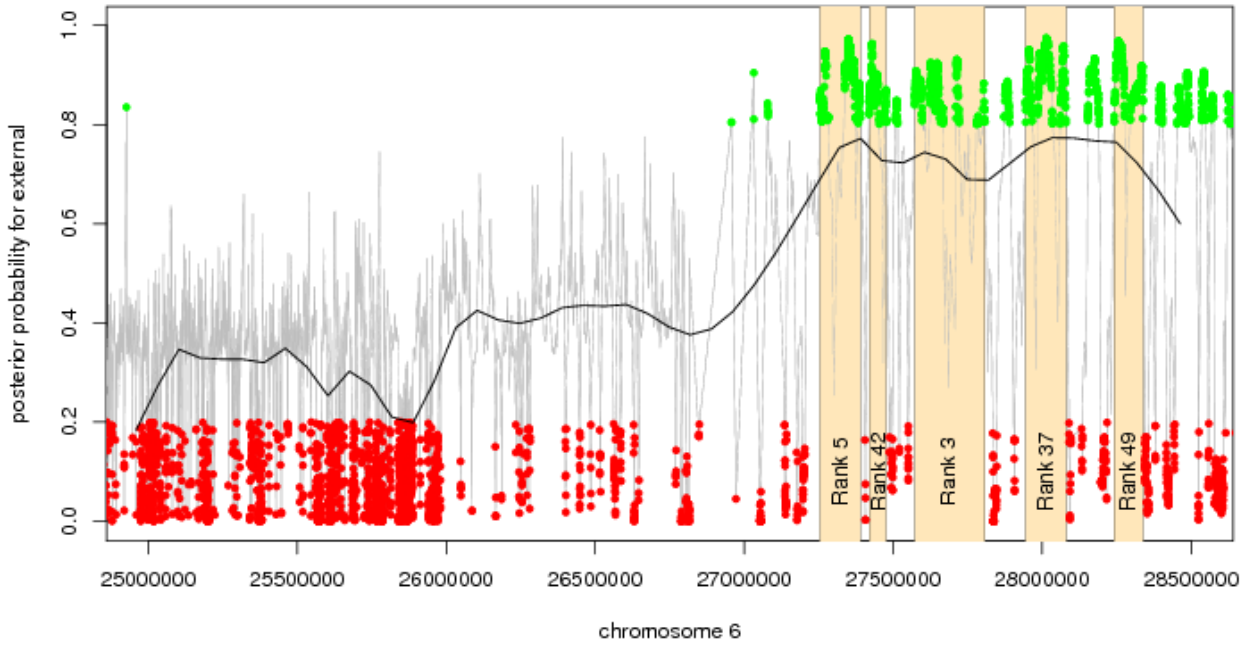


Figure S7.7: Posterior probabilities upstream of the MHC region on chromosome 6. The five external regions are shown in yellow. Gray lines give the posterior p for all positions; the black line is a smoothed curve on these values. Green points mark posterior p>0.8, red points p<0.2.

Supplementary Information 8

Speciation Times, Ancestral Population Sizes and Incomplete Lineage Sorting

Kasper Munch[1,*], Thomas Mailund[1], Kay Prüfer[2], Asger Hobolth[1], Julien Dutheil[1] and Mikkel H. Schierup[1,*]

1. Bioinformatics Research Center (BiRC), Aarhus University, Aarhus, Denmark
2. Max-Planck Institute for Evolutionary Anthropology, Leipzig, Germany

* To whom correspondence should be addressed (kaspermunch@birc.au.dk, mheide@birc.au.dk)

Along an alignment, divergence times between species differ due to segregating polymorphism in the ancestral species. For some species the population size of the ancestral species is sufficiently large and the time span between speciation events sufficiently small that ancestral polymorphism may lead to gene trees with a topology different from the species tree. This phenomenon is termed incomplete lineage sorting (ILS) and implies that segments of the genome will share an ancestor with a species other than its sister species in the species tree.

The CoalHMM framework [77] allows for inference of population genetic parameters and patterns of ILS. The framework is based on a hidden Markov model where the hidden states along the alignment represent gene trees with separate topologies and separate coalescent times. We apply the model to the full Bonobo genome to estimate the speciation times and the ancestral population sizes in the species tree of bonobo, chimpanzee and human. We demonstrate ILS between bonobo, chimpanzee and human and describe how the occurrence of ILS correlates with gene annotation and recombination rate.

Preparation of alignment data

The analysis is based on the four-way alignment of bonobo, chimpanzee, human and orangutan (HCBO, see SI 3). In this analysis the orangutan sequence serves only as outgroup. To establish the impact of sequence and assembly quality on the results of our analysis we performed a pilot analysis of chromosome 2, 21 and X to identify an appropriate filtering approach. We compared results of separate analyses based on raw un-filtered alignments as well as on alignments where called bases with phred scores [78] below cutoffs of 10, 30 and 50 were masked as 'N'. To evaluate the effect of filtering we compared the proportions of site patterns, the amount of predicted ILS and the estimated model parameters. We found that filtering had some effect on our results, especially on the X chromosome, but that using phred score cutoffs of 30 and 50 resulted in only marginal different results. To not remove more data than necessary we require a phred score of 30. Low complexity sequence regions are more prone to sequencing and assembly artifacts and were found to significantly affect the analysis. For this reason all regions annotated by the UCSC RepeatMasker track are masked as 'N'. Over-collapsing of regions due to duplications not recognized in the assembly stage will lead to a falsely inflated number of substitutions in such regions and lead to false predictions of Human-Chimpanzee ILS because Bonobo divergence is artificially increased. To this end a map of over-collapsed regions in the bonobo assembly (see SI 4) is used to remove such regions from the alignment. This filtering, however, only has negligible effects on overall results. Subsequent to all filtering procedures runs of more

than 100 masked positions are removed from the alignment.

The filtered four-way alignment used for the analysis consists of a large number of blocks each aligning a genomic segment from each species. Prior to analysis chunks separated by no more than 100 positions in bonobo, chimpanzee and human scaffold/chromosome coordinates are concatenated to form larger chunks of consecutive alignment. Only alignment chunks of minimum 500 base pairs are retained. An independent CoalHMM analysis is then run on each mega base of alignment chunks.

The CoalHMM method

The demographic model applied here is a three species isolation model (Figure S8.1). The parameters of the model are: three fixed extant population sizes, two ancestral population sizes, Ne1, Ne2, two speciation times, t1, t2 and a recombination rate. These parameters are all scaled with the substitution rate.

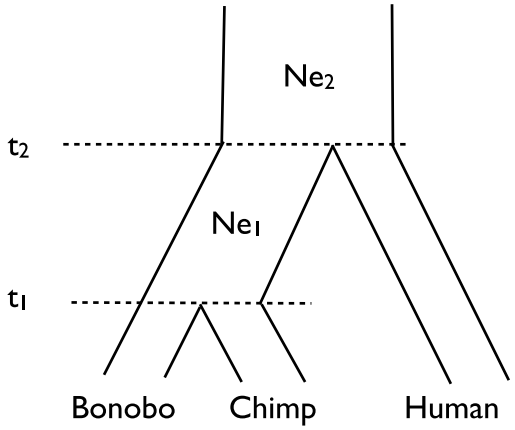


Figure S8.1: Isolation model used in the analysis. t1: speciation time of bonobo and chimpanzee. t2: speciation time of Human. Ne1: effective population size of the population size ancestral to bonobo and chimpanzee. Ne2 : effective population size of the ancestor to all three species.

The CoalHMM model applied operates with four different trees connecting three species: the bonobo and chimpanzee may find a common ancestor in their ancestral population (Figure S8.2 top left) or in the population ancestral to all three species (Figure S8.2 top right), and human may find a common ancestor with either the bonobo or the chimpanzee in the population ancestral to all three species (Figure S8.2 bottom left and right).

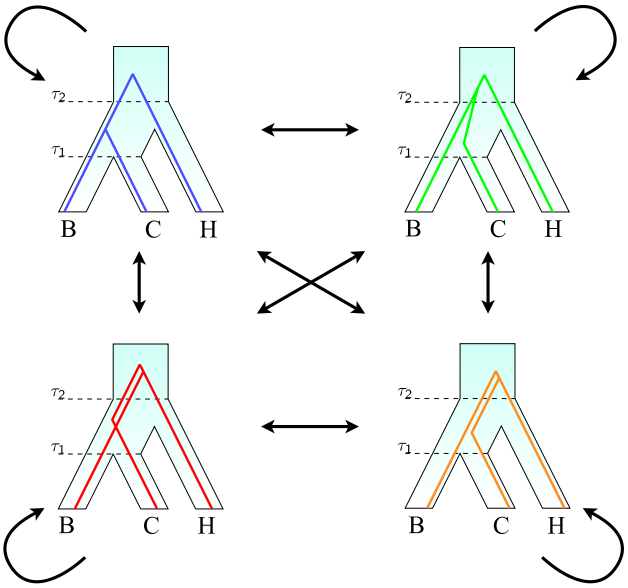


Figure S8.2: Outline of the four states in the HMM. The four states correspond to the four different trees describing the ancestry of an alignment column. Arrows indicate possible transitions.

The coalescent HMM method [77] used for this analysis employs a hidden Markov model with the four trees in figure S8.2 as hidden states. The transition matrix is parameterized using coalescent theory. The probability of emitting a column in the alignment from a state is calculated as the probability of the underlying tree given the four bases. The likelihood of the model given a set of model parameters is calculated using the forward algorithm. The maximum likelihood of the coalescent HMM is found using a modified Newton-Raphson algorithm.

Estimates of speciation times and ancestral population sizes

22% of analyses did not converge or resulted in radically deviating estimates. These were removed leaving 777 analyses. The raw parameter estimates are: Bonobo-Chimpanzee speciation time 0.31 Myr, Bonobo-Chimpanzee-Human speciation time 3.46 Myr, Bonobo-Chimpanzee population size 19.000, Bonobo-Chimpanzee-Human population size 60.000. These estimates are associated with a previously characterized bias [77]. The analysis tends to underestimate the size of the bonobo-chimpanzee population and overestimate the distance between the two speciation events. We predict and correct for bias in the estimation of model parameters by simulating data sets using a grid of relevant values of model parameters. Specifically, five alignments of 500 kbp are simulated for all combinations of the following model parameters: recombination rate 1.0e-8 per base per generation; mutation rate 1.0e-9 per year; generation time 20; Ne1: 2.0e+4, 4.0e+4, 6.0e+4, 8.0e+4; Ne2: 4.0e+4, 6.0e+4, 8.0e+4, 1.0e+5; t1: 2.0e+5, 5.0e+5, 1.0e+6, 1.5e+6; t2: 3.0e+6, 4.0e+6, 6.0e+6, 8.0e+6; divergence to orangutan out-group 17.0e+6. The CoalHMM analysis is then applied to each of these data sets to estimate model parameters. The bias on model paramerers in each analysis is computed as the deviation of the estimate from the true value used in the simulation. A linear model is fitted to explain bias from known values of parameters and their interaction.

The bias on each model parameter, estimated in each analysis of real data, is then predicted by the linear model and corrected for. The mean speciation times and effective population sizes over all corrected analyses are summarized in table S8.1 each estimate is rescaled assuming a generation time of 20 years and a per year mutation rate of 1e-9. Figure S8.3 shows the distribution of each estimate on autosomes. Figure S8.4 shows estimates across chromosomes.

Model parameter	Mean	Confidence
Bonobo-Chimpanzee speciation time	0.99 Myr	+/- 0.009 Myr
Bonobo-Chimpanzee-Human speciation time	4.50 Myr	+/- 0.04 Myr
Bonobo-Chimpanzee population size	27.000	400 +/-
Bonobo-Chimpanzee-Human population size	45.000	1100 +/-

Table S8.1: Rescaled parameter estimates using a generation time of 20 years and a mutation rate of 1e-9 per year. The 95% confidence interval is calculated as 1.96 times the standard error calculated using bootstrapping.

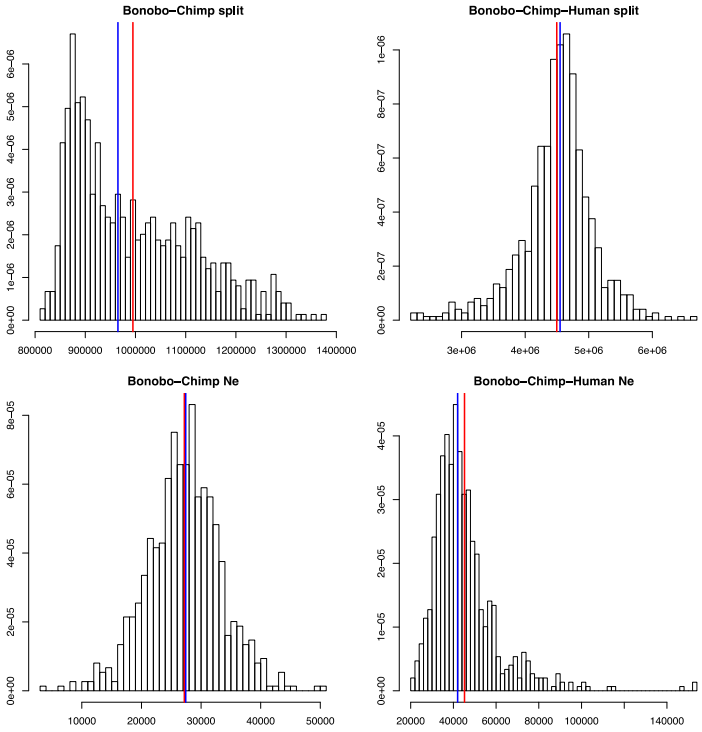


Figure S8.3: Distribution of estimates for 1Mb alignments. The large variances on the estimators reflect in part the variation in strength of forces such as recombination and selection. Red vertical line shows mean. Blue vertical line shows median.

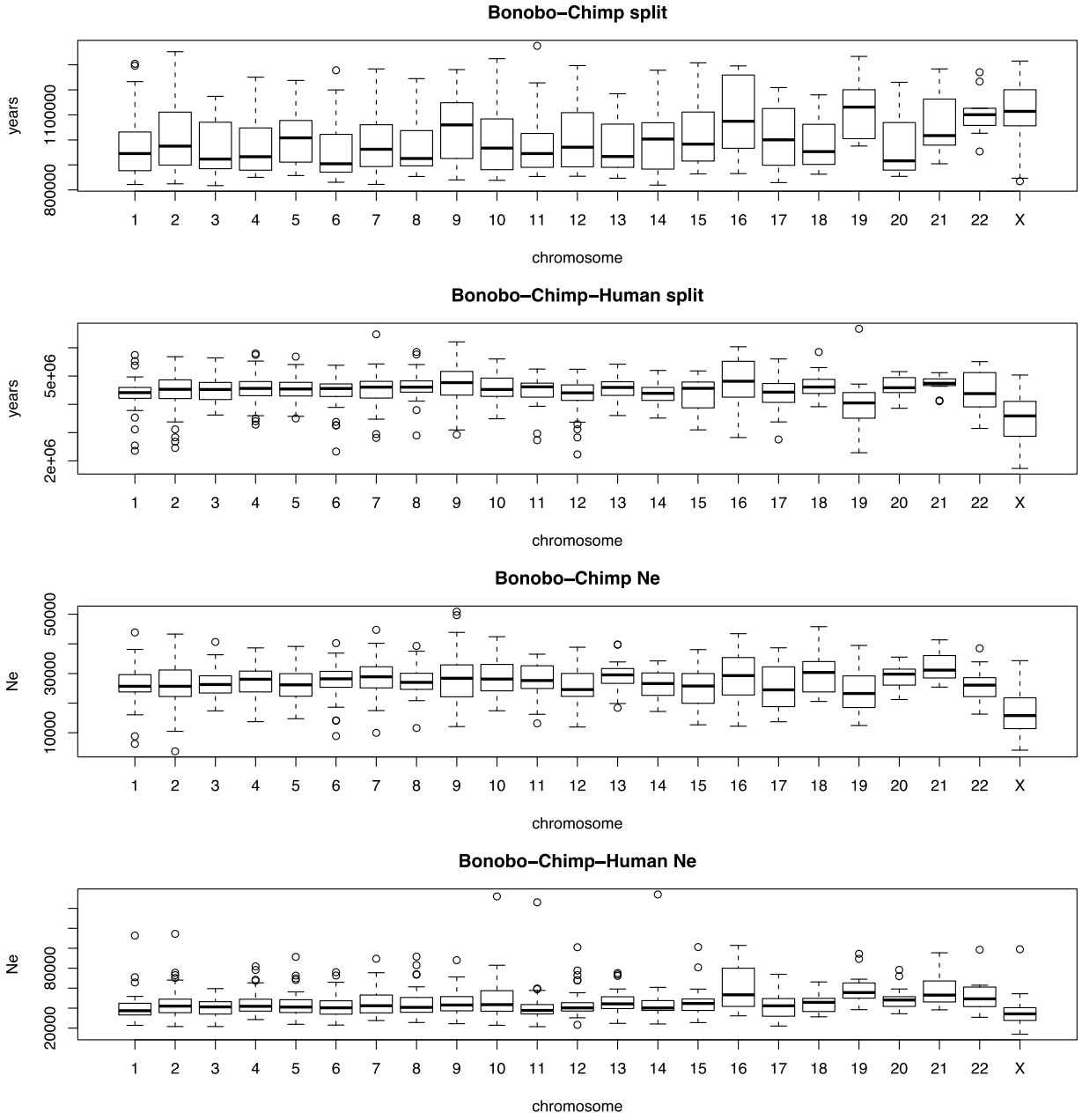


Figure S8.4: Distribution of estimates for individual chromosomes.

The rescaled estimates are linear functions of mutation rate and generation time. Alternative estimates of mutation rate or generation times may thus be applied to accommodate a different values of parameter estimates, e.g. using a mutation rate per year of only 0.6e-9 yields speciation times of 1.65 and 7.50, and population sizes of 45000 and 75000. Assuming that mutation rate per generation and generation time has remained constant across the phylogeny of the four species analyzed the relative values of split times and population sizes are not affected mutation rate and generation chosen. This assumption, however, may not be valid.

Prediction of incomplete lineage sorting

Posterior decoding is used to calculate the posterior probability of each state at each position in the alignment. The most probable state at each position is then identified as the predicted state. Prior to posterior decoding of each Mb of alignment the bias on model parameters for split times and ancestral population sizes are corrected and the recombination rate is re-optimized to accommodate this change. The mean proportion of analyzed autosomal alignment with bonobo-human most recent ancestry is 0.0157 +/- 0.0006. and the proportion of alignment with chimpanzee-human most recent ancestry is 0.0167 +/- 0.0006. The proportion of alignment with bonobo-chimpanzee most recent ancestry after the bonobo-chimpanzee-human split is 0.0180 +/- 0.0008.

The observed agreement between these proportions is expected as the three topologies are equally probable given that the bonobo and chimpanzee lineages have not coalesced before the bonobo-chimpanzee-human split. Using the bias-corrected model parameters to calculate the theoretically expected proportions of each type of ILS (exp[-(t2-t1)/(20*2*Ne1)]/3) yields 0.014 +/- 0.0005. We note that this expectation is slightly lower than the observed proportions. However, a reduction in the time between speciations by just 150.000 years *or* and an increase in chimpanzee-bonobo Ne by 3000 is sufficient to reconcile the theoretically expected and observed proportions of ILS. Together, these observations indicate internal consistency of the model, and support the validity of our findings.

Figure S8.5 shows the correlation between the predicted and theoretically expected proportion of ILS on chromosome one. Figure S8.6 shows the observed proportion of sites in each of the four HMM states. Table S8.2 shows the abundance of ILS-informative sites in each of the four site-classes.

	CH-sites	BH-sites	sites
CB	70718	64963	7,92E+08
CB2	4198	4176	14684330
BH	3166	33645	12942453
CH	33879	3187	13756464

Table S8.2: Abundance of ILS-informative sites in regions classified as CB, CB2, BH and CH. Informative-sites were counted on the HCBO alignments (see SI3). CH-sites were required to show the same base for chimpanzee-human and bonobo-orangutan. BH-informative-sites were required to show the same base for bonobo-human and chimpanzee-orangutan.

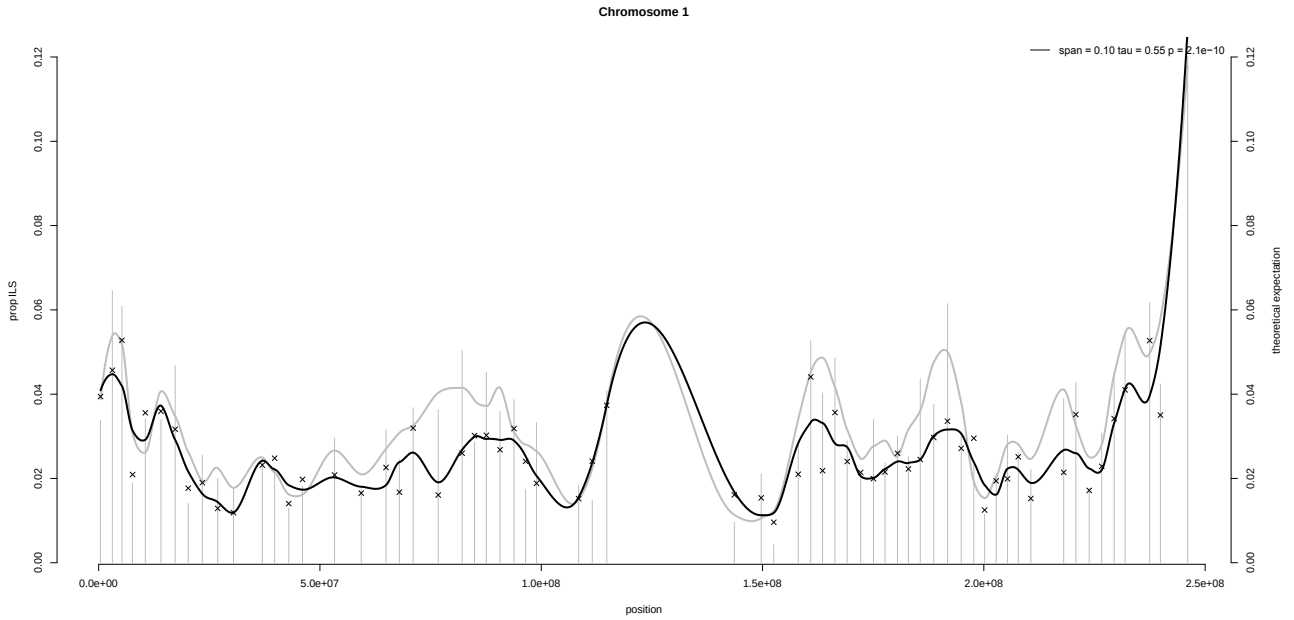


Figure S8.5: Correlation between predicted and theoretically expected proportion of ILS in each mega base of alignment. Gray bars and left y-axis shows predicted proportion of ILS. Black crosses and right y-axis shows the theoretically expected proportion. Solid lines are loess-curves with span chosen to maximize the correlation.

X / autosome ratio of ancestral population size

Calculating the ratio of X to autosome population sizes using the mean estimates of Nx and Na yields a ratio of 0.62 +/- 0.11. This does not take into account any difference in substitution rate between X and autosomes. One way to achieve this is to rescale population sizes with estimated bonobo-orangutan divergence. This yields a ratio close to the expected ¾: 0.76 +/- 0.10. Alternatively, the effective population sizes can also be estimated indirectly from the mean speciation times on autosomes and the observed amount of ILS in each analysis: -((T_12 – T_1)/generation_time)/(2 * (log(ILS_proportion) - log(2/3))). Using this approach we obtain a very similar ratio of 0.77 +/- 0.07 indicating internal consistency in model estimates and predictions of ILS.

The strength of the male mutation bias may have changed along the bonobo-orangutan lineage. If so, the average bias on the lineage from bonobo only back to the bonobo-human most recent common ancestor may be a more relevant correction. To address this we counted, for each 1Mb alignment analyzed, the number of substitutions on the branch (using orangutan as outgroup). Rescaling population sizes using this divergence yields a ratio of 0.83 +/- 0.09. This ratio is in agreement with the one found in bonobos (see SI 9). The rescaling is potentially also affected by a increased or reduced divergence on the X chromosome resulting from from a Nx/Na different from 0.75 in the bonobo-human ancestor. To investigate this we removed the substitutions attributed to divergence in the bonobo-human ancestor by scaling each substitution count with $\tau_{HC}/(\tau_{HC} + \theta_{HC}/2)$, where θ_{HC} and τ_{HC} are the estimated theta of the bonobo-chimpanzee-human population and the scaled estimated number of substitutions back to the bonobo-chimpanzee-human split, respectively. The remaining counts now represent the divergence back to the bonobo-human split i.e the

bonobo branch plus the Pan branch. This adjustment yields an almost identical Nx/Na ratio of 0.83 +/- 0.08 suggesting that the rescaling is not biased by Nx/Na in the bonobo-human ancestor. It should be noted, however, that the male mutation bias may have increased along this lineage. If the bias is higher in bonobos than in the bonobo-chimpanzee ancestor the re-scaling applied here will overcompensate for male mutation bias resulting in an inflated Nx/Na for the bonobo-chimpanzee ancestor.

As discussed in SI 9 the Nx/Na ratio is confounded with a number of effects in addition to male mutation bias. These include sex-specific reproductive variance and and migration behavior as well as selection potentially reducing variation on the X chromosome more effectively.

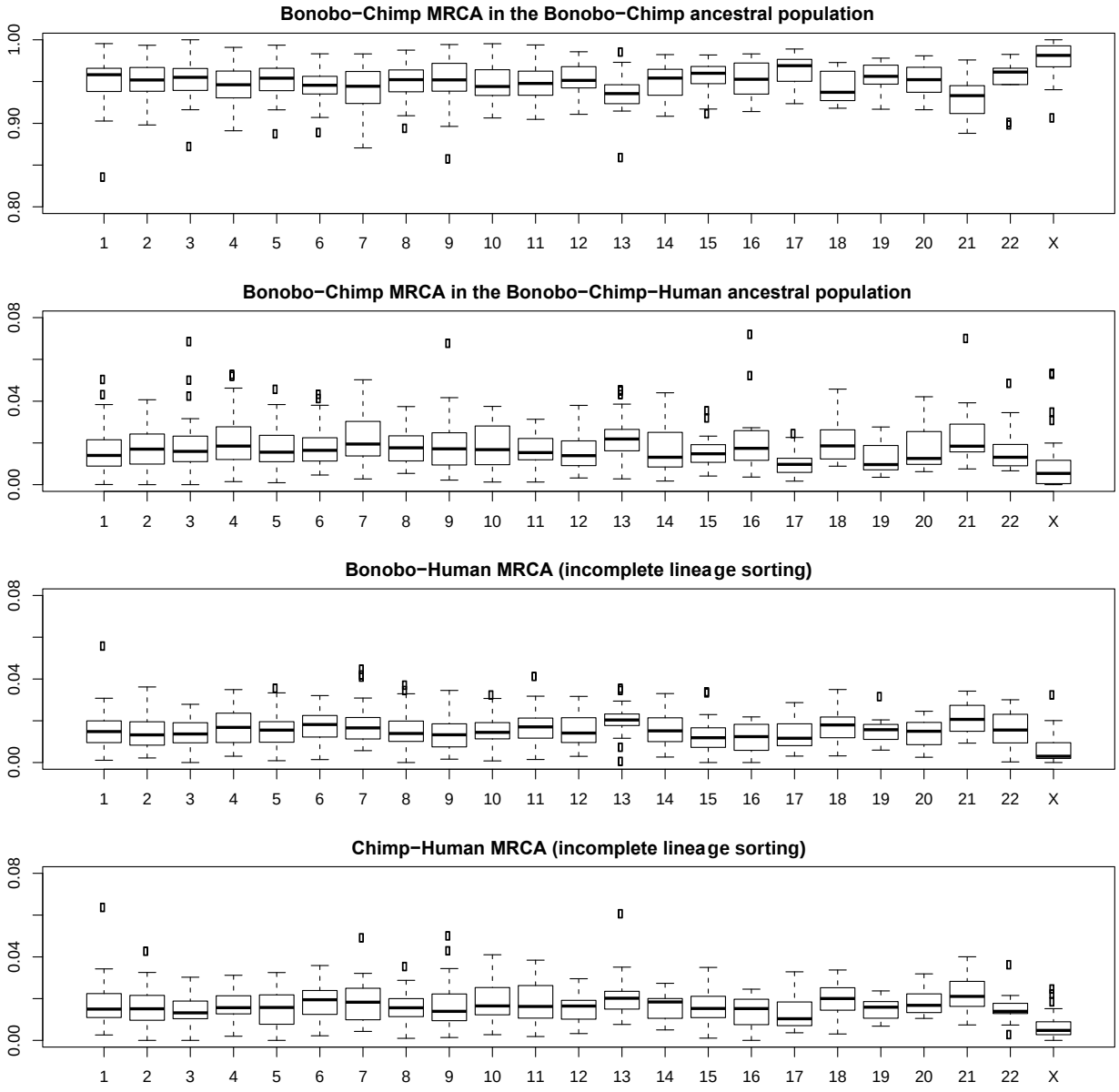


Figure S8.6: Distribution of observed proportion of sites in each of the four HMM states.

Correlation of ILS and human recombination rate

We downloaded the deCODE human recombination rate track from UCSC and plotted the proportion of ILS against the sex-averaged recombination rate for each mega base of analyzed alignment. The proportion of ILS is correlated with recombination rate (p-value < 4.3e-13) with a Pearson correlation coefficient of 0.16. A linear model fitted has a highly significant slope consistent with the notion that a higher recombination rate is inversely correlated with the effect of background selection reducing the ancestral population size (Figure S8.7). The grey interval in Figure S8.7 outlines the data points with recombination rates between the 2.5 and 97.5 percentiles. Over this span of recombination rates the proportion of ILS decreases from 0.040 at 3.6 cM/Mb to 0.027at 0.1 cM/Mb. This corresponds to a reduction in Ne of 12%. Higher recombination thus indirectly correlates with higher ancestral population size, which in turn increases the expected proportion of ILS. The correlation is also significant on individual chromosomes.

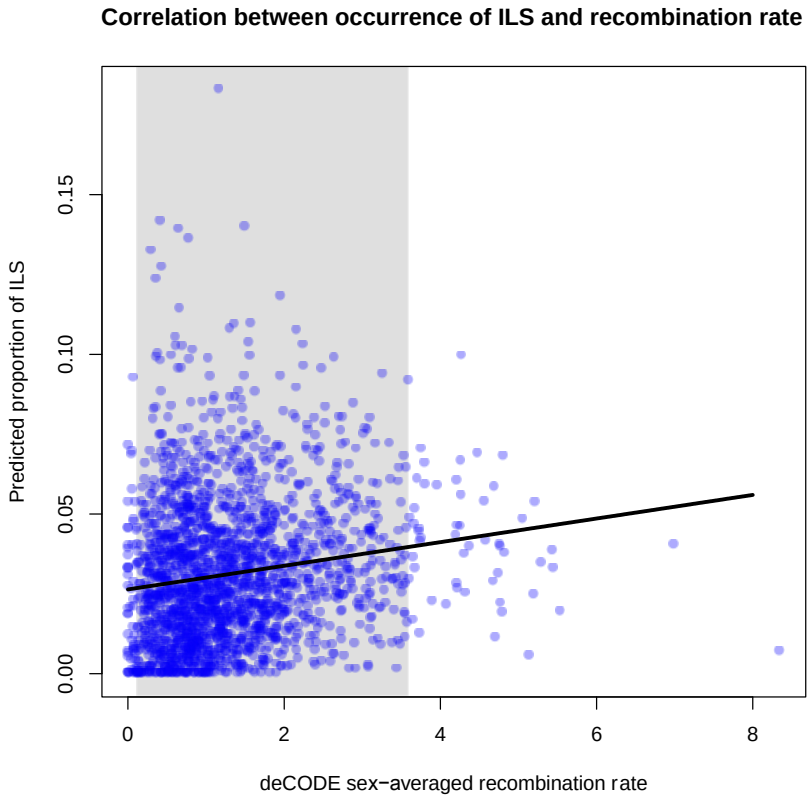


Figure S8.7: Correlation between Human recombination rate and proportion of ILS in observed in individual analyses of 1Mb of alignment.

Correlation of ILS and human gene annotation

We downloaded the knownGenes annotation track from UCSC and computed the fraction of exons and introns predicted as ILS regions. Figure S8.8 shows these fractions grouped by chromosome. The solid lines show a linear model for the proportion of sequence in exons and introns predicted as ILS regions, dashed lines outline the confidence interval computed from the linear model. The significantly smaller overlap to

exons is expected because background selection on exons reduce the bonobo-chimpanzee ancestral effective population size, which in turn lowers the probability of ILS. The difference in proportion of ILS predicted in autosomal exons and introns corresponds to a reduction in Ne of 12% (8%-15%). We do not see a stronger effect on the X chromosome. Calculating the Nx/Na ratio from the proportion of ILS and speciation times (see above) in all analyzed sequence and in exons yields 0.77 +/- 0.07 and 0.73 +/- 0.14 respectively.

Both overlap to exons and introns are (although not significantly so) inversely correlated with chromosome size. Considering the relation between ILS and population size this trend likely stems from the fact that the smaller recombination rate of larger chromosomes gives rise to a larger effect of background selection on the effective population size.

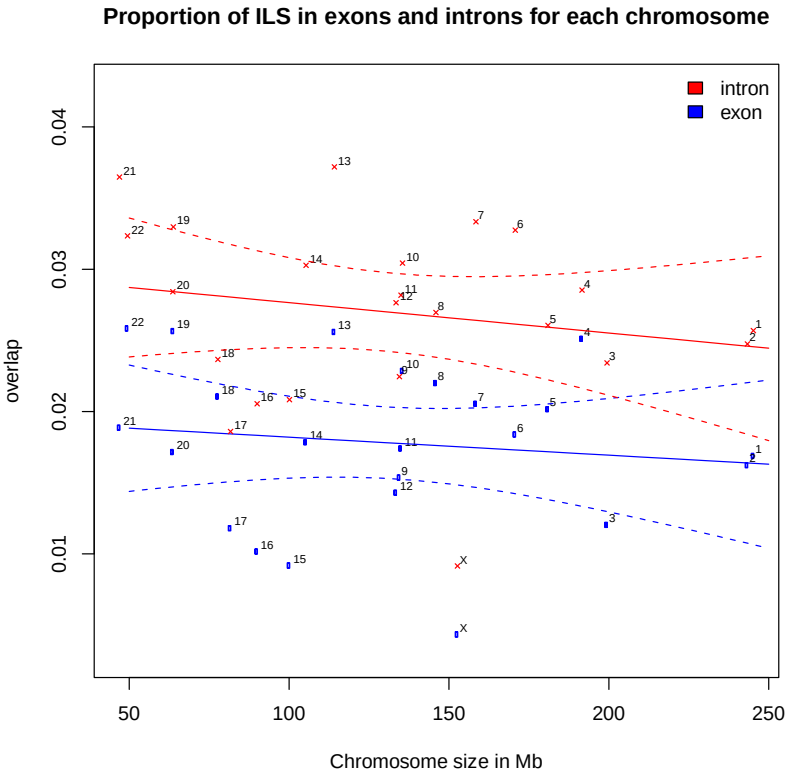


Figure S8.8: The proportion of exon and intron sequence that is predicted as ILS regions is plotted for each chromosome.

Wide ILS-void region on chromosome three suggests strong impact of selection

Since the probability of observing ILS at a genomic position is a function of Ne (in addition to the speciation times), long regions without any ILS may be suggestive of a stronger impact of selection locally reducing Ne. Figure S8.9 shows the length distribution of tracts on the autosomes completely void of ILS in the analyzed regions. The distribution has an exponential-like decay, which is expected if the process of state-change over the alignment is approximately Markov. One extreme outlier is observed, which corresponds to a 6.1Mb region (47.426.275-53.553.471) on chromosome three. This is almost twice the length of the second longest tract of 3.2Mb. The region, shown in Figure S8.10, likely has a very low recombination rate with an

average sex-averaged recombination rate in humans of 0.16 cM/Mb. In addition the region contains a cluster of immunity related genes reported to be under positive selection in humans [79] as well as the CCR5 gene involved in HIV resistance [80]. Both observations support the hypothesis that the depletion of ILS is a result of a selection depressing the local effective population size through hitchhiking and/or background selection.

We further characterize the longest ILS-void region by estimating diversity among the Illumina-sequenced chimpanzee and bonobos, and the divergence between chimpanzee and bonobo. Data was processed as described for SI 5 and a high quality base was sampled for each individual if more than one read covered a position. Diversity was estimated in 50kb windows as number of differences in all pairwise comparisons divided by total bases compared. Divergence between bonobo and chimpanzee was estimated in 50kb windows as the number of differences in all pairwise comparisons between bonobo and chimpanzee individuals. Both diversity and divergence estimates were then normalized by the divergence between human and orangutan to correct for differences in mutations rates along the genome. Figure S8.11 shows the results of the comparison. When comparing the ILS-void region with surrounding sequence, we observe a drop in bonobo-chimpanzee divergence in the region, as expected by the reduction of long coalescence trees due to the absence of incomplete lineage sorting. Interestingly, a drop in diversity is also visible in chimpanzee, while bonobos show no reduction in diversity.

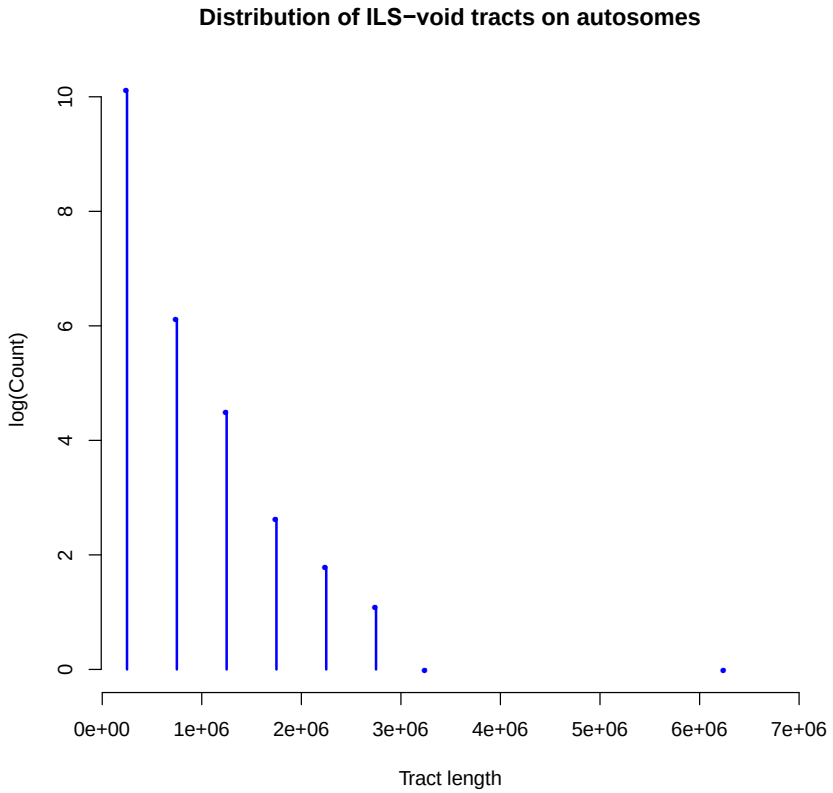


Figure S8.9: The length distribution of genomic tracts void of ILS predictions. Notice the log-scale on the second axis. The outlier corresponds to a 6.1Mb tract on chromosome three.

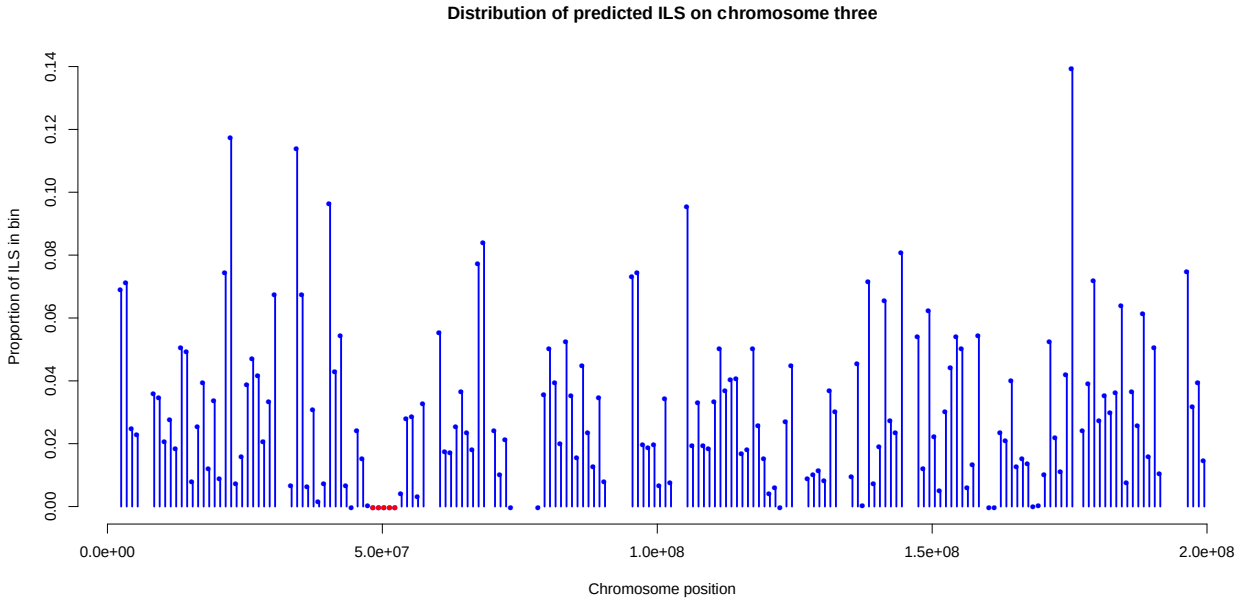


Figure S8.10: Proportion of ILS predicted in 1Mb windows across chromosome three. Windows with red dots outline the 6.1Mb region identified in Figure S8.9.

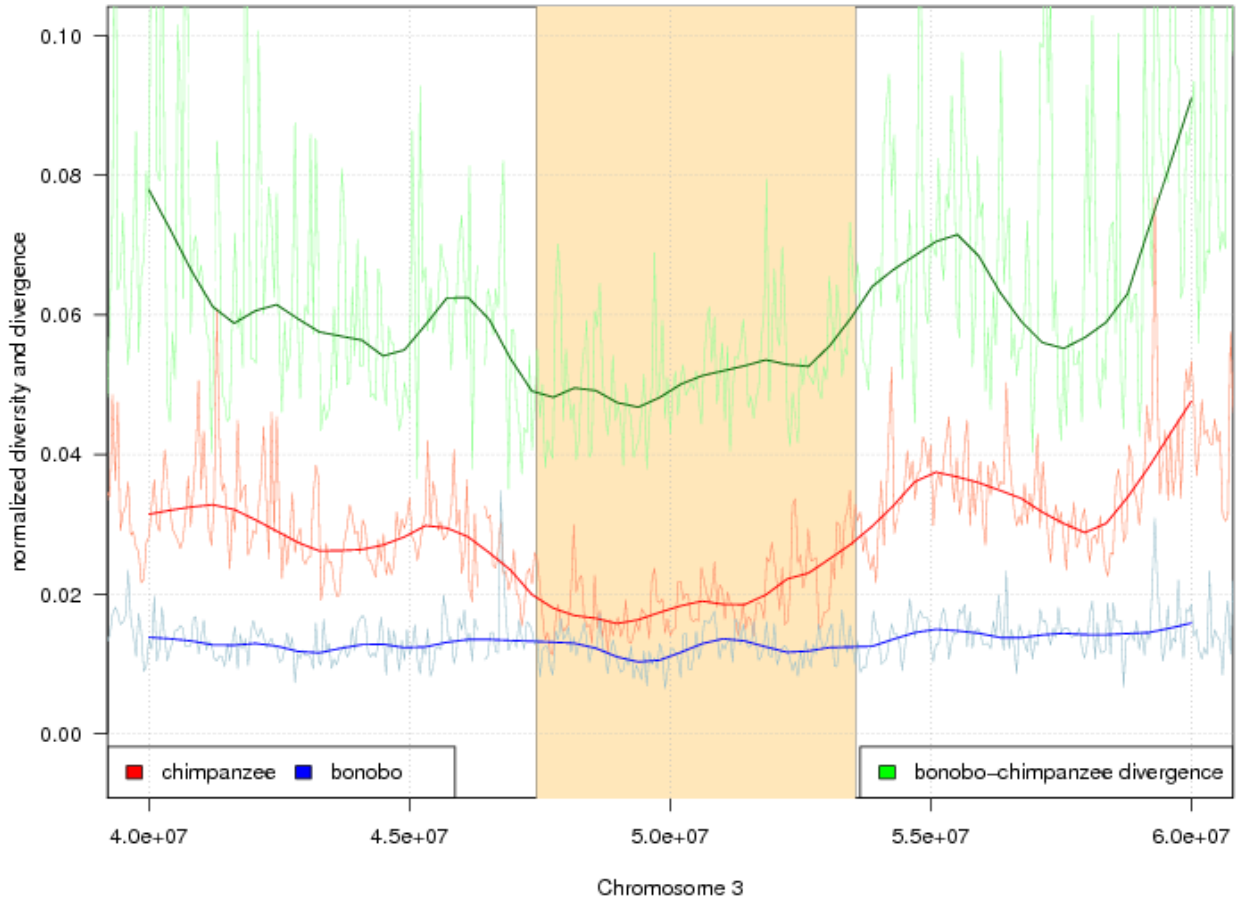


Figure S8.11: Bonobo and chimpanzee diversity, and bonobo-chimpanzee divergence normalized by human-orangutan divergence for a part of chromosome 3 include the longest ILS-void region. Solid thick lines give a smoothed polynomial curve fit to the original data shown as thin lines. The ILS-void region is shown as box in the middle of the plot.

Correlation between ILS and gene ontology classes

We use the gene annotation from the UCSC genome browser (release hg18) [25] to count the bases in each of the four ILS states for the entire length of genes including introns. The ILS-assignment excludes human annotated repetitive sequences (simple repeats and transposons provided as mask by the UCSC genome browser).

We carry out two gene ontology enrichment tests using FUNC [76]. We first test for enrichment of genes with a high fraction of ILS states. For this, we calculate the number of bases in the ILS states CH and BH over all bases annotated by the CoalHMM. We then use the Wilcoxon rank test to identify gene ontology categories that are either enriched or depleted for bases predicted as ILS (see Tables S8.3 and S8.4 for all categories with a family wise error rate (FWER) < 0.01).

We further our analysis by testing whether the result differs when short genes (which are expected to have a wider spread in %ILS due to stochasticity) are excluded. For this we exclude all genes with less than 20 kilobases assigned by the CoalHMM and retain 3412 of the original 9942 genes with a GO annotation. Tables S8.5 and S8.6 show the significant categories with FWER < 0.01. The restricted set yields fewer significant categories, but similar categories are found.

Genes depleted in ILS are often located intracellular and involved in transcription or translation, among other processes. Genes with a high fraction of ILS tend to encode for proteins integral to the membrane that are responsible for cell adhesion. The genes with low fraction ILS may be primarily genes evolving under strong purifying selection, while genes with a high fraction of ILS may either evolve under a relax of constraint or may be false positives due to assembly artifacts around highly duplicated gene classes.

The second test analyses genes that contain either a stretch of CH assignment (CH genes) or a stretch of BH assignment (BH genes), but not both. A total of 1280 genes are either BH genes or CH genes and have an associated Gene Ontology annotation. We then use a hypergeometric test to search for categories that contain either more CH genes than expected or more BH genes than expected. We find no significant enrichment (FDR<0.05) for genes that contain solely BH or CH states. This result is compatible with ILS being a random process that is not expected to preferentially sort specific classes of genes to lineages.

Taxonomy	GO-Category Name	GO-Id	FWER
biological_process	biosynthetic process	GO:0009058	<0.0001
biological_process	cellular biosynthetic process	GO:0044249	<0.0001
biological_process	cellular macromolecule biosynthetic process	GO:0034645	<0.0001
biological_process	cellular macromolecule metabolic process	GO:0044260	<0.0001
biological_process	cellular metabolic process	GO:0044237	<0.0001
biological_process	gene expression	GO:0010467	<0.0001
biological_process	macromolecule biosynthetic process	GO:0009059	<0.0001
biological_process	macromolecule metabolic process	GO:0043170	<0.0001
biological_process	nitrogen compound metabolic process	GO:0006807	<0.0001
biological_process	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	GO:0006139	<0.0001
biological_process	regulation of macromolecule metabolic process	GO:0060255	<0.0001
biological_process	RNA metabolic process	GO:0016070	<0.0001
biological_process	translation	GO:0006412	<0.0001

cellular_component	chromosomal part	GO:0044427	<0.0001
cellular_component	chromosome	GO:0005694	<0.0001
cellular_component	intracellular	GO:0005622	<0.0001
cellular_component	intracellular membrane-bounded organelle	GO:0043231	<0.0001
cellular_component	intracellular organelle	GO:0043229	<0.0001
cellular_component	intracellular organelle part	GO:0044446	<0.0001
cellular_component	intracellular part	GO:0044424	<0.0001
cellular_component	macromolecular complex	GO:0032991	<0.0001
cellular_component	membrane-bounded organelle	GO:0043227	<0.0001
cellular_component	nuclear part	GO:0044428	<0.0001
cellular_component	nucleus	GO:0005634	<0.0001
cellular_component	organelle	GO:0043226	<0.0001
cellular_component	organelle part	GO:0044422	<0.0001
cellular_component	ribonucleoprotein complex	GO:0030529	<0.0001
molecular_function	DNA binding	GO:0003677	<0.0001
molecular_function	nucleic acid binding	GO:0003676	<0.0001
molecular_function	RNA binding	GO:0003723	<0.0001
molecular_function	transcription regulator activity	GO:0030528	<0.0001
biological_process	chromatin assembly	GO:0031497	0.0001
biological_process	metabolic process	GO:0008152	0.0001
biological_process	nucleosome organization	GO:0034728	0.0001
biological_process	primary metabolic process	GO:0044238	0.0001
biological_process	regulation of biosynthetic process	GO:0009889	0.0001
biological_process	regulation of cellular biosynthetic process	GO:0031326	0.0001
biological_process	regulation of gene expression	GO:0010468	0.0001
biological_process	regulation of macromolecule biosynthetic process	GO:0010556	0.0001
biological_process	regulation of metabolic process	GO:0019222	0.0001
biological_process	RNA processing	GO:0006396	0.0001
biological_process	cellular macromolecular complex subunit organization	GO:0034621	0.0002
biological_process	DNA conformation change	GO:0071103	0.0002
biological_process	nucleosome assembly	GO:0006334	0.0002
biological_process	transcription	GO:0006350	0.0002
biological_process	regulation of cellular metabolic process	GO:0031323	0.0004
cellular_component	nucleosome	GO:0000786	0.0004
cellular_component	ribosome	GO:0005840	0.0004
biological_process	regulation of primary metabolic process	GO:0080090	0.0005
cellular_component	nuclear lumen	GO:0031981	0.0005
molecular_function	structural constituent of ribosome	GO:0003735	0.0005
biological_process	chromatin organization	GO:0006325	0.0007
biological_process	DNA packaging	GO:0006323	0.0007
biological_process	protein-DNA complex assembly	GO:0065004	0.0007
biological_process	regulation of transcription	GO:0045449	0.0008
biological_process	chromatin assembly or disassembly	GO:0006333	0.0014
biological_process	chromosome organization	GO:0051276	0.0015
biological_process	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	GO:0019219	0.0015
cellular_component	nucleoplasm	GO:0005654	0.0017
cellular_component	protein-DNA complex	GO:0032993	0.0018
biological_process	cellular macromolecular complex assembly	GO:0034622	0.0021

biological_process	regulation of nitrogen compound metabolic process	GO:0051171	0.0021
cellular_component	chromatin	GO:0000785	0.0023
molecular_function	cytokine activity	GO:0005125	0.0024
biological_process	RNA biosynthetic process	GO:0032774	0.0029
biological_process	transcription, DNA-dependent	GO:0006351	0.0032
molecular_function	cytokine receptor binding	GO:0005126	0.0036
biological_process	RNA splicing	GO:0008380	0.0039
biological_process	response to stress	GO:0006950	0.0041
biological_process	regulation of RNA metabolic process	GO:0051252	0.0042
biological_process	regulation of transcription, DNA-dependent	GO:0006355	0.0048
biological_process	organelle organization	GO:0006996	0.0049
biological_process	ncRNA metabolic process	GO:0034660	0.0095

Table S8.3: GO-Categories *depleted* for incomplete lineage sorting (BH+CH states) tested using the wilcoxon rank test implemented in FUNC. Shown are all categories with a FWER < 0.01.

Taxonomy	GO-Category Name	GO-Id	FWER
biological_process	biological adhesion	GO:0022610	<0.0001
biological_process	cell adhesion	GO:0007155	<0.0001
cellular_component	integral to membrane	GO:0016021	<0.0001
cellular_component	intrinsic to membrane	GO:0031224	<0.0001
cellular_component	membrane	GO:0016020	<0.0001
cellular_component	membrane part	GO:0044425	<0.0001
cellular_component	plasma membrane	GO:0005886	<0.0001
molecular_function	calcium ion binding	GO:0005509	<0.0001
cellular_component	plasma membrane part	GO:0044459	0.0002
molecular_function	ion binding	GO:0043167	0.0015
molecular_function	metal ion binding	GO:0046872	0.0019
molecular_function	diacylglycerol binding	GO:0019992	0.0022
molecular_function	cation binding	GO:0043169	0.0028
cellular_component	extracellular matrix part	GO:0044420	0.0036
cellular_component	proteinaceous extracellular matrix	GO:0005578	0.0038
biological_process	regulation of small GTPase mediated signal transduction	GO:0051056	0.0045
biological_process	cell communication	GO:0007154	0.0054
cellular_component	extracellular matrix	GO:0031012	0.0057
molecular_function	GTPase regulator activity	GO:0030695	0.0079

Table S8.4: GO-Categories *enriched* for incomplete lineage sorting (BH+CH states) tested using the wilcoxon rank test implemented in FUNC. Shown are all categories with a FWER < 0.01.

Taxonomy	GO-Category Name	GO-Id	FWER
biological_process	cellular macromolecule biosynthetic process	GO:0034645	<0.0001
biological_process	cellular macromolecule metabolic process	GO:0044260	<0.0001
biological_process	cellular metabolic process	GO:0044237	<0.0001
biological_process	gene expression	GO:0010467	<0.0001
biological_process	macromolecule biosynthetic process	GO:0009059	<0.0001
biological_process	macromolecule metabolic process	GO:0043170	<0.0001
biological_process	metabolic process	GO:0008152	<0.0001

biological_process	nitrogen compound metabolic process	GO:0006807	<0.0001
biological_process	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	GO:0006139	<0.0001
biological_process	primary metabolic process	GO:0044238	<0.0001
biological_process	regulation of biosynthetic process	GO:0009889	<0.0001
biological_process	regulation of cellular biosynthetic process	GO:0031326	<0.0001
biological_process	regulation of gene expression	GO:0010468	<0.0001
biological_process	regulation of macromolecule biosynthetic process	GO:0010556	<0.0001
biological_process	regulation of macromolecule metabolic process	GO:0060255	<0.0001
biological_process	regulation of metabolic process	GO:0019222	<0.0001
biological_process	regulation of nitrogen compound metabolic process	GO:0051171	<0.0001
biological_process	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	GO:0019219	<0.0001
biological_process	regulation of primary metabolic process	GO:0080090	<0.0001
biological_process	regulation of RNA metabolic process	GO:0051252	<0.0001
biological_process	regulation of transcription, DNA-dependent	GO:0006355	<0.0001
biological_process	regulation of transcription	GO:0045449	<0.0001
biological_process	RNA biosynthetic process	GO:0032774	<0.0001
biological_process	RNA metabolic process	GO:0016070	<0.0001
biological_process	transcription, DNA-dependent	GO:0006351	<0.0001
biological_process	transcription	GO:0006350	<0.0001
cellular_component	intracellular	GO:0005622	<0.0001
cellular_component	intracellular membrane-bounded organelle	GO:0043231	<0.0001
cellular_component	intracellular organelle	GO:0043229	<0.0001
cellular_component	intracellular organelle part	GO:0044446	<0.0001
cellular_component	intracellular part	GO:0044424	<0.0001
cellular_component	membrane-bounded organelle	GO:0043227	<0.0001
cellular_component	nuclear part	GO:0044428	<0.0001
cellular_component	nucleus	GO:0005634	<0.0001
cellular_component	organelle	GO:0043226	<0.0001
cellular_component	organelle part	GO:0044422	<0.0001
molecular_function	DNA binding	GO:0003677	<0.0001
molecular_function	nucleic acid binding	GO:0003676	<0.0001
molecular_function	transcription regulator activity	GO:0030528	0.0002
biological_process	regulation of cellular metabolic process	GO:0031323	0.0003
cellular_component	nuclear lumen	GO:0031981	0.0032
biological_process	cellular biosynthetic process	GO:0044249	0.0034
biological_process	organelle organization	GO:0006996	0.0062
biological_process	biosynthetic process	GO:0009058	0.0067
cellular_component	nucleoplasm	GO:0005654	0.0079
molecular_function	transcription activator activity	GO:0016563	0.0081
cellular_component	intracellular organelle lumen	GO:0070013	0.0091
cellular_component	organelle lumen	GO:0043233	0.0091

Table S8.5: GO-Categories *depleted* for incomplete lineage sorting (BH+CH states) genes with length >= 20 kilobases tested using the Wilcoxon rank test implemented in FUNC. Shown are all categories with FWER < 0.01.

Taxonomy	GO-Category Name	GO-Id	FWER
biological_process	biological adhesion	GO:0022610	<0.0001
biological_process	cell adhesion	GO:0007155	<0.0001
biological_process	multicellular organismal process	GO:0032501	<0.0001
cellular_component	extracellular region	GO:0005576	<0.0001
cellular_component	intrinsic to membrane	GO:0031224	<0.0001
cellular_component	membrane	GO:0016020	<0.0001
cellular_component	membrane part	GO:0044425	<0.0001
cellular_component	plasma membrane	GO:0005886	<0.0001
molecular_function	transmembrane receptor activity	GO:0004888	0.0001
cellular_component	integral to membrane	GO:0016021	0.0002
molecular_function	calcium ion binding	GO:0005509	0.0003
biological_process	system process	GO:0003008	0.0045
molecular_function	extracellular matrix structural constituent	GO:0005201	0.0066

Table S8.6: GO-Categories *enriched* for incomplete lineage sorting (BH+CH states) genes with length >= 20 kilobases tested using the wilcoxon rank test implemented in FUNC. Shown are all categories with FWER < 0.01.

Supplementary Information 9

Genome-wide Estimates of Nucleotide Diversity in Ulindi

Ines Hellmann[1,*] and Kay Prüfer[2]

1. Max F. Perutz Laboratories, University Vienna, Vienna, Austria
2. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

* To whom correspondence should be addressed (ines.hellmann@univie.ac.at)

Estimating diversity

Although our coverage is high, we can neither neglect sequencing errors nor the possibility that at some positions only one chromosome was sampled. In order to get unbiased estimates of nucleotide diversity we re-aligned all Ulindi reads to the human genome hg18 using bwa [46]. From these alignments we determine the true coverage using only 'reliable' base calls (see SI 2a). We further exclude clusters of clonal reads: If multiple reads have the same start position, we keep only the longest. Finally, we restrict our analysis to sites that have coverage of at least four and at most 40. Next, we use the program mlRho [81] (version 1.5) to obtain estimates of nucleotide diversity (π_{mlRho}) and error rates within our data. mlRho also provides 95% confidence intervals for diversity estimates. mlRho is designed to estimate nucleotide diversity from next generation sequencing data of one diploid individual. It concomitantly estimates the sequencing error rate and diversity also accounting for missing data.

This allows us to monitor diversity of Ulindi's two chromosomes with good confidence, however, the variance in coalescent times of two chromosomes is enormous and leaves little power to detect selection. Therefore, we also use a total of 10.4 gigabases of Illumina sequence data from 3 other bonobos (SI 5). In order for mlRho to work properly >4x coverage per diploid individual is required and therefore cannot be used for the low coverage data. Thus diversity for low coverage data that will be used for the selection screen is calculated differently from Ulindi's diversity, which we will primarily use for the X- to autosome comparison. To avoid ambiguity from sample size we estimate diversity from the low coverage bonobo data from a single allele per individual by choosing one random base from all reads at each position. We then calculate nucleotide diversity as π_{LC} by tabulating the number of pair-wise differences over the number of pair-wise comparisons per basepair.

Quality check

As for the SNP calling we use the 22kb from the Fischer dataset [16] to evaluate the precision of our estimates. We know that Ulindi has 35 heterozygote sites in these regions and the SNP calling

identifies 30, which – if taken at face value – would bias nucleotide diversity downwards. With mlRho we estimate π_{mlRh} = 1.6 x 10^{-3} which corresponds roughly to 35 segregating sites within 22kb.

Annotation based Filtering

Since we use the human genome as reference, we filter bonobo-duplicated regions out which could be over-collapsed (SI 4). For annotation-based filtering we use annotations from the human genome hg18 as downloaded from the UCSC Genome Browser. We filter against all repetitive regions to avoid mapping issues. Furthermore, we only want to analyze putatively neutral sites, so we only use intronic and intergenic sequences for our diversity estimates.

Selection screen

One hallmark of a selective sweep is that it locally reduces nucleotide diversity; that is nucleotide diversity increases with the distance from the selected site. Here we try to use these two features to enrich for regions in the genome that have recently experienced a selective sweep.

First, we estimate diversity for consecutive windows of 20kb using one Ulindi allele and the low coverage data from 4 other bonobos. In order to correct for mutation rate variation, we use the 5-way alignments described in SI 3 to estimate divergence to the orangutan. We standardize the orang divergence so that the mean corresponds to the observed bonobo diversity and treat this standardized orang divergence as expected bonobo diversity per chromosome. Next, we sort the windows into three classes: low, normal and high diversity, whereas we define low diversity as regions for which the probability of observing S_n segregating sites within a sample of n chromosomes, given θ $Pr(S_n = s| \theta)$ with [82].

(1) $$Pr(S_n = s|\theta) = \frac{n-1}{\theta} \sum_{j=1}^{n-1} (-1)^{(j-1)} \binom{n-2}{j-1} \left(\frac{\theta}{j+\theta}\right)^{(s+1)}$$

Whereas we approximate θ with the expected bonobo diversity, i.e. the scaled substitution rates, and get S_n from π_{LC} assuming the standard neutral model and a sample size of four. From (1) we can obtain a probability density function and we define windows as having low and high diversity, if S falls within the respective 20% tails of the distribution. Normal diversity is defined as anything that falls inbetween. Next, we use a HMM with three states: sweep, neutral and balancing with initial probabilities 0.05, 0.9 and 0.05. The emission and transition probabilities are then optimized using the Baum-Welch algorithm as available in the HMM-package of R, and the posterior probabilities (p) of each state for each window is calculated. We keep windows with p>=0.99 as candidates for further analysis.

With only few chromosomes diversity is expected to be low in some windows just by chance.

However, we would not expect diversity to increase with distance from these windows. Therefore, we calculate the Pearson's correlation coefficient (r) of log(π) ~ distance in bp for up to 20 x 20kb to the left and the right of the candidate window. For windows with r_{left}< -0.4 and r_{right}>0.4 we conducted neutral coalescent simulations with $\theta = \theta_{exp}$, Ne=10,000 and the recombination estimates from the human genetic map (Kong et al. 2001) and calculated how often the observed r_{left} is smaller and r_{right} is higher than the simulated data. We picked these cut-offs to give reasonable false / true positive ratios for a wide parameter range (Figure S9.1). We note that this test has no power for regions with recombination rates smaller than 0.5 cM/Mb. For each window that fulfilled the above candidate-criteria we conducted neutral simulations to obtain more precise false positive rates and restrict the candidates to <=1%. Like this we end up with a list of 13 candidate regions 9 of which do not overlap with genes (Table S9.1).

From this list the ephrin receptor A5 appears to be most interesting. Another member of this protein family, EphA6, had been implicated in a similar selection screen of the human genome [83]. Many ephrin receptors are involved in synaptogenesis [84]. EphA5 in particular has also been shown to be involved in the signaling between beta islet cells and to adjust their capacity to secrete insulin in response to glucose [85]. Another interesting candidate is the muscle gene dystrophin,

However, it might also be that we pick up regions under strong, localized background selection.

X-Autosome effective population sizes

Under random mating the effective populations size for the X-chromosome is expected to be ¾ that of autosomes. All things being equal we would therefore also expect that X-diversity is reduced by a factor of ¾. Deviations from this ratio can be due to differences in the variance of reproductive success between males and females. For example, if each female leaves one offspring, but in males one male has 5 and some have none, this would increase the X-effective population size (Nx) relative to the autosomes (Na). In other words, the female effective population size (Nf) is bigger than the male effective population size (Nm). However, the ratio of X- to autosome diversity depends also on other selectively neutral factors:

1. differences between male and female mutation rates
2. differences in male and female migration behavior
3. recent population size changes

We use the 5-species alignments (SI 3) to estimate the mutation rate on the bonobo lineage using parsimony, with orangutan and human as outgroups, so that we can use the divergence on the bonobo branch since the bonobo-human split to correct for differences between male and female mutation rates (1st factor). Thus we implicitly assume that the average male mutation bias since the bonobo-human split is a good approximation to the male mutation bias in bonobos over the last

~200 ky, which must not necessarily be true [86]

Concerning the bonobo demography: population genetic studies using nuclear re-sequencing data failed to find evidence for deviations from the standard neutral model [44]; This implies that Bonobos have an approximately constant size and have no or only very little population structure, at least within the time-span relevant to this analysis. Therefore, we expect neither population size changes (2nd factor) nor sex biased migration to affect X/Autosome diversity (3rd factor.), so that the only neutral interpretation would be differences the male and female effective population size.

To estimate Nx/Na, we use diversity estimates for non-overlapping 200kb windows with similar filtering criteria as in the selection screen, but a less exclusive annotation filtering: Now, we only filter against duplications and simple repeats. Furthermore, we assume that the species split of bonobos and humans happened 4.5 Mya, a generation time of 20 years and an ancestral population size (N_{anc}) of 45,000 (SI 8), whereas the X-chromosome has ¾ of this. If we assume a constant population size, no recombination within a locus and free recombination between loci, we can calculate the probability of observing S segregating sites given Ne and the mutation rate analytically [87].

This said, our maximum likelihood estimate for Ne is 12,000 with Nx/Na = 0.85 (95% C.I. 0.79–0.93; Figure 9.3), which corresponds to an Nf/Nm of 2.04. In bonobos the female effective population size is roughly double that of males and this difference is significant. It is also robust to model assumptions such as the ancestral Nx/Na (tested values: 0.63, 0.78, constant) as well as the bonobo-human split time and ancestral population size (tested values N_{anc}=55,000 & t = 4.1 Mya). The mean of our estimates also did not change, when applying more stringent filtering criteria excluding exons and interspersed repeats.

For comparison, we did the same analysis for one high coverage African Yoruba (NA19240) and one European CEPH individual (NA12878) from the 1000 Genomes pilot project [88]. We downloaded all Illumina reads from those two individuals and aligned them to hg18. We then continued exactly as for the bonobo reads with identical parameters for read and site filtering (see SI 5), SNP calling and mlRho. The Nx/Na is 0.8 for the Yoruban and 0.68 for the CEPH individual (Table S9.2).

The Nx/Na is significantly different between the two human populations, but the Yoruban estimate is neither significantly different from 0.75 and nor is it different from the bonobo estimate. The human result is consistent with a previous study where also a reduced Nx/Na is seen in non-African compared to African populations. Keinan et al. 2010 [89], using the HapMap data, found Nx/Na to not be significantly different from 0.75 in Africans and a reduced Nx/Na in out of African populations. They went on to show that the low Nx/Na in out-of-Africa populations cannot be explained by demographic scenarios with equal sex-ratios, but they could fit a model with strong

male migration from African to non-African populations after the out-of-Africa bottleneck. Our simulations under the Keinan-model yielded the same Nx/Na ratio as estimated from the CEPH individual (Nx/Na obs.= Nx/Na sim. =0.68). However, these simulations also yielded Nx/Na = 0.75 for the African individual, while the observed ratio is slightly, but not significantly, elevated (0.8 C.I. 0.74-0.87).

Hammer et al. 2010 [90] suggested an alternative interpretation for the low Nx/Na: They observed a strong positive correlation between the average genetic distance to a gene and diversity on the X chromosome. Furthermore the average genetic distance to a gene is shorter on the X as compared to autosomes (Figure 9.4), therefore selection will have a bigger impact on linked neutral sites and the low Nx/Na ratio cannot be solely interpreted as a sign for sex biased evolution.

In an attempt to distinguish selection from demography, we binned the data according to the background-selection factor (*B*) as estimated from human recombination and gene annotation [74]. Our most neutral bin, in which diversity is reduced by at most 0.9x, also yields most windows. Notably, Nx/Na estimates from this neutral bin are above 0.75 for all three individuals, but the genome-wide average falls below 0.75. This suggests that X-diversity is more reduced by selection than autosomal diversity (Figure S9.4 A. Furthermore, this difference between neutral (B>=0.9) and selected (B<0.9) bins appears to be strongest in Europeans and weakest for bonobos. Possible explanations for this pattern are:

1. The predicted effect of background selection is most accurate for Europeans.
2. There was more recent selection in Europeans subsequent to the bottleneck in which the efficacy of selection was reduced [91].
3. X/A diversity in Europeans is reduced due to demographic events (Gottipati et al., 2011).

In an attempt to distinguish these two possibilities, we investigate the relationship between expected amount of background selection and diversity, reasoning that the correlation should linearly increase with *B*. We find the correlation for the X-chromosome is always stronger than for the autosomes, suggesting that the X is likely to be more impacted by selection. The correlation for the bonobo is always much weaker, suggesting that assuming human recombination rates is inappropriate (Figure S9.4 B). Furthermore, the correlation in the CEPH seems weaker compared to the Yoruban. This may be due to the fact that drift had a bigger impact on the CEPH and hence diversity is expected to be more stochastic. Therefore explanation (1) appears unlikely. Furthermore, X/A diversity in the CEPH is almost always lower than diversity in the Yoruba, which is consistent with the somewhat more parsimonious explanation (3).

Putting the focus back on what we can say about bonobos: We clearly see an elevated Nx/Na (0.85). There is no evidence supporting that this increase is due to demographic effects; bonobos appear to be a constant-size, unstructured population. We also tried to correct for more diversity reducing selection on the X-chromosome and only saw a slight difference if only between putatively neutral regions are considered (B>=0.9; Nx/Na=0.87). This leads us to suggest that the female effective population size in bonobos is double that of males. This can be explained if every generation double as many females as males breed or depicting a less extreme scenario, if the males have a higher reproductive variance than females. The comparison of these numbers to human data is hampered by the more complex human demography. But looking at the Yoruba data, which is probably least likely to be biased by demographic effects, we also estimate an Nx/Na ratio of 0.85, suggesting that also in humans the female effective population size is roughly double that of males.

It will be interesting to be able to compare these results to gorillas and chimpanzees, both of which have more sexual dimorphism and also less egalitarian mating systems than humans and bonobos.

Table S9.1: Regions that are likely to have experienced a recent selective sweep in Bonobos. Genome coordinates and annotations are from hg18.

bed	Pr of rejecting neutrality	rr (human cM/Mb)	refGene (human)
chr4:65960000-66060000	0	0.61	EPHA5
chr4:8220000-8280000	0	1.91	SH3TC1
chr7:101120000-101180000	0	0.85	
chr8:4940000-4980000	0.001	3.17	
chr9:135240000-135260000	0.001	2.77	C9orf96
chrX:32620000-32640000	0.001	2.10	DMD
chr11:1940000-1980000	0.002	1.52	LOC100133545, H19, MIR675
chr4:60480000-60500000	0.002	0.52	
chr7:2200000-2220000	0.002	1.91	MAD1L1
chr11:63580000-63620000	0.003	0.92	MACROD1
chr1:105380000-105400000	0.004	1.04	MIR548H3
chr19:2060000-2080000	0.005	4.79	AP3D1
chr17:17180000-17200000	0.01	0.85	NT5M

Figure S9.1: Power analysis for selective sweeps in a population with Ne=10,000. The advantageous allele that was just fixed occurred in the middle of the examined region. The sample size was 4. The curve represents different symmetric cut-offs for Pearsons correlation coefficients for distance and diversity of 20 20kb windows left and right of the selected site, ranging from b= -0.3-0.9. The different colors indicate varying recombination rates.

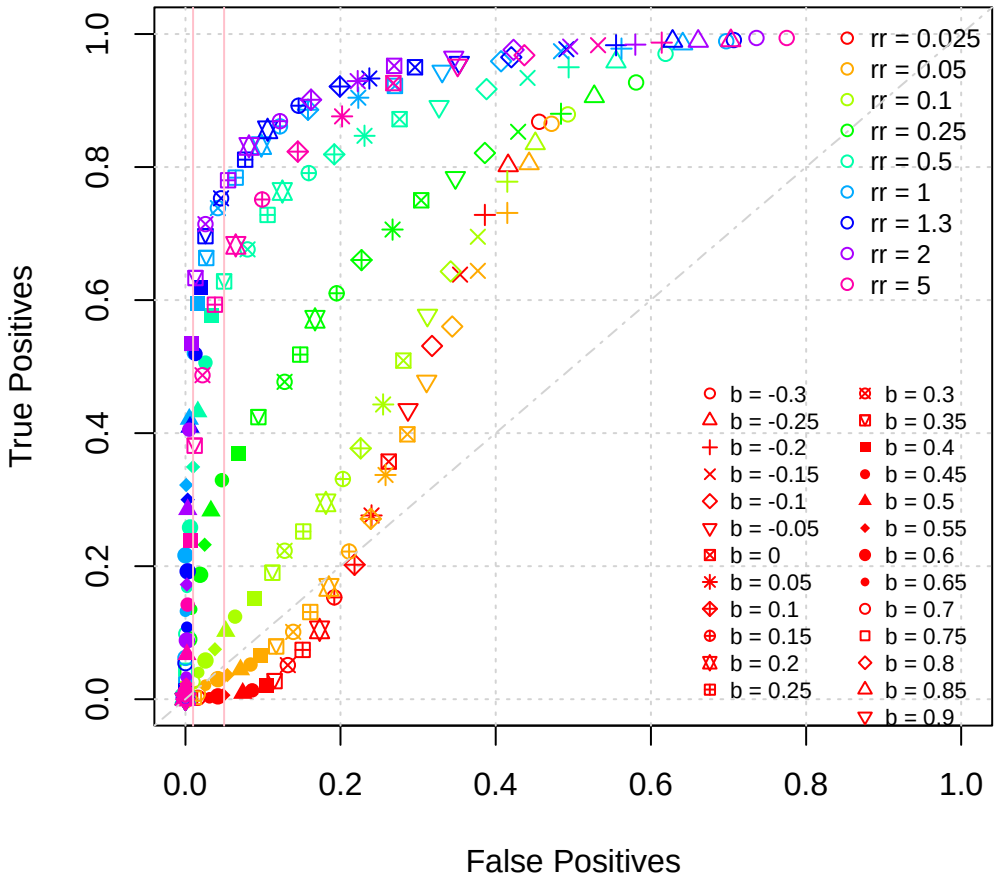


Figure S9.2: Observed diversity (red) and expected diversity (blue) in one of the candidate regions. The bottom plot gives the posterior probabilities from the HMM diversity low (green), high (red) and intermediate diversity (grey).

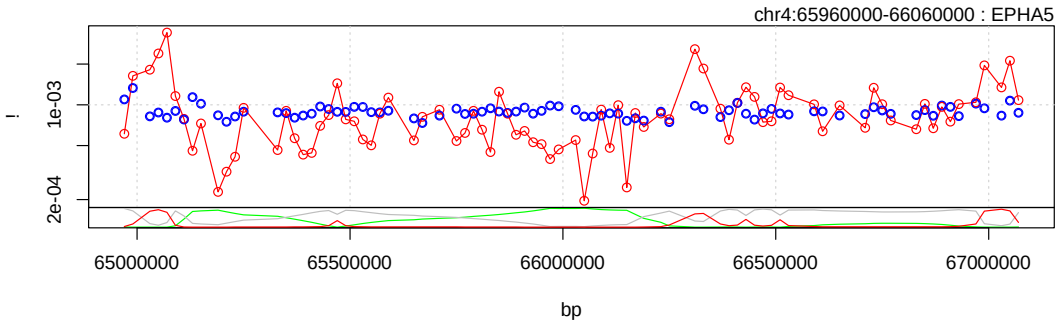


Figure S9.3: Joint Log-Likelihoods for the effective population size of extant bonobos (Ne) and the ratio of female over male effective population size.

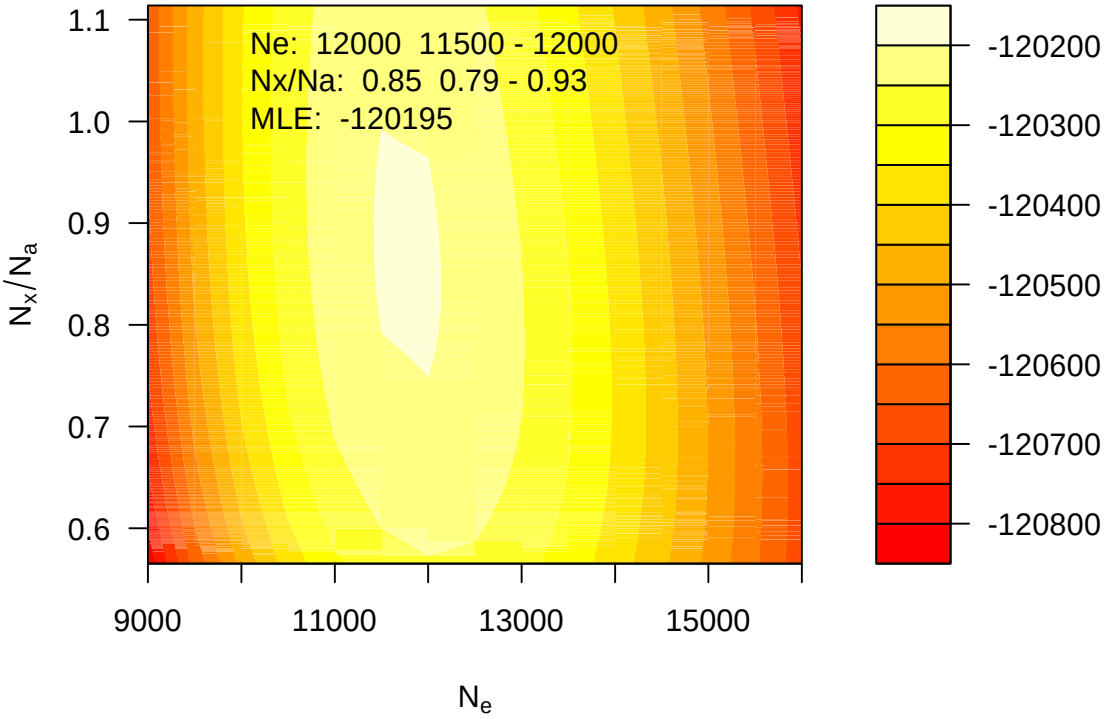


Table S9.2 Composite Maximum Likelihood estimates of Ne and Nf/Nm. The ancestral Ne of bonobos and humans was assumed to be 45,000 and the split time 4.5Mya.

		Ne	Nx/Na	Nf/Nm	Liklihood x 10^{-3}	# of A-windows	# of X-windows
bonobo	all	12000	0.85 (0.79-0.93)	2.04 (1.32-3.89)	-120	11450	636
	bs 0.9-1	12000	0.87 (0.76-1.02)	2.45 (1.05-8.32)	-20	2918	177
yoruba	all	10500	0.8 (0.74-0.87)	1.51 (0.95-2.4)	-116	11520	632
	bs 0.9-1	11000	0.85 (0.73-0.98)	2.04 (0.87-6.03)	-20	3971	179
ceph	all	8500	0.68 (0.63-0.74)	0.55 (0.28-0.91)	-112	11139	635
	bs 0.9-1	9000	0.78 (0.68-0.91)	1.26 (0.5-3.16)	-19	2812	181

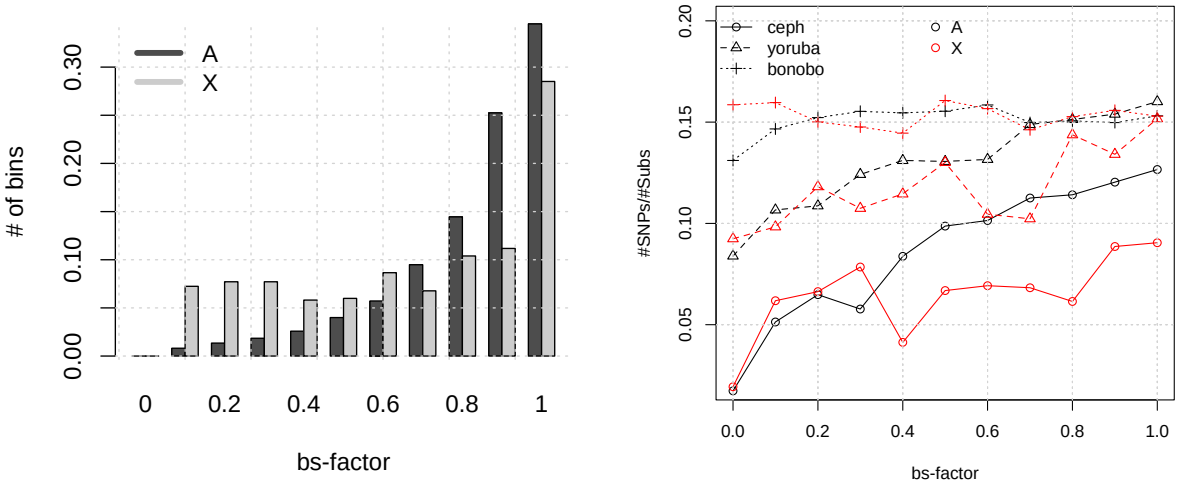


Figure S9.4 A) Relative number of windows sorted according the putative effect of background selection. Overall neutrally evolving sites are more likely to be effected by selection on linked sites on the X than on autosomes. B) The median of the #SNPs over the number of substitutions for each background-selection bin.

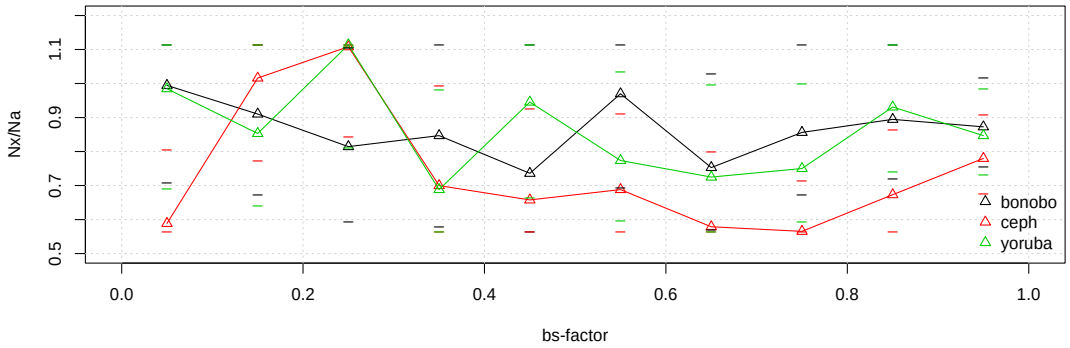


Figure S9.5. We binned the data according to the average effect of background selection, given human recombination rates (McVicker et al., 2008) and estimated the Nx/Na for each bin separately. Dashes mark the 95% confidence intervals. Most data were available for the bin where selection had the least effect on linked sites: bs-factor 0.9-1. Note that for this bin Nx/Na estimates for all three populations are similar and above 0.75.

Supplementary Information 10

Divergence, Site Pattern Analysis and Signals of Admixture

Kay Prüfer[1,*] and Ines Hellmann[2]

1. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
2. Max F. Perutz Laboratories, University Vienna, Vienna, Austria

* To whom correspondence should be addressed (pruefer@eva.mpg.de)

We use divergence and a site-pattern based statistic to gain insight into the relationship between and within bonobo and chimpanzee populations. Our results show no evidence of preferential admixture of bonobos with either eastern or central chimpanzee subpopulation. Our comparison with western chimpanzees, on the other hand, hints towards a closer relationship between western chimpanzees and bonobos than between bonobos and eastern or central chimpanzees. This difference is, however, not significant in any of the individual comparisons.

Western chimpanzee are an outgroup to eastern and central chimpanzee. Using the site-pattern based statistic and divergence estimates, we test whether western chimpanzees show signals of preferential admixture with either central or eastern chimpanzee. We find consistent signals for a closer relationship between eastern and western chimpanzees as compared to central and western chimpanzees. Interestingly, we also find a signal dividing the central individuals into two distinct groups with a different relationship to eastern chimpanzees. This result indicates subpopulation structure within central chimpanzees.

Data

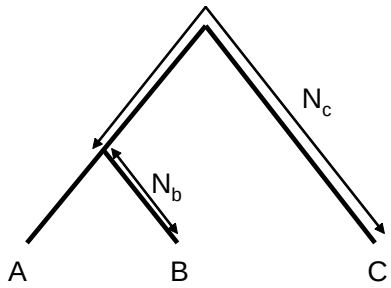
We use the filtered Illumina reads of 16 chimpanzees and 3 Bonobos, the 454 reads of Ulindi and the Sanger sequencing reads of Clint (the source of shotgun reads for the chimpanzee assembly) for divergence estimates and the site pattern analysis. Filtering of Illumina reads have been carried out as described in SI 5. Briefly, we filter reads according to mapping and base quality, and sample one read per position and individual (to exclude a bias against SNPs which would be introduced if a consensus were called from the low-coverage data). Reads from Ulindi has been filtered and positions of SNPs have been assigned to two copies of the genome sequence based on the number of reads in support (see SI 2a for details). Regions of overcollapsed duplications were excluded (see SI 4). Clint reads have been processed with the same set of parameters as 454 reads, but were not split in separate alleles at SNP positions due to the lower coverage. All reads were mapped to the human genome hg18. Using the whole genome alignment HCBOR (see SI 3), the orangutan and rhesus macaque genome sequence was added based on their alignment with the human genome. CpG sites were flagged when they were present in either human, orangutan or rhesus macaque. A full list of all included individuals is shown in Table S10.1.

Individual	Species	Alignment
CC1	Central Chimpanzee	BWA
CC2	Central Chimpanzee	BWA
CC3	Central Chimpanzee	BWA
CC4	Central Chimpanzee	BWA
CC5	Central Chimpanzee	BWA
CC6	Central Chimpanzee	BWA
CC7	Central Chimpanzee	BWA
EC1	Eastern Chimpanzee	BWA
EC2	Eastern Chimpanzee	BWA
EC3	Eastern Chimpanzee	BWA
EC4	Eastern Chimpanzee	BWA
EC5	Eastern Chimpanzee	BWA
EC6	Eastern Chimpanzee	BWA
EC7	Eastern Chimpanzee	BWA
WC1	Western Chimpanzee	BWA
WC2	Western Chimpanzee	BWA
Clint	Western Chimpanzee	BWA
B1	Bonobo	BWA
B2	Bonobo	BWA
B3	Bonobo	BWA
Ulindi1	Bonobo	BWA
Ulindi2	Bonobo	BWA
hg18	Human	-
PonAbe2	Orangutan	WGA
RheMac2	Rhesus Macaque	WGA

Table S10.1: Sequences used for the site pattern test. See SI 2a, SI 3 and SI 5 for details on mapping and alignment procedures.

Calculating Divergence

We calculated divergence based on 3-way alignments between individuals. We use the previously described method of divergence triangulation [92]. When appropriate filtering procedures are used, this method can give stable divergence estimates even when a large excess of sequence error is present in one individual [93]. With three individuals A, B and C, we count sites in which A equals B, and C is different (C specific changes N_c) and sites where A equals C, and B is different (B specific changes N_b). The divergence between A and B can then be calculated as $2N_b/(N_b+N_c)$, giving the relative lineage length to the common ancestor of A and C (or B and C).



Bonobo-Chimpanzee Divergence

We calculated the divergence of all chimpanzee individuals to Ulindi1. We restrict our analysis to human non-repetitive sequence and alignments to human autosomes (see Table S10.2). Estimates are consistently around 2.2 million years for all individuals. However, Sanger sequencing data from Clint give a lower divergence estimate than all estimates from Illumina reads. This difference may be explained by the higher quality of Sanger sequencing reads as compared to Illumina sequencing reads and by the increased power to align to the human orthologous position due to the longer read length in Sanger sequencing. The view that read length may influence correct placement in the alignment is further supported by the fact that two of the three chimpanzees sequenced with 76 base pair read length are among the highest divergent chimpanzees in our test.

Individual	Divergence to bonobo relative to human divergence	lower CI	upper CI	Divergence time in million years
CC1	0.347	0.345	0.349	2.25
CC2*	0.349	0.347	0.351	2.27
CC3	0.345	0.343	0.347	2.24
CC4	0.347	0.345	0.349	2.25
CC5	0.347	0.345	0.349	2.25
CC6	0.346	0.344	0.348	2.25
CC7**	0.350	0.348	0.352	2.27
EC1*	0.350	0.348	0.352	2.27
EC2*	0.347	0.344	0.349	2.25
EC3	0.347	0.345	0.349	2.25
EC4	0.346	0.344	0.348	2.25
EC5	0.353	0.351	0.356	2.30
EC6	0.348	0.346	0.350	2.26
EC7**	0.350	0.348	0.352	2.28
WC1**	0.349	0.347	0.351	2.27
WC2**	0.350	0.347	0.352	2.27
Clint	0.333	0.331	0.335	2.16

Table S10.2: Divergence of chimpanzee individuals to Ulindi1 for non-repeatmasked bases in human. Divergence is given relative to human-chimpanzee/bonobo common ancestor. CI: 95% confidence intervals calculated from 5000 bootstrap replicas on 100kb blocks. Divergence time is assuming 6.5 million years average divergence time between human and chimpanzee/bonobo. *Individuals with 76 cycle read length. **Individuals with less than 3GB of aligned and filtered data.

We further our analysis by investigating the distribution of divergence in blocks of 100 kilo bases along the human genome sequence. As before, analysis is restricted to autosomal sequence and filters repeatmasked sequence based on human annotation. Additionally, we require a minimum of 50 informative sites for the calculation of divergence for each block.

A more recent shared ancestry with Bonobo for some individuals may be visible as either a shift in the distribution towards smaller divergence to Bonobo or an excess of blocks with small divergence to Bonobo. We observe that Illumina sequenced individuals show a generally wider distribution compared to Clint reads. The three 76 cycle sequenced individuals show the widest distribution, potentially caused by the lower amount of data available for these individuals but also consistent with an excess of mapping artifacts

affecting the distribution (see Figure S10.1). Similarly, 101 cycle individuals with less than 3 giga bases sequence coverage after filtering (see SI 5) give a wider distribution as higher coverage individuals. When excluding these individuals, we find no evidence for a recent shared ancestry with bonobo for any individual (see Figure S10.2).

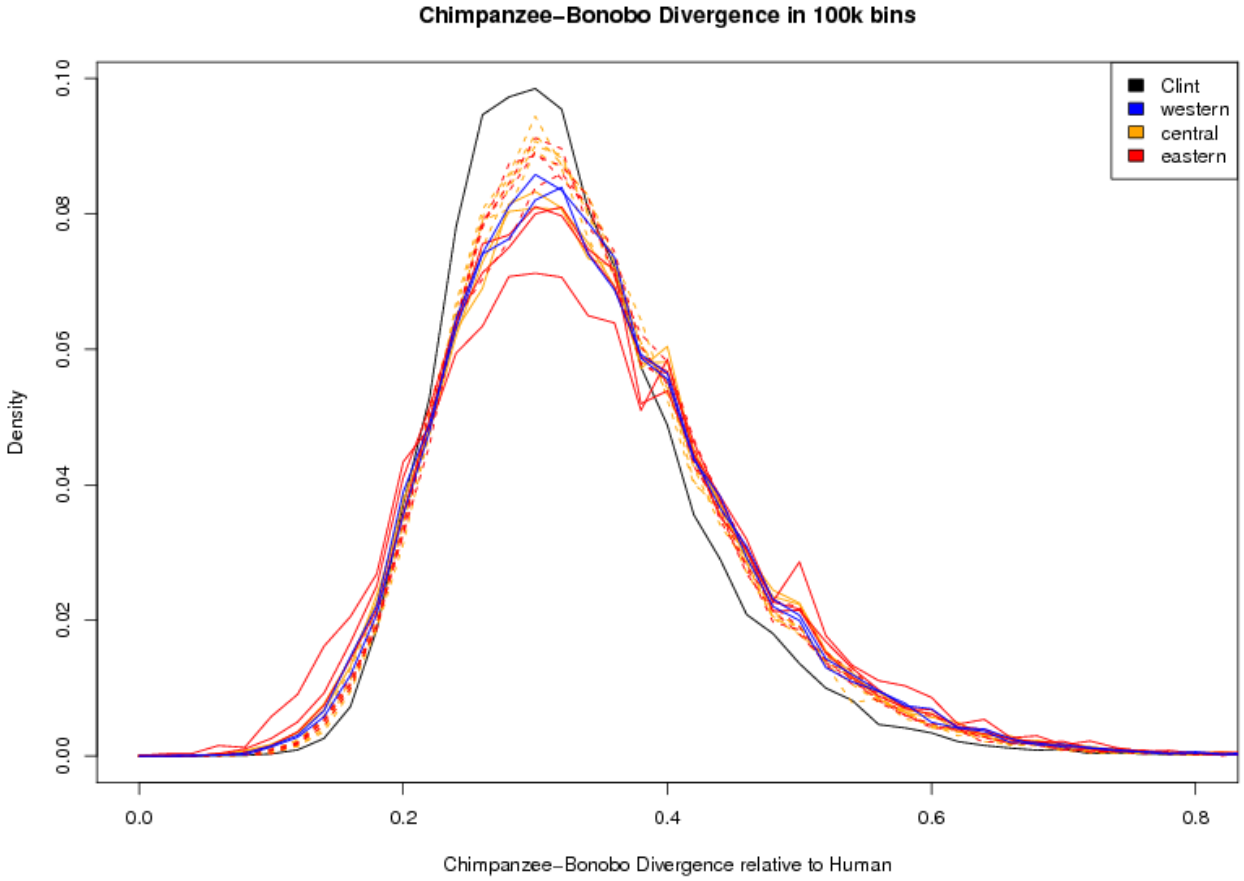


Figure S10.1: Divergence of chimpanzee individuals to Ulindi1 in 100 kilo base bins along the human autosomes. Clint divergence is shown in black, divergence for western individuals is shown in blue, divergence for central chimpanzee individuals is shown in green and divergence for eastern chimpanzee individuals is shown in red. The Illumina sequenced individuals are further divided into those with high coverage and long read length (> 3 giga bases sequence data and 101 cycle read-length) shown as dashed lines and the remaining sequences with lower coverage or smaller read length (solid lines).

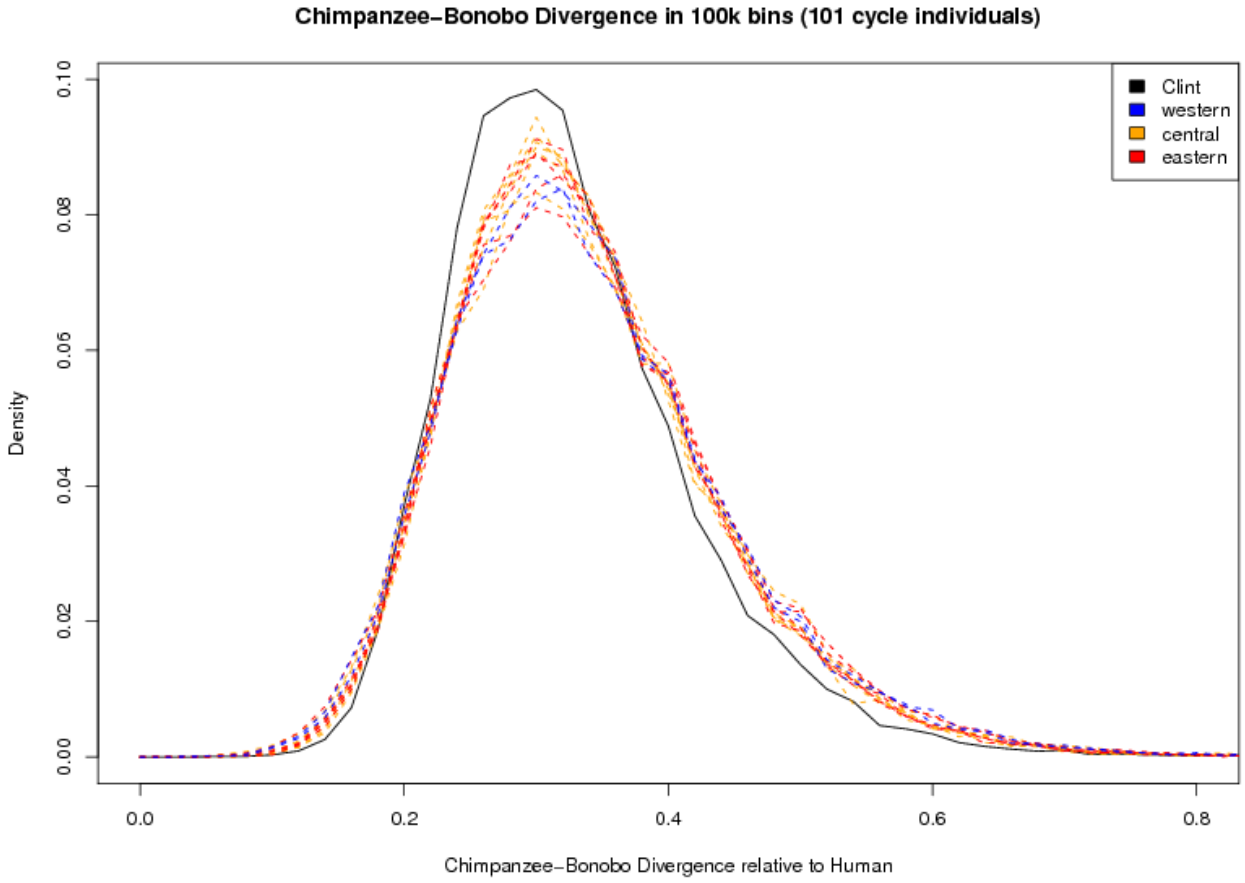


Figure S10.2: Divergence of chimpanzee individuals to Ulindi1 in 100 kilo base bins along the human autosomes. Clint divergence is shown in black, western chimpanzee divergence is shown in blue, central chimpanzee divergence is shown in green and eastern chimpanzee divergence is shown in red. Only 101 cycle Illumina individuals are included.

In a next step, we compared the divergence of Bonobo individuals to Clint. Two Bonobo individuals were sequenced on a 76 cycle Illumina run (B1 and B2) and one individual was sequenced on a 101 cycle Illumina run (B3). We also include data from mapped Ulindi reads (Ulindi1 and Ulindi2, separating SNP positions in two states). Table S10.2 gives the divergence times and Figure S10.3 shows the distribution of divergence in 100 kilo base blocks. Similar to the results for the chimpanzee divergence, the divergence of the Illumina sequence is wider than the distribution for the longer reads from 454 sequences. The distribution for B1 also appears wider as the distribution of B2. B1 has been sequenced to ca. 1x coverage, while B2 reaches ca. 2x coverage. This difference may explain the larger variation for B1. On the other hand, individual B3 was sequenced to lower coverage compared to individual B2, but at a read length of 101 cycles. Despite the lower coverage, the distribution of divergence is smaller for B3. This suggests that a read-length dependent mapping bias may contribute to the variation of divergence. The divergence estimates for B1, B2 and B3 are not significantly different according to the bootstrap confidence interval. We thus see no evidence for chimpanzee introgression in any of our sequenced bonobo individuals.

Individual	Divergence to Clint relative to human divergence	lower CI	upper CI	Divergence time in million years
B1	0.345	0.342	0.347	2.24
B2	0.346	0.343	0.348	2.25
B3	0.343	0.341	0.345	2.23
Ulindi1	0.324	0.322	0.325	2.10
Ulindi2	0.323	0.321	0.325	2.10

Table S10.3: Divergence of bonobo individuals to Clint for non-repeatmasked bases in human. Divergence is given relative to human-chimpanzee/bonobo common ancestor. CI: 95% confidence intervals calculated from 5000 bootstrap replicas on 100kb blocks. Divergence time is assuming 6.5 million years average divergence time between human and chimpanzee/bonobo.

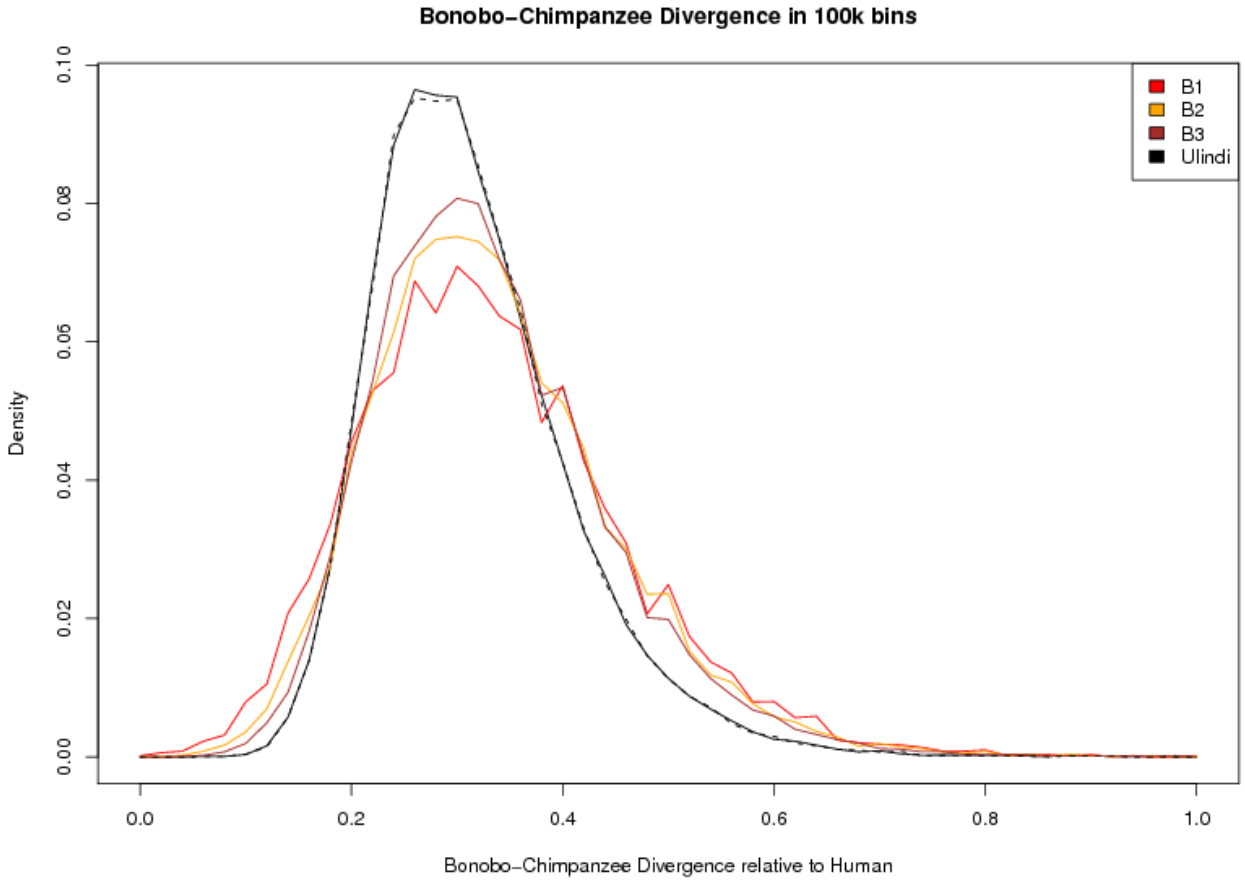


Figure S10.3: Divergence of bonobo individuals to Clint in 100 kilo base bins along the human autosomes. Ulindi divergence is shown separately for Ulindi1 (solid black) and Ulindi2 (dashed black).

Divergence of Chimpanzee Subpopulations

We also calculated divergence between eastern, central and western individuals in relation to the western chimpanzee individual Clint. In order to avoid potential biases caused by admixture of bonobo and eastern and/or central chimpanzees, we used human as an outgroup to calculate divergence. Table S10.4 shows the average divergence relative to Clint-Human divergence and Figure S10.4 shows the distribution of divergence over 100 kilo base blocks. The divergence between Illumina-sequenced western individuals and Clint is, as expected, significantly smaller than the divergence between eastern and central chimpanzees and Clint. Divergence of eastern to western chimpanzees is on average lower compared to central-western

chimpanzee divergence. This tendency is also visible as a shift towards lower divergence in the distribution of divergence.

Individual	Divergence to Clint relative to human divergence	lower CI	upper CI	Divergence time in million years
CC1	0.201	0.200	0.203	1.31
CC2	0.201	0.199	0.202	1.30
CC3	0.196	0.195	0.198	1.28
CC4	0.202	0.201	0.204	1.32
CC5	0.201	0.200	0.203	1.31
CC6	0.199	0.197	0.200	1.29
CC7	0.198	0.197	0.200	1.29
EC1	0.196	0.194	0.198	1.27
EC2	0.195	0.194	0.197	1.27
EC3	0.194	0.193	0.196	1.26
EC4	0.194	0.192	0.195	1.26
EC5	0.197	0.196	0.199	1.28
EC6	0.196	0.194	0.197	1.27
EC7	0.196	0.195	0.198	1.28
WC1	0.079	0.078	0.080	0.52
WC2	0.079	0.078	0.080	0.51

Table S10.4: Divergence of chimpanzee individuals to Clint for non-repeatmasked bases in human. Divergence is given relative to human-chimpanzee/bonobo common ancestor. CI: 95% confidence intervals calculated from 5000 bootstrap replicas on 100kb blocks. Divergence time is assuming 6.5 million years average divergence time between human and chimpanzee/bonobo.

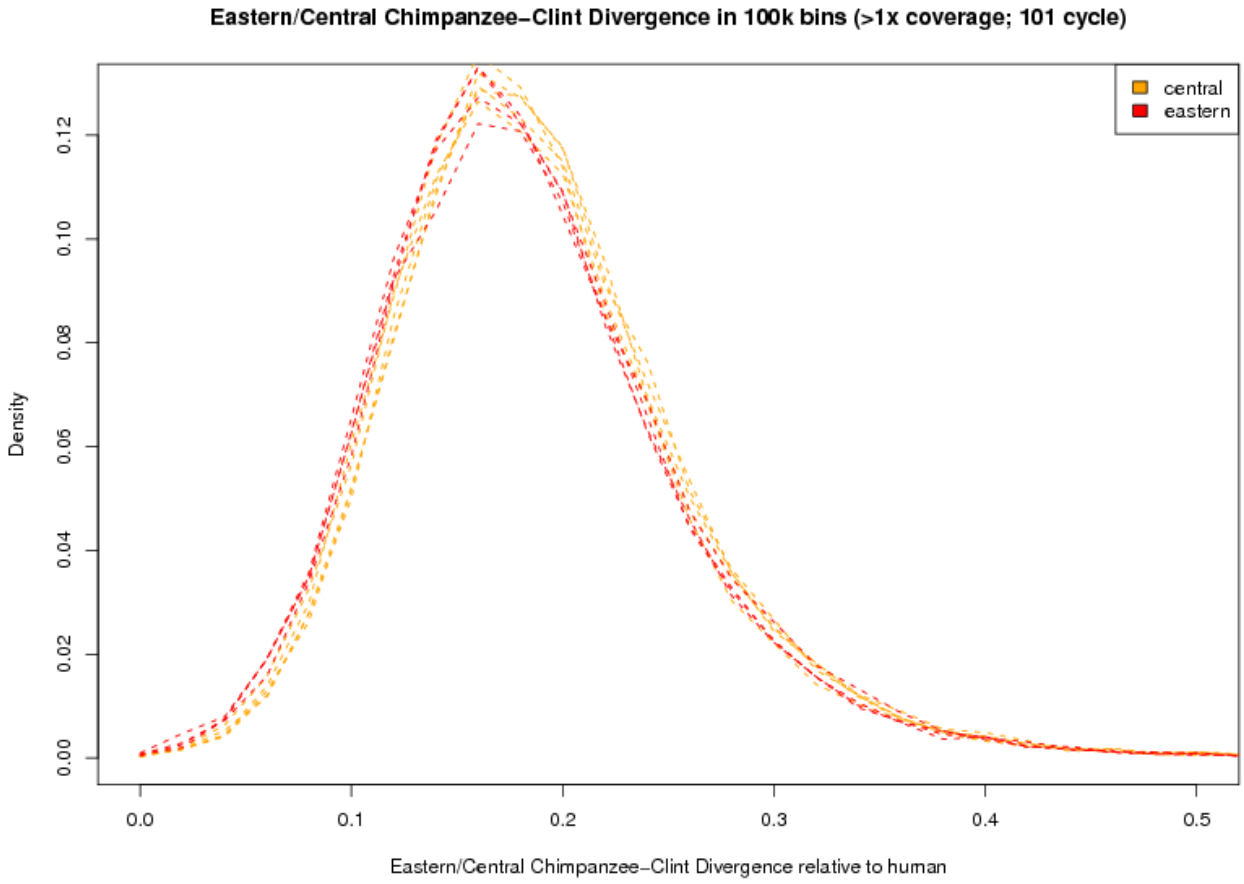


Figure S10.4: Divergence of eastern and central chimpanzee individuals to Clint in 100 kilo base blocks along the human autosomes. Central chimpanzee divergence is shown in orange and eastern chimpanzee divergence is shown in red. Only 101 cycle Illumina individuals with more than 1x coverage mapped sequence data are included.

Divergence of Bonobo Individuals

Next, we calculate divergence for the Illumina sequenced bonobo individuals to Ulindi1. We used again the human genome sequence as an outgroup. Divergence of the two individuals shows a significantly closer relationship of individuals B1 to Ulindi than individuals B2 and B3 (see Table S10.5). Similarly, the divergence distribution shows an excess of closely related segments between B1 and Ulindi (see Figure S10.5).

We test whether the closer relationship of B1 is an artifact of sequence quality by using solely sequence data of B1 and B2 that have been sequenced on the same lane. Due to the small amount of data and the generally closer relationship among bonobo individuals, the divergence within 100 kilo base blocks vary too much to give a reliable distribution. We therefore calculated divergence in blocks of 500 kilo bases and chosen larger bins for plotting (see Figure S10.6). Both divergence estimates and distribution plot support the closer relationship of individual B1 to Ulindi compared to B2-Ulindi divergence.

Individual	Divergence to Ulindi relative to human divergence	lower CI	upper CI	Divergence time in million years
B1 (separate lane)	0.079	0.078	0.080	0.51
B2 (separate lanes)	0.086	0.085	0.086	0.56
B3	0.082	0.082	0.083	0.54
B1 (mixed sequencing)	0.081	0.078	0.084	0.53
B2 (mixed sequencing)	0.086	0.084	0.089	0.56

Table S10.5: Divergence of Illumina-sequenced bonobos to Ulindi for non-repeatmasked bases in human. Divergence is given relative to human-chimpanzee/bonobo common ancestor. Divergence time is assuming 6.5 million years average divergence time between human and chimpanzee/bonobo. Sequencing data was partitioned in two sets for B1 and B2: separate lane for data acquired on different lanes and mixed sequencing for data acquired on a single lane with indexed reads from B1 and B2. The 95% confidence intervals were calculated from 5000 bootstrap replicas on 100kb blocks (separate lane) or 500kb blocks (mixed sequencing).

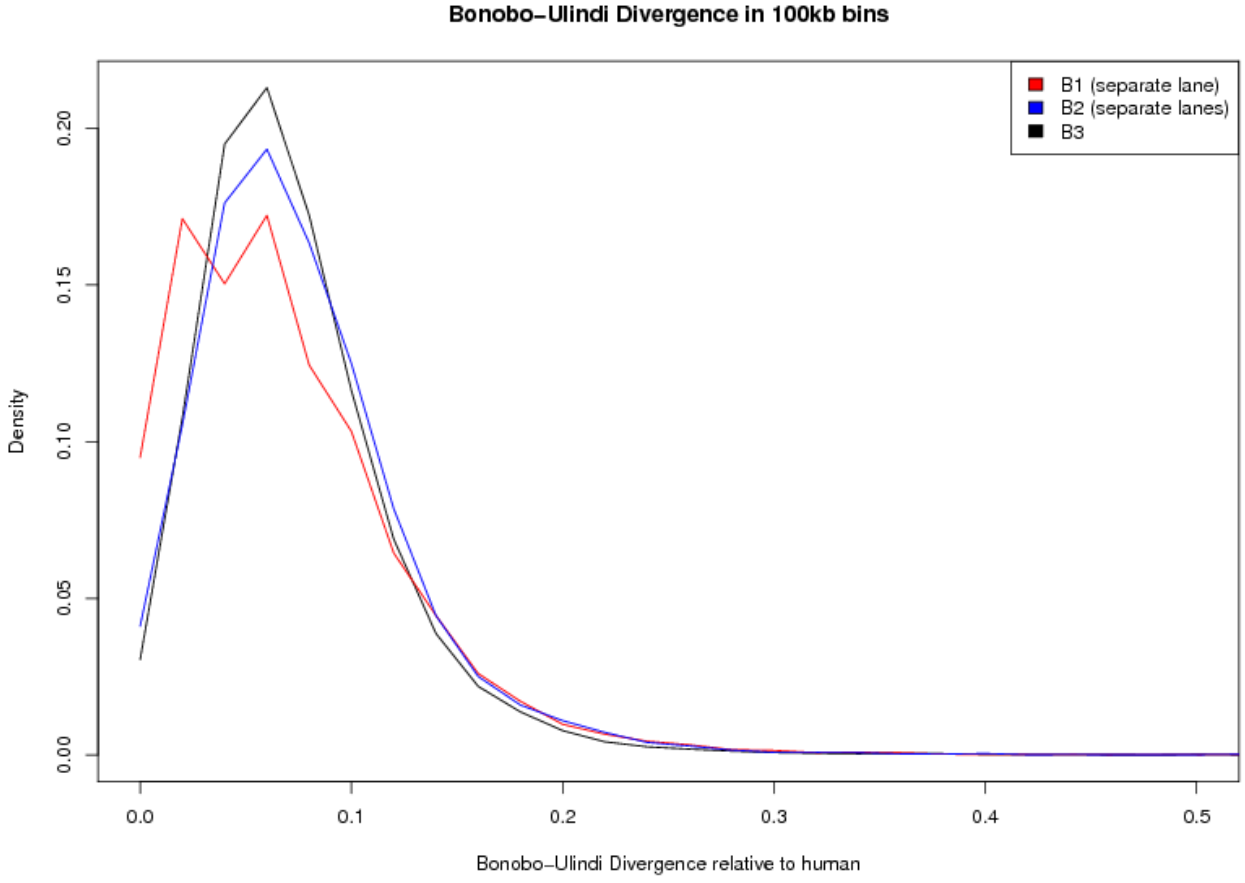


Figure S10.5: Divergence of B1, B2 and B3 to Ulindi in 100 kilo base blocks along the human autosomes. Divergence of B1 is shown in red, divergence of B2 in blue and divergence of B3 in black.

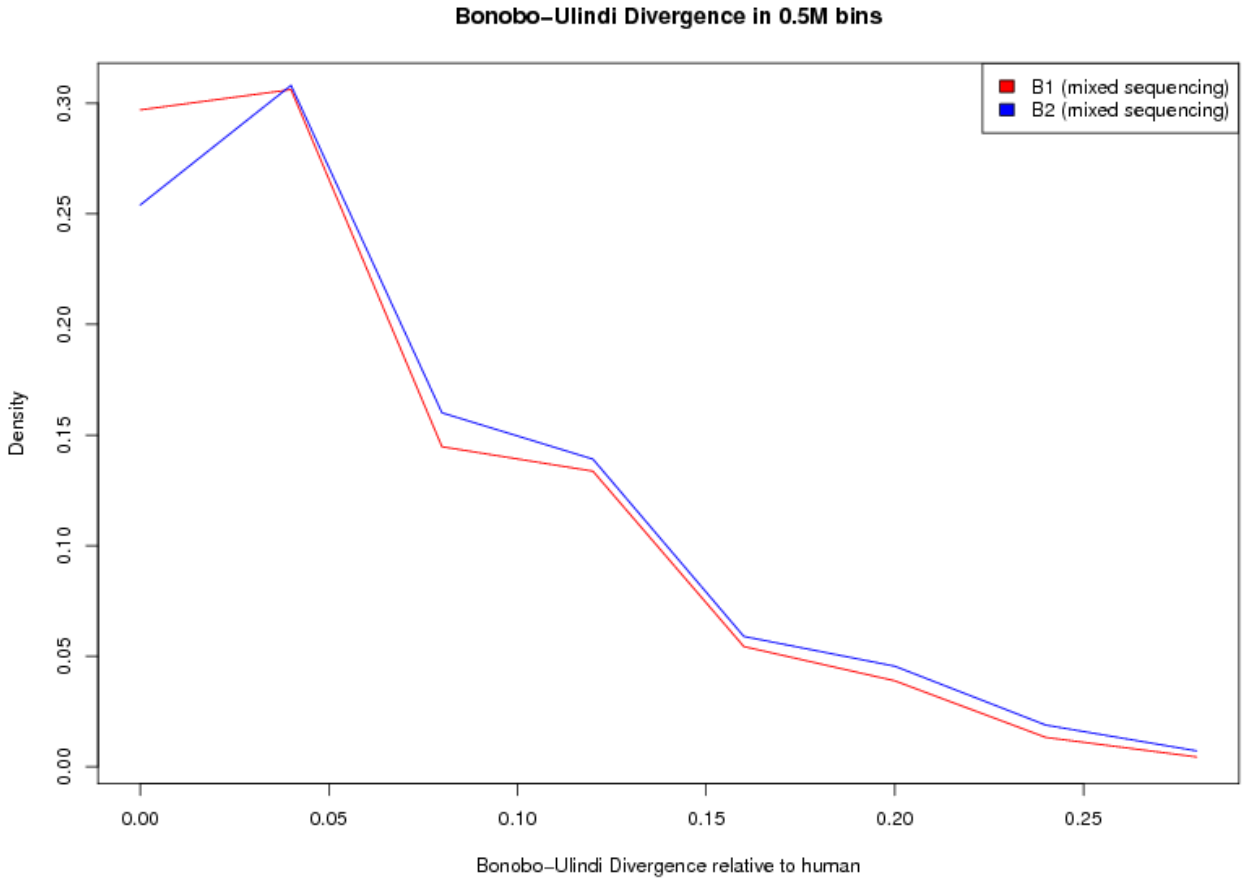
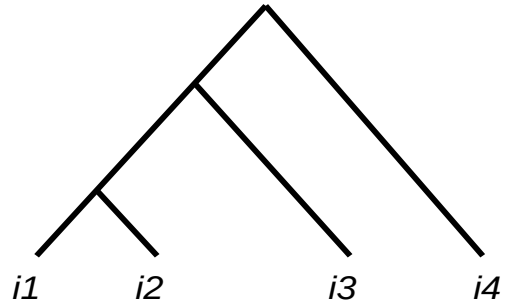


Figure S10.6: Divergence of B1 and B2 to Ulindi in 500 kilo base blocks along the human autosomes. Data for B1 and B2 was generated on one Illumina lane with indexed sequences. Divergence of B1 is shown in red, divergence of B2 in blue.

Site-Pattern Test Statistics



In order to test for admixture or ancient population structure, we followed the approach from Green et al., 2010 [94] (see also SI 15 in this reference). In this approach, the site patterns "ABBA" and "BABA" are counted in an alignment of four individuals, in the following denoted *i1, i2, i3, i4*. Individual *i4* is always used as an outgroup to assign the ancestral state and the test is based on the assumption that *i4* is not more closely related to any of the other three individuals. With this, the ABBA counts correspond to the number of shared derived sites between *i2* and *i3*, while BABA corresponds to the number of shared derived alleles between *i1* and *i3*. Based on these two counts we calculate D=(ABBA-BABA)/(ABBA+BABA), a value between -1 and 1 expressing the excess of shared sites between *i3* and either *i1* (negative values) or *i2* (positive values).

In order to determine the statistical significance of a D-value, we used a weighted block jackknife procedure to calculate the standard error [95]. For this, we count ABBA and BABA sites in blocks of 5 mega bases along the reference genome sequence. The weight for each block is the sum of informative sites

(ABBA+BABA) and the standard error is estimated for the D-value over all blocks with weight > 0.

The jackknifing procedure gives standard error estimates that can be used to define cutoffs for significance in case of multiple tests. For this, the distribution of neutral variation of D-values must be known or approximated. Since both ABBA and BABA counts can be interpreted as a randomly drawn subset of the true number derived sites in *i3*, both counts can be approximated by a binomial distribution, with the probability *p* equal to the expected proportion of *i3* derived sites present in *i1* and *i2*. Due to the usually large count of *i3* derived sites, this binomial distribution is approaching a normal distribution. The D-statistics is then calculated from two random samples from this distribution for ABBA and BABA counts. The distribution of D-statistics values under neutrality is thus approximately normal.

We used the site-pattern test on different combinations of individuals. We counted a total of 2040 tests when we consider all non-redundant (i.e. removing all redundant cases due to symmetry in the D-measure: D(A,B,C,D)=-D(B,A,C,D)) D-statistics using human as an outgroup and choosing the other parameters from all chimpanzee and bonobo individuals. In order to correct for multiple testing, we assumed the neutral expectation of D-values to be approximately normally distributed with mean zero and standard deviation equal to the standard error. We then used the Bonferroni correction and set our two-sided 5% confidence intervals to 4.4 standard errors deviation from zero. We calculated the Z-score (D-value divided by standard error) for each result to test against this interval.

All D-statistics were calculated on human autosomal sequence, excluding repeatmasked sequence. We also excluded over-collapsed duplicated regions in the Bonobo genome and CpG sites, as indicated by either the human, orangutan or rhesus macaque genomes.

Limitations of the D-Statistics due to Differences in Sequencing Platform

The D-statistics can give significant results due to differences in sequence or alignment. One major factor causing these differences is the choice of sequencing platform. In order to demonstrate the effect on our alignments and investigate the source of problems closer, we analyzed D-values of the structure D(Illumina-sequenced chimpanzee, Clint, human, orangutan) (see Table S10.6). The underlying assumption for this test is that no influence of ancient population structure or admixture is expected at the long range of divergence between chimpanzee and human. When we compare Illumina-sequenced chimpanzee data of at least 1x genome coverage and the Sanger-sequencing data from Clint, we observe a significantly closer relationship of human to Illumina sequences as compared to human to Clint sequences. This difference could potentially be explained by the difference in mapping or a difference in quality score filtering (we use NQS 20/15 for Sanger sequencing and 454 reads and a cutoff of Q30 for Illumina sequences).

We further investigated the effect of read-length and quality score filtering using the Sanger sequences from Clint. First, we sampled 101 base pair long sequences from Clint reads and remapped using the short read alignment algorithm BWA. The D-statistics using the shortened Sanger sequences does not show significant differences between the sequencing platforms. Thus, read-length and mapping are a major factor explaining the observed significant signal.

In a second test, we adjusted the quality score filtering of the Sanger reads to identical procedures as

used for Illumina. The matching filtering yields comparable significant results as for our default filtering procedures. We conclude that difference in sequencing platform can cause significant enrichment in the D-statistics due to difference in read length and the associated mapping errors. We therefore exclude cross-platform comparisons from our analysis.

	Standard filtering			Clint reads 101 basepairs			Clint reads Q30 filtering		
Comparison	D-value %	std. err. %	Z-score	D-value %	std. err. %	Z-score	D-value %	std. err. %	Z-score
D(CC1,Clint,Human,Orang)	-7.86%	0.71%	-11.06	0.28%	0.80%	0.35	-7.76%	0.71%	-10.87
D(CC2,Clint,Human,Orang)	-7.66%	0.83%	-9.28	-0.23%	0.93%	-0.25	-7.94%	0.82%	-9.73
D(CC3,Clint,Human,Orang)	-8.35%	0.69%	-12.17	-0.04%	0.74%	-0.06	-8.36%	0.69%	-12.16
D(CC4,Clint,Human,Orang)	-8.30%	0.77%	-10.81	-0.85%	0.86%	-0.98	-8.77%	0.79%	-11.16
D(CC5,Clint,Human,Orang)	-8.38%	0.77%	-10.92	-0.05%	0.78%	-0.06	-8.35%	0.76%	-10.95
D(CC6,Clint,Human,Orang)	-9.13%	0.79%	-11.56	0.19%	0.87%	0.22	-9.22%	0.79%	-11.61
D(EC1,Clint,Human,Orang)	-6.62%	1.10%	-6.00	0.24%	1.27%	0.19	-6.63%	1.13%	-5.88
D(EC2,Clint,Human,Orang)	-5.96%	0.86%	-6.97	1.78%	1.02%	1.75	-6.15%	0.88%	-6.98
D(EC3,Clint,Human,Orang)	-8.26%	0.73%	-11.29	-0.21%	0.83%	-0.25	-8.38%	0.74%	-11.28
D(EC4,Clint,Human,Orang)	-8.02%	0.77%	-10.42	-0.03%	0.83%	-0.04	-8.34%	0.75%	-11.09
D(EC5,Clint,Human,Orang)	-7.47%	0.73%	-10.18	-0.36%	0.90%	-0.39	-7.38%	0.76%	-9.76
D(EC6,Clint,Human,Orang)	-8.20%	0.72%	-11.31	-0.21%	0.83%	-0.25	-8.11%	0.73%	-11.05

Table S10.6: Effect of read-length and quality score filtering on D-statistics for comparison of Illumina-sequenced and Sanger-sequenced chimpanzee individuals. D-value and jackknife standard error estimates are given in per cent.

Limitations of the D-statistics due to Differences in Error Rates

In addition to a difference in sequencing platform the D-statistics may also be influenced by different error rates between individuals sequenced on the same platform. Illumina sequencing has been previously noted to differ in error rates between runs and lanes. We investigated this effect by comparing D-values for all pairwise comparisons of Illumina-sequenced Bonobo and Chimpanzee individuals to human with orangutan as outgroup (comparisons of the form D(Illumina-sequence, Illumina-sequence, Human, Orangutan)). When we compared the resulting Z-scores with the difference in error rate estimates between individuals (see SI 3), we found a weak, significant correlation[1] (Spearman's ρ=0.30, p-value<0.001). However, none of the Z-scores was significant for our cutoff of 4.4 standard errors deviation.

One potential source of the correlation between error rate differences and Z-scores can be a difference in read length. We therefore divided our set of comparisons up into 3 parts: Comparisons of 101 cycle sequenced individuals, comparisons of 76 cycle sequenced individuals and comparisons between 101 cycle and 76 cycle sequenced individuals (see Figure S10.7). We observed that comparisons of different read length data shows the strongest correlation (Spearman's ρ=0.22) followed by comparisons of 101 cycle sequenced individuals (Spearman's ρ=0.20) and 76 cycle sequenced individuals (Spearman's ρ=0.13). However, none of these individual correlations is significantly different from zero.

We conclude, that error rate differences have an impact on the D-statistics and may lead to borderline significant results. We therefore preferentially use datasets of identical read-length. Additionally, in our data

[1] The D-values (and thus the Z-scores) are symmetrical in the first two parameters: D(A,B,…) = -D(B,A,…). For the calculation of correlation, we restrict the analysis to negative Z-scores among all comparisons (no Z-score of 0 existed).

we have two sets of individuals that were indexed and sequenced on identical lanes: {B1, B2} and {B3, CC7, EC7, WC1, WC2}. This form of sequencing excludes variation in error between runs and lanes. Where possible, we re-examine our results by focusing on comparisons within these sets.

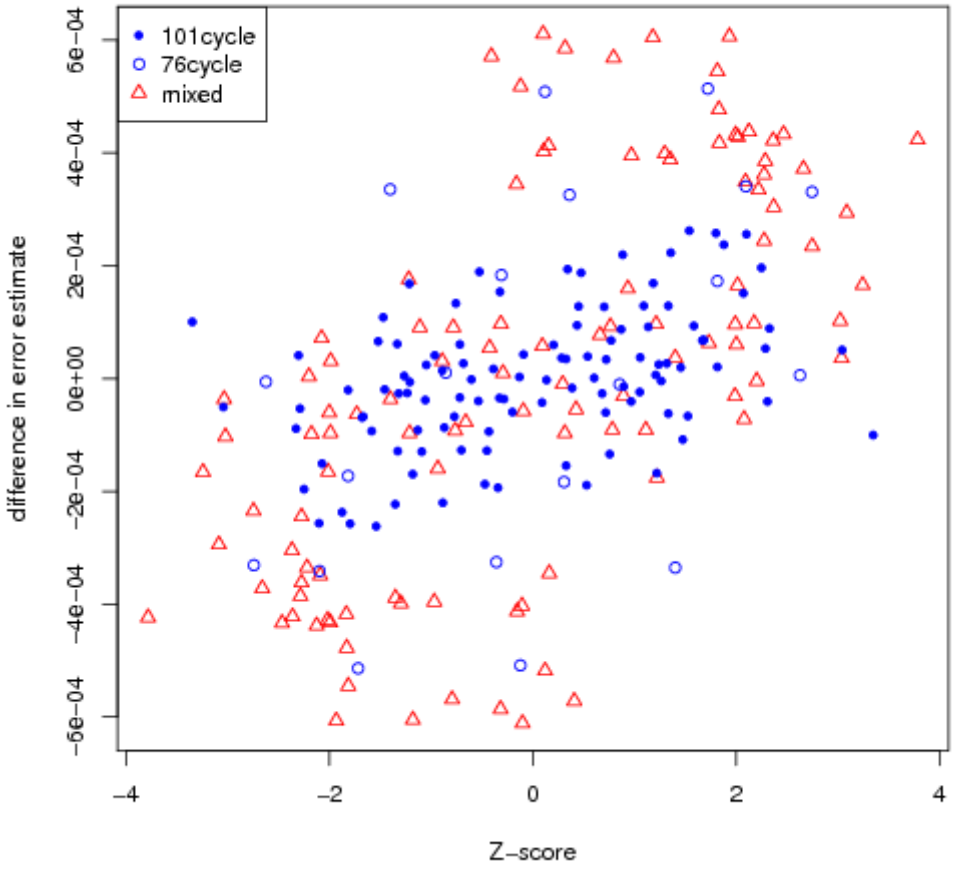


Figure S10.7: Correlation of Z-score and error-rate difference for pairwise comparisons of Illumina-sequenced individuals. Comparisons within a read-length are drawn in blue (open circles for 76 cycle, solid circles for 101 cycle read length). Comparisons between two different read length are shown as red triangles. Note that the plot has a two-fold rotational symmetry due to the symmetric D-statistics for pairwise comparisons with identical 3rd and 4th argument: D(A,B,C,D)=D(B,A,C,D).

Test for Admixture between Eastern and Central Chimpanzee and Bonobo

We first sought to test whether eastern or central chimpanzees show evidence for a closer relationship with bonobos. For this purpose, we calculated D-values in all pairwise comparisons of central and eastern chimpanzees to bonobo. The results are summarized in Table S10.7. We observe no consistent difference in relatedness to bonobo between eastern and central chimpanzees. We find, however, four central individuals that appear to share more derived alleles with Ulindi than one eastern individual. Since the signal is not supported by other comparisons with eastern individuals and is in all instances a comparison between different read length Illumina data, we consider it more likely that the signal is an artifact related to mapping or sequencing error differences.

Comparison	D-value %	std. err. %	Z-score	Comparison	D-value %	std. err. %	Z-score
D(CC1,EC1,B1,Human)	-4.3%	1.9%	-2.23	D(CC4,EC4,Ulindi1,Human)	1.3%	0.6%	2.08
D(CC1,EC1,B2,Human)	-1.7%	1.4%	-1.18	D(CC4,EC4,Ulindi2,Human)	1.1%	0.6%	1.86
D(CC1,EC1,B3,Human)	-1.7%	1.3%	-1.34	D(CC4,EC5,B1,Human)	1.8%	1.4%	1.34
D(CC1,EC1,Ulindi1,Human)	**-5.0%**	**0.9%**	**-5.54**	D(CC4,EC5,B2,Human)	1.5%	1.0%	1.60
D(CC1,EC1,Ulindi2,Human)	**-5.1%**	**0.9%**	**-5.60**	D(CC4,EC5,B3,Human)	-0.1%	0.9%	-0.11
D(CC1,EC2,B1,Human)	-0.1%	1.5%	-0.09	D(CC4,EC5,Ulindi1,Human)	0.3%	0.7%	0.47
D(CC1,EC2,B2,Human)	-0.8%	1.1%	-0.78	D(CC4,EC5,Ulindi2,Human)	0.3%	0.7%	0.50
D(CC1,EC2,B3,Human)	0.8%	0.9%	0.85	D(CC4,EC6,B1,Human)	2.0%	1.2%	1.71
D(CC1,EC2,Ulindi1,Human)	-1.2%	0.7%	-1.70	D(CC4,EC6,B2,Human)	1.2%	1.0%	1.23
D(CC1,EC2,Ulindi2,Human)	-1.3%	0.7%	-1.86	D(CC4,EC6,B3,Human)	1.6%	0.8%	2.01
D(CC1,EC3,B1,Human)	0.5%	1.2%	0.41	D(CC4,EC6,Ulindi1,Human)	0.9%	0.6%	1.51
D(CC1,EC3,B2,Human)	0.9%	0.9%	1.01	D(CC4,EC6,Ulindi2,Human)	0.9%	0.6%	1.49
D(CC1,EC3,B3,Human)	2.3%	0.8%	2.69	D(CC4,EC7,B1,Human)	1.0%	1.5%	0.67
D(CC1,EC3,Ulindi1,Human)	0.6%	0.6%	0.94	D(CC4,EC7,B2,Human)	2.1%	1.0%	2.07
D(CC1,EC3,Ulindi2,Human)	0.6%	0.6%	0.94	D(CC4,EC7,B3,Human)	2.6%	0.9%	2.79
D(CC1,EC4,B1,Human)	0.5%	1.2%	0.40	D(CC4,EC7,Ulindi1,Human)	1.5%	0.7%	2.08
D(CC1,EC4,B2,Human)	1.8%	0.9%	2.00	D(CC4,EC7,Ulindi2,Human)	1.4%	0.7%	1.95
D(CC1,EC4,B3,Human)	0.6%	0.8%	0.84	D(CC5,EC1,B1,Human)	-2.4%	1.8%	-1.30
D(CC1,EC4,Ulindi1,Human)	0.0%	0.6%	-0.05	D(CC5,EC1,B2,Human)	-1.8%	1.4%	-1.26
D(CC1,EC4,Ulindi2,Human)	0.1%	0.6%	0.14	D(CC5,EC1,B3,Human)	-2.9%	1.3%	-2.33
D(CC1,EC5,B1,Human)	0.8%	1.3%	0.60	**D(CC5,EC1,Ulindi1,Human)**	**-4.6%**	**0.9%**	**-5.02**
D(CC1,EC5,B2,Human)	-0.2%	1.0%	-0.25	**D(CC5,EC1,Ulindi2,Human)**	**-4.6%**	**0.9%**	**-5.12**
D(CC1,EC5,B3,Human)	1.0%	0.9%	1.10	D(CC5,EC2,B1,Human)	-1.2%	1.4%	-0.80
D(CC1,EC5,Ulindi1,Human)	-0.8%	0.7%	-1.28	D(CC5,EC2,B2,Human)	1.1%	1.1%	1.05
D(CC1,EC5,Ulindi2,Human)	-0.7%	0.7%	-1.04	D(CC5,EC2,B3,Human)	-1.8%	1.0%	-1.83
D(CC1,EC6,B1,Human)	1.2%	1.1%	1.11	D(CC5,EC2,Ulindi1,Human)	-2.0%	0.7%	-2.69
D(CC1,EC6,B2,Human)	0.5%	0.9%	0.53	D(CC5,EC2,Ulindi2,Human)	-2.0%	0.7%	-2.67
D(CC1,EC6,B3,Human)	1.2%	0.8%	1.62	D(CC5,EC3,B1,Human)	0.4%	1.2%	0.37
D(CC1,EC6,Ulindi1,Human)	-0.2%	0.6%	-0.38	D(CC5,EC3,B2,Human)	0.8%	0.9%	0.97
D(CC1,EC6,Ulindi2,Human)	-0.1%	0.6%	-0.23	D(CC5,EC3,B3,Human)	0.4%	0.8%	0.53
D(CC1,EC7,B1,Human)	-1.2%	1.4%	-0.89	D(CC5,EC3,Ulindi1,Human)	0.2%	0.6%	0.28
D(CC1,EC7,B2,Human)	0.4%	1.0%	0.41	D(CC5,EC3,Ulindi2,Human)	0.2%	0.6%	0.35
D(CC1,EC7,B3,Human)	2.9%	0.9%	3.15	D(CC5,EC4,B1,Human)	-0.6%	1.2%	-0.52
D(CC1,EC7,Ulindi1,Human)	-0.1%	0.7%	-0.09	D(CC5,EC4,B2,Human)	0.6%	0.9%	0.72
D(CC1,EC7,Ulindi2,Human)	-0.1%	0.7%	-0.09	D(CC5,EC4,B3,Human)	-0.7%	0.8%	-0.86
D(CC2,EC1,B1,Human)	5.1%	2.3%	2.23	D(CC5,EC4,Ulindi1,Human)	-0.5%	0.6%	-0.73
D(CC2,EC1,B2,Human)	-3.0%	1.6%	-1.88	D(CC5,EC4,Ulindi2,Human)	-0.5%	0.6%	-0.81
D(CC2,EC1,B3,Human)	-1.1%	1.5%	-0.76	D(CC5,EC5,B1,Human)	-0.6%	1.3%	-0.44
D(CC2,EC1,Ulindi1,Human)	-2.2%	1.0%	-2.10	D(CC5,EC5,B2,Human)	0.5%	1.0%	0.50
D(CC2,EC1,Ulindi2,Human)	-2.2%	1.0%	-2.16	D(CC5,EC5,B3,Human)	-0.5%	0.8%	-0.62
D(CC2,EC2,B1,Human)	0.1%	1.6%	0.05	D(CC5,EC5,Ulindi1,Human)	-1.2%	0.7%	-1.80
D(CC2,EC2,B2,Human)	-1.9%	1.2%	-1.56	D(CC5,EC5,Ulindi2,Human)	-1.1%	0.7%	-1.70
D(CC2,EC2,B3,Human)	-2.5%	1.1%	-2.19	D(CC5,EC6,B1,Human)	0.7%	1.1%	0.61
D(CC2,EC2,Ulindi1,Human)	-2.0%	0.8%	-2.35	D(CC5,EC6,B2,Human)	0.8%	0.9%	0.91
D(CC2,EC2,Ulindi2,Human)	-1.9%	0.8%	-2.25	D(CC5,EC6,B3,Human)	0.7%	0.8%	0.85
D(CC2,EC3,B1,Human)	2.3%	1.4%	1.62	D(CC5,EC6,Ulindi1,Human)	0.1%	0.6%	0.13
D(CC2,EC3,B2,Human)	-0.7%	1.0%	-0.74	D(CC5,EC6,Ulindi2,Human)	0.1%	0.6%	0.12
D(CC2,EC3,B3,Human)	1.6%	1.0%	1.62	D(CC5,EC7,B1,Human)	0.5%	1.4%	0.36
D(CC2,EC3,Ulindi1,Human)	1.1%	0.7%	1.50	D(CC5,EC7,B2,Human)	1.1%	1.0%	1.07
D(CC2,EC3,Ulindi2,Human)	1.2%	0.7%	1.64	D(CC5,EC7,B3,Human)	0.4%	1.0%	0.37
D(CC2,EC4,B1,Human)	1.2%	1.4%	0.85	D(CC5,EC7,Ulindi1,Human)	0.1%	0.7%	0.16
D(CC2,EC4,B2,Human)	-1.3%	1.1%	-1.26	D(CC5,EC7,Ulindi2,Human)	-0.1%	0.7%	-0.19
D(CC2,EC4,B3,Human)	1.8%	0.9%	1.93	D(CC6,EC1,B1,Human)	-2.4%	2.1%	-1.12
D(CC2,EC4,Ulindi1,Human)	1.1%	0.7%	1.60	D(CC6,EC1,B2,Human)	0.7%	1.5%	0.45
D(CC2,EC4,Ulindi2,Human)	1.1%	0.7%	1.55	D(CC6,EC1,B3,Human)	-1.6%	1.3%	-1.21

D(CC2,EC5,B1,Human)	3.2%	1.6%	1.98	D(CC6,EC1,Ulindi1,Human)	-2.6%	1.0%	-2.69
D(CC2,EC5,B2,Human)	-0.4%	1.2%	-0.34	D(CC6,EC1,Ulindi2,Human)	-2.7%	1.0%	-2.80
D(CC2,EC5,B3,Human)	0.1%	1.0%	0.06	D(CC6,EC2,B1,Human)	-0.4%	1.6%	-0.24
D(CC2,EC5,Ulindi1,Human)	-0.1%	0.8%	-0.10	D(CC6,EC2,B2,Human)	-0.3%	1.2%	-0.27
D(CC2,EC5,Ulindi2,Human)	-0.2%	0.8%	-0.21	D(CC6,EC2,B3,Human)	-2.1%	1.1%	-1.95
D(CC2,EC6,B1,Human)	1.6%	1.4%	1.16	D(CC6,EC2,Ulindi1,Human)	-1.4%	0.8%	-1.74
D(CC2,EC6,B2,Human)	0.6%	1.0%	0.60	D(CC6,EC2,Ulindi2,Human)	-1.7%	0.8%	-2.19
D(CC2,EC6,B3,Human)	1.0%	0.9%	1.14	D(CC6,EC3,B1,Human)	-0.7%	1.3%	-0.56
D(CC2,EC6,Ulindi1,Human)	1.4%	0.7%	2.05	D(CC6,EC3,B2,Human)	-0.1%	0.9%	-0.09
D(CC2,EC6,Ulindi2,Human)	1.3%	0.7%	2.02	D(CC6,EC3,B3,Human)	0.4%	0.8%	0.45
D(CC2,EC7,B1,Human)	1.6%	1.6%	0.96	D(CC6,EC3,Ulindi1,Human)	0.4%	0.7%	0.55
D(CC2,EC7,B2,Human)	0.2%	1.2%	0.18	D(CC6,EC3,Ulindi2,Human)	0.6%	0.6%	0.86
D(CC2,EC7,B3,Human)	2.9%	1.1%	2.72	D(CC6,EC4,B1,Human)	-0.9%	1.3%	-0.71
D(CC2,EC7,Ulindi1,Human)	2.4%	0.8%	2.93	D(CC6,EC4,B2,Human)	0.8%	1.0%	0.77
D(CC2,EC7,Ulindi2,Human)	2.3%	0.8%	2.86	D(CC6,EC4,B3,Human)	0.5%	0.9%	0.54
D(CC3,EC1,B1,Human)	-1.6%	1.8%	-0.85	D(CC6,EC4,Ulindi1,Human)	1.0%	0.6%	1.49
D(CC3,EC1,B2,Human)	-3.0%	1.4%	-2.19	D(CC6,EC4,Ulindi2,Human)	0.9%	0.6%	1.38
D(CC3,EC1,B3,Human)	-2.5%	1.2%	-2.00	D(CC6,EC5,B1,Human)	-0.3%	1.3%	-0.26
D(CC3,EC1,Ulindi1,Human)	**-4.9%**	**0.9%**	**-5.52**	D(CC6,EC5,B2,Human)	0.1%	1.0%	0.11
D(CC3,EC1,Ulindi2,Human)	**-5.1%**	**0.9%**	**-5.87**	D(CC6,EC5,B3,Human)	-0.1%	0.9%	-0.08
D(CC3,EC2,B1,Human)	0.3%	1.5%	0.23	D(CC6,EC5,Ulindi1,Human)	-0.1%	0.7%	-0.12
D(CC3,EC2,B2,Human)	-0.7%	1.0%	-0.67	D(CC6,EC5,Ulindi2,Human)	0.0%	0.7%	-0.05
D(CC3,EC2,B3,Human)	-1.7%	1.0%	-1.67	D(CC6,EC6,B1,Human)	0.0%	1.2%	-0.03
D(CC3,EC2,Ulindi1,Human)	-2.4%	0.7%	-3.27	D(CC6,EC6,B2,Human)	-0.2%	1.0%	-0.21
D(CC3,EC2,Ulindi2,Human)	-2.4%	0.7%	-3.31	D(CC6,EC6,B3,Human)	0.5%	0.9%	0.53
D(CC3,EC3,B1,Human)	0.2%	1.2%	0.17	D(CC6,EC6,Ulindi1,Human)	0.5%	0.7%	0.69
D(CC3,EC3,B2,Human)	-1.2%	0.9%	-1.29	D(CC6,EC6,Ulindi2,Human)	0.4%	0.7%	0.59
D(CC3,EC3,B3,Human)	-0.2%	0.8%	-0.29	D(CC6,EC7,B1,Human)	-0.9%	1.5%	-0.58
D(CC3,EC3,Ulindi1,Human)	-0.9%	0.6%	-1.43	D(CC6,EC7,B2,Human)	1.1%	1.2%	0.94
D(CC3,EC3,Ulindi2,Human)	-0.8%	0.6%	-1.30	D(CC6,EC7,B3,Human)	1.1%	1.0%	1.08
D(CC3,EC4,B1,Human)	1.5%	1.2%	1.23	D(CC6,EC7,Ulindi1,Human)	0.4%	0.8%	0.50
D(CC3,EC4,B2,Human)	-0.4%	0.9%	-0.50	D(CC6,EC7,Ulindi2,Human)	0.2%	0.8%	0.22
D(CC3,EC4,B3,Human)	-0.7%	0.8%	-0.84	D(CC7,EC1,B1,Human)	-3.6%	2.3%	-1.56
D(CC3,EC4,Ulindi1,Human)	-0.7%	0.6%	-1.24	D(CC7,EC1,B2,Human)	-2.4%	1.6%	-1.44
D(CC3,EC4,Ulindi2,Human)	-0.9%	0.6%	-1.53	D(CC7,EC1,B3,Human)	-2.9%	1.5%	-1.89
D(CC3,EC5,B1,Human)	0.2%	1.3%	0.17	**D(CC7,EC1,Ulindi1,Human)**	**-5.0%**	**1.1%**	**-4.53**
D(CC3,EC5,B2,Human)	-0.2%	1.0%	-0.17	**D(CC7,EC1,Ulindi2,Human)**	**-5.1%**	**1.1%**	**-4.70**
D(CC3,EC5,B3,Human)	-0.4%	0.9%	-0.43	D(CC7,EC2,B1,Human)	-3.2%	1.8%	-1.82
D(CC3,EC5,Ulindi1,Human)	-1.7%	0.7%	-2.55	D(CC7,EC2,B2,Human)	0.2%	1.3%	0.14
D(CC3,EC5,Ulindi2,Human)	-1.7%	0.6%	-2.68	D(CC7,EC2,B3,Human)	-0.5%	1.2%	-0.44
D(CC3,EC6,B1,Human)	-0.1%	1.1%	-0.13	D(CC7,EC2,Ulindi1,Human)	-1.9%	0.9%	-2.11
D(CC3,EC6,B2,Human)	0.6%	0.8%	0.70	D(CC7,EC2,Ulindi2,Human)	-2.1%	0.9%	-2.42
D(CC3,EC6,B3,Human)	0.8%	0.8%	1.00	D(CC7,EC3,B1,Human)	-1.1%	1.5%	-0.75
D(CC3,EC6,Ulindi1,Human)	-0.5%	0.6%	-0.75	D(CC7,EC3,B2,Human)	0.0%	1.1%	0.01
D(CC3,EC6,Ulindi2,Human)	-0.7%	0.6%	-1.09	D(CC7,EC3,B3,Human)	1.3%	1.0%	1.29
D(CC3,EC7,B1,Human)	1.4%	1.4%	1.03	D(CC7,EC3,Ulindi1,Human)	0.5%	0.7%	0.75
D(CC3,EC7,B2,Human)	0.3%	1.0%	0.28	D(CC7,EC3,Ulindi2,Human)	0.5%	0.7%	0.74
D(CC3,EC7,B3,Human)	1.2%	0.9%	1.35	D(CC7,EC4,B1,Human)	-0.9%	1.5%	-0.60
D(CC3,EC7,Ulindi1,Human)	-0.6%	0.7%	-0.95	D(CC7,EC4,B2,Human)	0.6%	1.1%	0.54
D(CC3,EC7,Ulindi2,Human)	-0.7%	0.7%	-1.09	D(CC7,EC4,B3,Human)	1.1%	1.0%	1.12
D(CC4,EC1,B1,Human)	-1.1%	2.0%	-0.55	D(CC7,EC4,Ulindi1,Human)	0.4%	0.7%	0.48
D(CC4,EC1,B2,Human)	-0.4%	1.4%	-0.31	D(CC7,EC4,Ulindi2,Human)	0.4%	0.7%	0.58
D(CC4,EC1,B3,Human)	-1.1%	1.2%	-0.87	D(CC7,EC5,B1,Human)	-1.8%	1.6%	-1.14
D(CC4,EC1,Ulindi1,Human)	-3.1%	0.9%	-3.51	D(CC7,EC5,B2,Human)	-0.2%	1.2%	-0.13
D(CC4,EC1,Ulindi2,Human)	-3.3%	0.9%	-3.67	D(CC7,EC5,B3,Human)	0.3%	1.1%	0.31
D(CC4,EC2,B1,Human)	2.6%	1.4%	1.83	D(CC7,EC5,Ulindi1,Human)	-1.2%	0.8%	-1.56
D(CC4,EC2,B2,Human)	0.7%	1.1%	0.63	D(CC7,EC5,Ulindi2,Human)	-1.1%	0.8%	-1.48

D(CC4,EC2,B3,Human)	0.5%	1.0%	0.45	D(CC7,EC6,B1,Human)	-0.9%	1.4%	-0.62
D(CC4,EC2,Ulindi1,Human)	-0.9%	0.7%	-1.16	D(CC7,EC6,B2,Human)	-1.3%	1.0%	-1.25
D(CC4,EC2,Ulindi2,Human)	-1.1%	0.7%	-1.50	D(CC7,EC6,B3,Human)	-0.2%	0.9%	-0.17
D(CC4,EC3,B1,Human)	0.7%	1.2%	0.58	D(CC7,EC6,Ulindi1,Human)	-0.7%	0.7%	-1.05
D(CC4,EC3,B2,Human)	0.6%	0.9%	0.62	D(CC7,EC6,Ulindi2,Human)	-1.0%	0.7%	-1.41
D(CC4,EC3,B3,Human)	0.7%	0.8%	0.91	D(CC7,EC7,B1,Human)	2.5%	1.7%	1.51
D(CC4,EC3,Ulindi1,Human)	0.6%	0.6%	1.01	D(CC7,EC7,B2,Human)	1.6%	1.2%	1.31
D(CC4,EC3,Ulindi2,Human)	0.6%	0.6%	1.05	D(CC7,EC7,B3,Human)	3.2%	1.1%	2.92
D(CC4,EC4,B1,Human)	0.9%	1.3%	0.68	D(CC7,EC7,Ulindi1,Human)	2.1%	0.8%	2.55
D(CC4,EC4,B2,Human)	2.0%	0.9%	2.19	D(CC7,EC7,Ulindi2,Human)	2.0%	0.8%	2.45
D(CC4,EC4,B3,Human)	1.9%	0.8%	2.32				

Table S10.7: D-statistics for comparison with eastern and western chimpanzee to bonobo. D-value and jackknife standard error estimates are given in per cent. Rows with a significant enrichment are marked in bold and comparisons between individuals sequenced on identical lanes are shown with green background.

Comparison between Western and Central/Eastern Chimpanzees to Bonobo

We furthered our analysis by testing whether western chimpanzees differ from eastern or central chimpanzee in their relationship to bonobos. Except for a comparison between individuals of different read length, we find no individual signal supporting a significantly closer relationship between western chimpanzees and bonobos as compared to eastern and central chimpanzee individuals (see Table S10.8). However, a total of 62 out of 70 comparisons between western and central (p-value=1.8 x 10^{-11}; binomial test with p=0.5) and 63 out of 70 comparisons between western and eastern individuals (p-value=2.2 x 10^{-12}; binomial test with p=0.5) give a trend towards closer relationship (D>0) between western individuals and bonobo. This result also holds when the analysis is restricted to individuals with 101 cycle read length (eastern comparison: 45/50, p-value= 4.2 x 10^{-9}; central comparison: 52/60, p-value=5.2 x 10^{-9}). When we restrict the analysis to individuals sequenced on identical lanes (CC7, EC7, WC1 and WC2) the trend remains, but is only significant in the comparison with central individuals (eastern comparison: 8/10, p-value=0.11; central comparison: 10/10, p-value=0.002).

Comparison	D-value %	std. err. %	Z-score	Comparison	D-value %	std. err. %	Z-score
D(CC1,WC1,B1,Human)	-1.4%	1.3%	-1.08	D(CC1,WC2,B1,Human)	0.8%	1.3%	0.60
D(CC1,WC1,B2,Human)	1.7%	1.0%	1.80	D(CC1,WC2,B2,Human)	0.4%	1.0%	0.39
D(CC1,WC1,B3,Human)	1.8%	0.9%	2.10	D(CC1,WC2,B3,Human)	2.8%	0.9%	3.17
D(CC1,WC1,Ulindi1,Human)	0.7%	0.7%	0.96	D(CC1,WC2,Ulindi1,Human)	1.1%	0.7%	1.57
D(CC1,WC1,Ulindi2,Human)	0.8%	0.7%	1.18	D(CC1,WC2,Ulindi2,Human)	1.2%	0.7%	1.68
D(CC2,WC1,B1,Human)	0.1%	1.5%	0.09	D(CC2,WC2,B1,Human)	4.5%	1.5%	2.98
D(CC2,WC1,B2,Human)	1.5%	1.2%	1.29	D(CC2,WC2,B2,Human)	0.8%	1.2%	0.62
D(CC2,WC1,B3,Human)	1.1%	1.0%	1.04	D(CC2,WC2,B3,Human)	2.5%	1.1%	2.26
D(CC2,WC1,Ulindi1,Human)	1.8%	0.8%	2.34	D(CC2,WC2,Ulindi1,Human)	2.8%	0.8%	3.43
D(CC2,WC1,Ulindi2,Human)	2.0%	0.7%	2.65	D(CC2,WC2,Ulindi2,Human)	2.9%	0.8%	3.60
D(CC3,WC1,B1,Human)	-0.2%	1.3%	-0.12	D(CC3,WC2,B1,Human)	0.0%	1.4%	0.00
D(CC3,WC1,B2,Human)	0.8%	0.9%	0.81	D(CC3,WC2,B2,Human)	0.4%	1.0%	0.41
D(CC3,WC1,B3,Human)	0.8%	0.8%	0.89	D(CC3,WC2,B3,Human)	1.2%	0.9%	1.30
D(CC3,WC1,Ulindi1,Human)	-0.4%	0.7%	-0.55	D(CC3,WC2,Ulindi1,Human)	0.2%	0.7%	0.28
D(CC3,WC1,Ulindi2,Human)	-0.4%	0.7%	-0.63	D(CC3,WC2,Ulindi2,Human)	0.1%	0.7%	0.14
D(CC4,WC1,B1,Human)	-0.5%	1.3%	-0.39	D(CC4,WC2,B1,Human)	1.0%	1.5%	0.69
D(CC4,WC1,B2,Human)	1.2%	1.0%	1.18	D(CC4,WC2,B2,Human)	1.2%	1.1%	1.14

D(CC4,WC1,B3,Human)	0.7%	0.9%	0.82	D(CC4,WC2,B3,Human)	1.2%	0.9%	1.25
D(CC4,WC1,Ulindi1,Human)	1.0%	0.7%	1.45	D(CC4,WC2,Ulindi1,Human)	1.0%	0.7%	1.42
D(CC4,WC1,Ulindi2,Human)	1.0%	0.7%	1.57	D(CC4,WC2,Ulindi2,Human)	0.9%	0.7%	1.25
D(CC5,WC1,B1,Human)	0.0%	1.3%	0.00	D(CC5,WC2,B1,Human)	1.6%	1.4%	1.14
D(CC5,WC1,B2,Human)	2.2%	1.0%	2.30	D(CC5,WC2,B2,Human)	0.4%	1.0%	0.40
D(CC5,WC1,B3,Human)	0.2%	0.9%	0.24	D(CC5,WC2,B3,Human)	0.4%	0.9%	0.44
D(CC5,WC1,Ulindi1,Human)	0.7%	0.7%	0.95	D(CC5,WC2,Ulindi1,Human)	0.6%	0.7%	0.79
D(CC5,WC1,Ulindi2,Human)	0.8%	0.7%	1.13	D(CC5,WC2,Ulindi2,Human)	0.6%	0.7%	0.80
D(CC6,WC1,B1,Human)	-0.1%	1.3%	-0.08	D(CC6,WC2,B1,Human)	0.8%	1.5%	0.54
D(CC6,WC1,B2,Human)	1.9%	1.0%	1.85	D(CC6,WC2,B2,Human)	2.1%	1.0%	2.03
D(CC6,WC1,B3,Human)	0.7%	0.9%	0.72	D(CC6,WC2,B3,Human)	1.1%	1.0%	1.12
D(CC6,WC1,Ulindi1,Human)	0.9%	0.7%	1.22	D(CC6,WC2,Ulindi1,Human)	2.0%	0.8%	2.56
D(CC6,WC1,Ulindi2,Human)	1.0%	0.7%	1.34	D(CC6,WC2,Ulindi2,Human)	1.9%	0.8%	2.43
D(CC7,WC1,B1,Human)	0.2%	1.5%	0.15	D(CC7,WC2,B1,Human)	1.1%	1.6%	0.68
D(CC7,WC1,B2,Human)	1.3%	1.1%	1.15	D(CC7,WC2,B2,Human)	1.2%	1.2%	0.95
D(CC7,WC1,B3,Human)	1.1%	1.0%	1.16	D(CC7,WC2,B3,Human)	1.5%	1.1%	1.40
D(CC7,WC1,Ulindi1,Human)	1.5%	0.7%	2.01	D(CC7,WC2,Ulindi1,Human)	1.0%	0.8%	1.13
D(CC7,WC1,Ulindi2,Human)	1.3%	0.7%	1.81	D(CC7,WC2,Ulindi2,Human)	0.7%	0.8%	0.82
D(EC1,WC1,B1,Human)	2.0%	2.0%	0.98	D(EC1,WC2,B1,Human)	-0.5%	2.2%	-0.24
D(EC1,WC1,B2,Human)	4.0%	1.5%	2.66	D(EC1,WC2,B2,Human)	0.0%	1.7%	-0.03
D(EC1,WC1,B3,Human)	5.3%	1.4%	3.85	D(EC1,WC2,B3,Human)	3.2%	1.5%	2.09
D(EC1,WC1,Ulindi1,Human)	**4.9%**	**1.0%**	**4.86**	D(EC1,WC2,Ulindi1,Human)	4.0%	1.1%	3.84
D(EC1,WC1,Ulindi2,Human)	**5.0%**	**1.0%**	**5.05**	D(EC1,WC2,Ulindi2,Human)	3.9%	1.0%	3.68
D(EC2,WC1,B1,Human)	2.4%	1.6%	1.50	D(EC2,WC2,B1,Human)	2.6%	1.8%	1.46
D(EC2,WC1,B2,Human)	2.1%	1.2%	1.82	D(EC2,WC2,B2,Human)	1.1%	1.3%	0.91
D(EC2,WC1,B3,Human)	1.3%	1.1%	1.23	D(EC2,WC2,B3,Human)	2.6%	1.2%	2.13
D(EC2,WC1,Ulindi1,Human)	2.7%	0.8%	3.30	D(EC2,WC2,Ulindi1,Human)	2.5%	0.8%	2.99
D(EC2,WC1,Ulindi2,Human)	2.9%	0.8%	3.47	D(EC2,WC2,Ulindi2,Human)	2.5%	0.8%	2.91
D(EC3,WC1,B1,Human)	0.5%	1.3%	0.36	D(EC3,WC2,B1,Human)	1.2%	1.4%	0.81
D(EC3,WC1,B2,Human)	1.8%	1.0%	1.89	D(EC3,WC2,B2,Human)	1.8%	1.0%	1.75
D(EC3,WC1,B3,Human)	0.6%	0.9%	0.65	D(EC3,WC2,B3,Human)	0.5%	1.0%	0.48
D(EC3,WC1,Ulindi1,Human)	0.8%	0.7%	1.08	D(EC3,WC2,Ulindi1,Human)	0.9%	0.7%	1.27
D(EC3,WC1,Ulindi2,Human)	0.7%	0.7%	0.97	D(EC3,WC2,Ulindi2,Human)	0.7%	0.7%	0.90
D(EC4,WC1,B1,Human)	0.8%	1.3%	0.58	D(EC4,WC2,B1,Human)	0.0%	1.5%	-0.01
D(EC4,WC1,B2,Human)	2.9%	1.0%	2.89	D(EC4,WC2,B2,Human)	1.1%	1.1%	1.05
D(EC4,WC1,B3,Human)	0.7%	1.0%	0.69	D(EC4,WC2,B3,Human)	1.3%	1.1%	1.23
D(EC4,WC1,Ulindi1,Human)	1.3%	0.7%	1.88	D(EC4,WC2,Ulindi1,Human)	1.3%	0.8%	1.64
D(EC4,WC1,Ulindi2,Human)	1.4%	0.7%	2.04	D(EC4,WC2,Ulindi2,Human)	1.2%	0.8%	1.54
D(EC5,WC1,B1,Human)	0.4%	1.5%	0.25	D(EC5,WC2,B1,Human)	0.5%	1.6%	0.30
D(EC5,WC1,B2,Human)	1.0%	1.1%	0.91	D(EC5,WC2,B2,Human)	0.1%	1.2%	0.10
D(EC5,WC1,B3,Human)	1.3%	1.0%	1.24	D(EC5,WC2,B3,Human)	1.1%	1.0%	1.07
D(EC5,WC1,Ulindi1,Human)	1.8%	0.8%	2.28	D(EC5,WC2,Ulindi1,Human)	1.5%	0.8%	1.85
D(EC5,WC1,Ulindi2,Human)	1.8%	0.8%	2.35	D(EC5,WC2,Ulindi2,Human)	1.5%	0.8%	1.91
D(EC6,WC1,B1,Human)	-0.3%	1.3%	-0.26	D(EC6,WC2,B1,Human)	-0.2%	1.4%	-0.13
D(EC6,WC1,B2,Human)	1.3%	0.9%	1.42	D(EC6,WC2,B2,Human)	0.9%	1.1%	0.87
D(EC6,WC1,B3,Human)	0.3%	0.8%	0.38	D(EC6,WC2,B3,Human)	0.9%	1.0%	0.89
D(EC6,WC1,Ulindi1,Human)	1.0%	0.7%	1.50	D(EC6,WC2,Ulindi1,Human)	1.1%	0.7%	1.41
D(EC6,WC1,Ulindi2,Human)	1.2%	0.7%	1.77	D(EC6,WC2,Ulindi2,Human)	1.1%	0.8%	1.48
D(EC7,WC1,B1,Human)	-1.7%	1.6%	-1.05	D(EC7,WC2,B1,Human)	1.5%	1.7%	0.91
D(EC7,WC1,B2,Human)	0.6%	1.1%	0.56	D(EC7,WC2,B2,Human)	0.8%	1.2%	0.64
D(EC7,WC1,B3,Human)	-1.4%	1.0%	-1.38	D(EC7,WC2,B3,Human)	0.6%	1.1%	0.52
D(EC7,WC1,Ulindi1,Human)	0.0%	0.8%	0.05	D(EC7,WC2,Ulindi1,Human)	1.2%	0.9%	1.39
D(EC7,WC1,Ulindi2,Human)	0.3%	0.8%	0.33	D(EC7,WC2,Ulindi2,Human)	1.5%	0.9%	1.68

Table S10.8: D-statistics for comparison of westerm individuals to eastern and central individuals in their relationship to bonobo. D-value and jackknife standard error estimates are given in per cent. Significant

results are shown in bold font. Comparisons of individuals sequenced on the same lanes are highlighted green.

Test for Admixture between Bonobo Individuals and Chimpanzee Subgroups

We also used the site-pattern test to investigate the relationship of the sequenced bonobo individuals to western, central and eastern chimpanzee subgroups. Table S10.9 summarizes the results. We find no evidence for a closer relationship of bonobo individuals to any chimpanzee subgroup.

Comparison	D-value %	std. err. %	Z-score	Comparison	D-value %	std. err %.	Z-score
D(B1,B2,CC1,Human)	-0.3%	2.4%	-0.11	D(B2,Ulindi2,CC1,Human)	-1.0%	1.2%	-0.83
D(B1,B2,CC2,Human)	1.4%	2.9%	0.50	D(B2,Ulindi2,CC2,Human)	-2.3%	1.4%	-1.67
D(B1,B2,CC3,Human)	0.5%	2.3%	0.22	D(B2,Ulindi2,CC3,Human)	1.9%	1.1%	1.70
D(B1,B2,CC4,Human)	-1.0%	2.5%	-0.41	D(B2,Ulindi2,CC4,Human)	0.7%	1.3%	0.53
D(B1,B2,CC5,Human)	-1.0%	2.4%	-0.39	D(B2,Ulindi2,CC5,Human)	1.6%	1.2%	1.32
D(B1,B2,CC6,Human)	-4.8%	2.5%	-1.89	D(B2,Ulindi2,CC6,Human)	0.9%	1.3%	0.69
D(B1,B2,CC7,Human)	0.4%	2.8%	0.14	D(B2,Ulindi2,CC7,Human)	-2.0%	1.5%	-1.36
D(B1,B2,EC1,Human)	7.3%	3.9%	1.90	D(B2,Ulindi2,EC1,Human)	-4.4%	1.9%	-2.33
D(B1,B2,EC2,Human)	-1.2%	2.9%	-0.43	D(B2,Ulindi2,EC2,Human)	-2.9%	1.5%	-1.90
D(B1,B2,EC3,Human)	-1.1%	2.7%	-0.40	D(B2,Ulindi2,EC3,Human)	-0.2%	1.2%	-0.18
D(B1,B2,EC4,Human)	-4.5%	2.5%	-1.79	D(B2,Ulindi2,EC4,Human)	1.1%	1.3%	0.89
D(B1,B2,EC5,Human)	0.5%	2.8%	0.18	D(B2,Ulindi2,EC5,Human)	-1.5%	1.3%	-1.10
D(B1,B2,EC6,Human)	-5.1%	2.5%	-2.07	D(B2,Ulindi2,EC6,Human)	0.8%	1.2%	0.66
D(B1,B2,EC7,Human)	-0.4%	2.9%	-0.14	D(B2,Ulindi2,EC7,Human)	0.1%	1.4%	0.09
D(B1,B2,WC1,Human)	1.1%	2.7%	0.39	D(B2,Ulindi2,WC1,Human)	3.3%	1.3%	2.50
D(B1,B2,WC2,Human)	-2.6%	2.9%	-0.87	D(B2,Ulindi2,WC2,Human)	2.6%	1.4%	1.82
D(B1,B3,CC1,Human)	-2.0%	2.4%	-0.84	D(B3,B2,CC1,Human)	-2.8%	1.7%	-1.61
D(B1,B3,CC2,Human)	-0.5%	3.0%	-0.18	D(B3,B2,CC2,Human)	-1.6%	2.0%	-0.80
D(B1,B3,CC3,Human)	5.0%	2.3%	2.19	D(B3,B2,CC3,Human)	-3.2%	1.6%	-1.96
D(B1,B3,CC4,Human)	-0.6%	2.5%	-0.23	D(B3,B2,CC4,Human)	-2.4%	1.7%	-1.37
D(B1,B3,CC5,Human)	0.7%	2.6%	0.26	D(B3,B2,CC5,Human)	-3.7%	1.7%	-2.21
D(B1,B3,CC6,Human)	-1.1%	2.5%	-0.43	D(B3,B2,CC6,Human)	-5.2%	1.8%	-2.87
D(B1,B3,CC7,Human)	0.4%	2.9%	0.14	D(B3,B2,CC7,Human)	-4.1%	2.0%	-2.09
D(B1,B3,EC1,Human)	0.5%	4.0%	0.11	D(B3,B2,EC1,Human)	-2.8%	2.8%	-1.01
D(B1,B3,EC2,Human)	-2.6%	3.2%	-0.83	D(B3,B2,EC2,Human)	2.3%	2.1%	1.14
D(B1,B3,EC3,Human)	-1.2%	2.5%	-0.46	D(B3,B2,EC3,Human)	-2.1%	1.7%	-1.29
D(B1,B3,EC4,Human)	-2.8%	2.5%	-1.13	D(B3,B2,EC4,Human)	-3.3%	1.7%	-1.90
D(B1,B3,EC5,Human)	1.6%	2.6%	0.60	D(B3,B2,EC5,Human)	-2.8%	1.9%	-1.50
D(B1,B3,EC6,Human)	3.1%	2.3%	1.34	D(B3,B2,EC6,Human)	-3.7%	1.7%	-2.13
D(B1,B3,EC7,Human)	-1.4%	2.8%	-0.51	D(B3,B2,EC7,Human)	-3.0%	2.0%	-1.52
D(B1,B3,WC1,Human)	-1.1%	2.6%	-0.44	D(B3,B2,WC1,Human)	-3.8%	1.8%	-2.14
D(B1,B3,WC2,Human)	4.3%	2.8%	1.52	D(B3,B2,WC2,Human)	-4.5%	2.0%	-2.28
D(B1,Ulindi1,CC1,Human)	-0.5%	1.6%	-0.31	D(B3,Ulindi1,CC1,Human)	-2.3%	1.2%	-1.88
D(B1,Ulindi1,CC2,Human)	-1.5%	1.8%	-0.83	D(B3,Ulindi1,CC2,Human)	-0.8%	1.3%	-0.63
D(B1,Ulindi1,CC3,Human)	3.0%	1.6%	1.94	D(B3,Ulindi1,CC3,Human)	-2.0%	1.1%	-1.85
D(B1,Ulindi1,CC4,Human)	-1.3%	1.7%	-0.73	D(B3,Ulindi1,CC4,Human)	-1.9%	1.2%	-1.59
D(B1,Ulindi1,CC5,Human)	0.6%	1.6%	0.40	D(B3,Ulindi1,CC5,Human)	-1.2%	1.2%	-1.06
D(B1,Ulindi1,CC6,Human)	-0.9%	1.8%	-0.47	D(B3,Ulindi1,CC6,Human)	-3.0%	1.3%	-2.39
D(B1,Ulindi1,CC7,Human)	-1.8%	2.0%	-0.88	D(B3,Ulindi1,CC7,Human)	-4.3%	1.3%	-3.21
D(B1,Ulindi1,EC1,Human)	-5.4%	2.8%	-1.94	D(B3,Ulindi1,EC1,Human)	-5.8%	1.8%	-3.18
D(B1,Ulindi1,EC2,Human)	-2.8%	2.1%	-1.32	D(B3,Ulindi1,EC2,Human)	-1.8%	1.5%	-1.22
D(B1,Ulindi1,EC3,Human)	-0.4%	1.7%	-0.24	D(B3,Ulindi1,EC3,Human)	-2.9%	1.1%	-2.51
D(B1,Ulindi1,EC4,Human)	-0.1%	1.7%	-0.03	D(B3,Ulindi1,EC4,Human)	-2.2%	1.2%	-1.89
D(B1,Ulindi1,EC5,Human)	-2.1%	1.9%	-1.12	D(B3,Ulindi1,EC5,Human)	-3.4%	1.2%	-2.72

D(B1,Ulindi1,EC6,Human)	0.7%	1.7%	0.41	D(B3,Ulindi1,EC6,Human)	-2.3%	1.2%	-1.94
D(B1,Ulindi1,EC7,Human)	1.6%	1.9%	0.83	D(B3,Ulindi1,EC7,Human)	-1.2%	1.4%	-0.86
D(B1,Ulindi1,WC1,Human)	0.7%	1.8%	0.37	D(B3,Ulindi1,WC1,Human)	-2.4%	1.3%	-1.93
D(B1,Ulindi1,WC2,Human)	0.9%	2.0%	0.43	D(B3,Ulindi1,WC2,Human)	-3.3%	1.4%	-2.39
D(B1,Ulindi2,CC1,Human)	-2.2%	1.7%	-1.31	D(B3,Ulindi2,CC1,Human)	-2.8%	1.2%	-2.43
D(B1,Ulindi2,CC2,Human)	-2.1%	1.9%	-1.11	D(B3,Ulindi2,CC2,Human)	-0.2%	1.3%	-0.16
D(B1,Ulindi2,CC3,Human)	1.9%	1.7%	1.12	D(B3,Ulindi2,CC3,Human)	-2.3%	1.1%	-2.08
D(B1,Ulindi2,CC4,Human)	1.4%	1.6%	0.84	D(B3,Ulindi2,CC4,Human)	-1.2%	1.2%	-1.02
D(B1,Ulindi2,CC5,Human)	1.8%	1.7%	1.06	D(B3,Ulindi2,CC5,Human)	-1.8%	1.1%	-1.57
D(B1,Ulindi2,CC6,Human)	-0.5%	1.8%	-0.27	D(B3,Ulindi2,CC6,Human)	-1.9%	1.2%	-1.56
D(B1,Ulindi2,CC7,Human)	-3.2%	2.0%	-1.62	D(B3,Ulindi2,CC7,Human)	-3.4%	1.3%	-2.61
D(B1,Ulindi2,EC1,Human)	-6.6%	2.6%	-2.53	D(B3,Ulindi2,EC1,Human)	-5.4%	1.8%	-2.96
D(B1,Ulindi2,EC2,Human)	-2.5%	2.1%	-1.18	D(B3,Ulindi2,EC2,Human)	-3.5%	1.4%	-2.61
D(B1,Ulindi2,EC3,Human)	-0.4%	1.7%	-0.25	D(B3,Ulindi2,EC3,Human)	-2.5%	1.1%	-2.20
D(B1,Ulindi2,EC4,Human)	-0.9%	1.6%	-0.54	D(B3,Ulindi2,EC4,Human)	-2.1%	1.1%	-1.91
D(B1,Ulindi2,EC5,Human)	-3.3%	1.8%	-1.83	D(B3,Ulindi2,EC5,Human)	-3.8%	1.2%	-3.08
D(B1,Ulindi2,EC6,Human)	-1.4%	1.6%	-0.86	D(B3,Ulindi2,EC6,Human)	-2.4%	1.1%	-2.16
D(B1,Ulindi2,EC7,Human)	0.2%	1.9%	0.11	D(B3,Ulindi2,EC7,Human)	-2.4%	1.3%	-1.85
D(B1,Ulindi2,WC1,Human)	1.3%	1.7%	0.75	D(B3,Ulindi2,WC1,Human)	-2.2%	1.2%	-1.86
D(B1,Ulindi2,WC2,Human)	2.2%	2.0%	1.09	D(B3,Ulindi2,WC2,Human)	-4.7%	1.3%	-3.46
D(B2,Ulindi1,CC1,Human)	-0.4%	1.2%	-0.31	D(Ulindi1,Ulindi2,CC1,Human)	-1.3%	0.7%	-1.77
D(B2,Ulindi1,CC2,Human)	-1.4%	1.4%	-1.05	D(Ulindi1,Ulindi2,CC2,Human)	-0.4%	0.9%	-0.41
D(B2,Ulindi1,CC3,Human)	2.1%	1.1%	1.89	D(Ulindi1,Ulindi2,CC3,Human)	-0.5%	0.7%	-0.66
D(B2,Ulindi1,CC4,Human)	0.3%	1.3%	0.25	D(Ulindi1,Ulindi2,CC4,Human)	-0.1%	0.8%	-0.10
D(B2,Ulindi1,CC5,Human)	2.4%	1.2%	2.01	D(Ulindi1,Ulindi2,CC5,Human)	-1.1%	0.8%	-1.43
D(B2,Ulindi1,CC6,Human)	0.5%	1.3%	0.40	D(Ulindi1,Ulindi2,CC6,Human)	-0.6%	0.9%	-0.73
D(B2,Ulindi1,CC7,Human)	-0.8%	1.4%	-0.58	D(Ulindi1,Ulindi2,CC7,Human)	-1.4%	0.9%	-1.55
D(B2,Ulindi1,EC1,Human)	-3.5%	1.8%	-1.91	D(Ulindi1,Ulindi2,EC1,Human)	0.7%	1.3%	0.52
D(B2,Ulindi1,EC2,Human)	-1.2%	1.5%	-0.80	D(Ulindi1,Ulindi2,EC2,Human)	-3.0%	1.1%	-2.73
D(B2,Ulindi1,EC3,Human)	0.4%	1.2%	0.36	D(Ulindi1,Ulindi2,EC3,Human)	-0.7%	0.8%	-0.90
D(B2,Ulindi1,EC4,Human)	1.5%	1.3%	1.19	D(Ulindi1,Ulindi2,EC4,Human)	-1.3%	0.8%	-1.62
D(B2,Ulindi1,EC5,Human)	-1.1%	1.3%	-0.85	D(Ulindi1,Ulindi2,EC5,Human)	-0.7%	0.8%	-0.84
D(B2,Ulindi1,EC6,Human)	1.2%	1.2%	0.96	D(Ulindi1,Ulindi2,EC6,Human)	-1.0%	0.8%	-1.28
D(B2,Ulindi1,EC7,Human)	1.5%	1.4%	1.07	D(Ulindi1,Ulindi2,EC7,Human)	-2.1%	0.9%	-2.26
D(B2,Ulindi1,WC1,Human)	3.6%	1.3%	2.72	D(Ulindi1,Ulindi2,WC1,Human)	-0.7%	0.8%	-0.85
D(B2,Ulindi1,WC2,Human)	2.6%	1.4%	1.82	D(Ulindi1,Ulindi2,WC2,Human)	-0.7%	0.9%	-0.78

Table S10.9: D-statistics for comparison of Bonobo individuals to eastern, central and western chimpanzees. D-value and jackknife standard error estimates are given in per cent.

Relationship between Bonobo Individuals

We used the D-statistics to revisit our earlier result on the closer relationship of B1 to Ulindi than B2. As before, we divide the data up in two sets: B1 and B2 reads that were sequenced on separate lanes and B1 and B2 reads that were sequenced on the same lane. The analysis of the separately sequenced data gives a highly significant signal for closer relationship between B1 and Ulindi (see Table S10.10). The mixed sequencing data encompasses only a fraction of the separately sequenced data. We observe the same directionality of the signal. However, the signal is not significant. The comparison to individual B3 gives a significant signal for the relationship of B1 and Ulindi. However, B3, in turn, shows a signal for closer relationship with Ulindi as compared to B2. These last comparisons involve a long read length individual (B3) and short read length individuals (B1 and B2) and have to be interpreted with caution.

B1 and B2 sequencing	Comparison	D-value %	std.err. %	Z-score
separate sequencing	**D(B1,B2,Ulindi1,Human)**	**-13.69%**	**1.21%**	**-11.34**
separate sequencing	**D(B1,B2,Ulindi2,Human)**	**-12.14%**	**1.18%**	**-10.25**
mixed sequencing	D(B1,B2,Ulindi1,Human)	-15.15%	5.10%	-2.97
mixed sequencing	D(B1,B2,Ulindi2,Human)	-2.15%	4.90%	-0.44
-	**D(B1,B3,Ulindi1,Human)**	**-7.3%**	**1.2%**	**-5.87**
-	**D(B1,B3,Ulindi2,Human)**	**-5.7%**	**1.3%**	**-4.45**
-	**D(B2,B3,Ulindi1,Human)**	**7.6%**	**1.0%**	**7.66**
-	**D(B2,B3,Ulindi2,Human)**	**8.1%**	**1.0%**	**8.21**

Table S10.10: D-statistics for comparison of Illumina-sequenced Bonobo individuals. D-value and jackknife standard error estimates are given in per cent. Columns with a significant enrichment are marked in bold.

Relationship within and between Chimpanzee Sub-Populations

We furthered our analysis by calculating D-values for the comparison of Illumina-sequenced chimpanzee individuals. We first compared eastern and central individuals in their relationship to Clint. We observed a consistent trend for a closer relationship of eastern individuals to Clint as compared to central individuals (see Table S10.11). The signal is significant for 120 of 147 comparisons and positive for 143 of 147 comparisons (binomial with p=0.5: p-value<2.2 x 10^{-16}). The signal is also significant for all three tests involving solely individuals sequenced over identical lanes.

Comparison	D-value %	std. err. %	Z-score	Comparison	D-value %	std. err %.	Z-score
D(CC7,EC1,Clint,Human)	-1.3%	0.8%	-1.55	**D(CC2,EC3,WC1,Human)**	**7.1%**	**0.7%**	**10.18**
D(CC3,EC1,Clint,Human)	-0.2%	0.7%	-0.35	**D(CC2,EC6,WC1,Human)**	**6.5%**	**0.6%**	**10.48**
D(CC7,EC5,Clint,Human)	1.3%	0.6%	2.15	**D(CC6,EC3,WC1,Human)**	**6.9%**	**0.6%**	**10.59**
D(CC7,EC2,Clint,Human)	1.6%	0.7%	2.44	**D(CC6,EC6,WC1,Human)**	**6.5%**	**0.6%**	**10.60**
D(CC3,EC5,Clint,Human)	1.3%	0.5%	2.52	**D(CC2,EC4,WC1,Human)**	**7.3%**	**0.7%**	**10.82**
D(CC3,EC2,Clint,Human)	1.4%	0.5%	2.54	**D(CC5,EC2,WC1,Human)**	**7.8%**	**0.7%**	**11.07**
D(CC2,EC1,Clint,Human)	3.6%	0.9%	4.22	**D(CC1,EC2,WC1,Human)**	**7.9%**	**0.7%**	**11.11**
D(CC6,EC1,Clint,Human)	3.2%	0.7%	4.32	**D(CC6,EC4,WC1,Human)**	**7.7%**	**0.6%**	**12.22**
D(CC7,EC6,Clint,Human)	**2.6%**	**0.6%**	**4.59**	**D(CC5,EC5,WC1,Human)**	**8.3%**	**0.7%**	**12.74**
D(CC7,EC3,Clint,Human)	**2.8%**	**0.6%**	**4.83**	**D(CC4,EC2,WC1,Human)**	**9.4%**	**0.7%**	**12.81**
D(CC3,EC6,Clint,Human)	**2.5%**	**0.5%**	**4.96**	**D(CC1,EC5,WC1,Human)**	**8.3%**	**0.6%**	**13.51**
D(CC3,EC3,Clint,Human)	**2.8%**	**0.5%**	**5.66**	**D(CC5,EC7,WC1,Human)**	**9.7%**	**0.7%**	**14.51**
D(CC5,EC1,Clint,Human)	**4.5%**	**0.7%**	**6.12**	**D(CC5,EC6,WC1,Human)**	**8.5%**	**0.6%**	**14.69**
D(CC7,EC4,Clint,Human)	**3.8%**	**0.6%**	**6.70**	**D(CC5,EC4,WC1,Human)**	**8.9%**	**0.6%**	**15.12**
D(CC3,EC7,Clint,Human)	**3.5%**	**0.5%**	**6.70**	**D(CC5,EC3,WC1,Human)**	**9.0%**	**0.6%**	**15.14**
D(CC3,EC4,Clint,Human)	**3.5%**	**0.5%**	**6.99**	**D(CC4,EC5,WC1,Human)**	**10.1%**	**0.7%**	**15.20**
D(CC7,EC7,Clint,Human)	**4.1%**	**0.6%**	**7.12**	**D(CC1,EC7,WC1,Human)**	**10.3%**	**0.7%**	**15.54**
D(CC1,EC1,Clint,Human)	**5.3%**	**0.7%**	**7.18**	**D(CC1,EC3,WC1,Human)**	**9.4%**	**0.6%**	**15.67**
D(CC2,EC2,Clint,Human)	**5.5%**	**0.7%**	**8.33**	**D(CC1,EC6,WC1,Human)**	**9.1%**	**0.6%**	**15.78**
D(CC6,EC2,Clint,Human)	**5.2%**	**0.6%**	**8.84**	**D(CC1,EC4,WC1,Human)**	**10.1%**	**0.6%**	**16.95**
D(CC6,EC5,Clint,Human)	**5.1%**	**0.5%**	**9.28**	**D(CC4,EC7,WC1,Human)**	**12.0%**	**0.7%**	**17.37**
D(CC4,EC1,Clint,Human)	**7.0%**	**0.8%**	**9.32**	**D(CC4,EC3,WC1,Human)**	**10.9%**	**0.6%**	**17.67**
D(CC2,EC5,Clint,Human)	**5.9%**	**0.6%**	**9.86**	**D(CC4,EC6,WC1,Human)**	**11.2%**	**0.6%**	**17.86**
D(CC6,EC7,Clint,Human)	**5.7%**	**0.6%**	**9.98**	**D(CC4,EC4,WC1,Human)**	**11.8%**	**0.6%**	**18.78**
D(CC6,EC3,Clint,Human)	**5.7%**	**0.5%**	**11.19**	D(CC7,EC1,WC2,Human)	-0.4%	1.2%	-0.30
D(CC6,EC6,Clint,Human)	**5.9%**	**0.5%**	**11.62**	D(CC7,EC2,WC2,Human)	0.4%	0.9%	0.41
D(CC2,EC7,Clint,Human)	**7.5%**	**0.6%**	**12.43**	D(CC3,EC1,WC2,Human)	1.1%	0.9%	1.20
D(CC2,EC4,Clint,Human)	**7.4%**	**0.6%**	**12.80**	D(CC7,EC5,WC2,Human)	1.6%	0.8%	1.91
D(CC2,EC6,Clint,Human)	**7.1%**	**0.5%**	**13.26**	D(CC3,EC5,WC2,Human)	1.6%	0.7%	2.32
D(CC5,EC2,Clint,Human)	**7.6%**	**0.6%**	**13.33**	D(CC2,EC1,WC2,Human)	3.1%	1.1%	2.71

D(CC5,EC5,Clint,Human)	**7.3%**	**0.5%**	**13.45**	D(CC3,EC2,WC2,Human)	2.1%	0.7%	2.77
D(CC1,EC2,Clint,Human)	**7.7%**	**0.6%**	**13.50**	D(CC6,EC1,WC2,Human)	3.3%	1.0%	3.12
D(CC2,EC3,Clint,Human)	**7.8%**	**0.6%**	**13.61**	D(CC7,EC3,WC2,Human)	2.7%	0.8%	3.44
D(CC6,EC4,Clint,Human)	**7.4%**	**0.5%**	**14.44**	D(CC7,EC6,WC2,Human)	2.6%	0.7%	3.59
D(CC1,EC5,Clint,Human)	**7.6%**	**0.5%**	**14.63**	**D(CC3,EC6,WC2,Human)**	**2.9%**	**0.6%**	**4.57**
D(CC5,EC7,Clint,Human)	**8.4%**	**0.6%**	**14.98**	**D(CC7,EC4,WC2,Human)**	**3.6%**	**0.8%**	**4.83**
D(CC5,EC6,Clint,Human)	**8.2%**	**0.5%**	**16.49**	**D(CC2,EC2,WC2,Human)**	**4.5%**	**0.9%**	**4.85**
D(CC4,EC5,Clint,Human)	**9.6%**	**0.6%**	**16.99**	**D(CC3,EC3,WC2,Human)**	**3.2%**	**0.7%**	**4.87**
D(CC4,EC2,Clint,Human)	**10.3%**	**0.6%**	**17.23**	**D(CC7,EC7,WC2,Human)**	**4.1%**	**0.8%**	**5.27**
D(CC1,EC7,Clint,Human)	**9.3%**	**0.5%**	**17.46**	**D(CC3,EC7,WC2,Human)**	**3.9%**	**0.7%**	**5.63**
D(CC5,EC3,Clint,Human)	**8.7%**	**0.5%**	**17.61**	**D(CC3,EC4,WC2,Human)**	**3.8%**	**0.7%**	**5.76**
D(CC1,EC6,Clint,Human)	**8.7%**	**0.5%**	**17.73**	**D(CC6,EC2,WC2,Human)**	**4.8%**	**0.8%**	**5.82**
D(CC5,EC4,Clint,Human)	**9.0%**	**0.5%**	**17.98**	**D(CC2,EC5,WC2,Human)**	**4.8%**	**0.8%**	**6.09**
D(CC1,EC3,Clint,Human)	**9.3%**	**0.5%**	**18.18**	**D(CC5,EC1,WC2,Human)**	**6.7%**	**1.0%**	**6.96**
D(CC1,EC4,Clint,Human)	**9.3%**	**0.5%**	**18.58**	**D(CC1,EC1,WC2,Human)**	**6.7%**	**0.9%**	**7.17**
D(CC4,EC7,Clint,Human)	**11.7%**	**0.6%**	**20.33**	**D(CC6,EC5,WC2,Human)**	**5.2%**	**0.7%**	**7.56**
D(CC4,EC6,Clint,Human)	**10.9%**	**0.5%**	**21.07**	**D(CC6,EC7,WC2,Human)**	**5.8%**	**0.7%**	**7.98**
D(CC4,EC3,Clint,Human)	**11.4%**	**0.5%**	**21.95**	**D(CC2,EC7,WC2,Human)**	**6.7%**	**0.8%**	**8.18**
D(CC4,EC4,Clint,Human)	**11.5%**	**0.5%**	**22.53**	**D(CC4,EC1,WC2,Human)**	**8.3%**	**1.0%**	**8.41**
D(CC7,EC1,WC1,Human)	-1.8%	1.1%	-1.69	**D(CC6,EC3,WC2,Human)**	**5.5%**	**0.6%**	**8.59**
D(CC3,EC1,WC1,Human)	1.0%	0.9%	1.11	**D(CC2,EC3,WC2,Human)**	**6.3%**	**0.7%**	**8.87**
D(CC7,EC2,WC1,Human)	1.3%	0.8%	1.51	**D(CC2,EC6,WC2,Human)**	**6.4%**	**0.7%**	**9.17**
D(CC7,EC5,WC1,Human)	1.2%	0.7%	1.70	**D(CC6,EC6,WC2,Human)**	**6.0%**	**0.7%**	**9.25**
D(CC3,EC2,WC1,Human)	1.3%	0.7%	1.94	**D(CC2,EC4,WC2,Human)**	**7.1%**	**0.7%**	**9.46**
D(CC2,EC1,WC1,Human)	3.2%	1.0%	3.12	**D(CC6,EC4,WC2,Human)**	**7.0%**	**0.7%**	**10.50**
D(CC3,EC5,WC1,Human)	2.2%	0.6%	3.73	**D(CC5,EC2,WC2,Human)**	**8.3%**	**0.8%**	**10.70**
D(CC7,EC3,WC1,Human)	2.8%	0.7%	3.92	**D(CC5,EC5,WC2,Human)**	**7.8%**	**0.7%**	**11.03**
D(CC6,EC1,WC1,Human)	3.6%	0.9%	4.03	**D(CC1,EC2,WC2,Human)**	**8.5%**	**0.8%**	**11.05**
D(CC7,EC6,WC1,Human)	**3.0%**	**0.7%**	**4.64**	**D(CC4,EC2,WC2,Human)**	**9.0%**	**0.8%**	**11.50**
D(CC7,EC7,WC1,Human)	**3.7%**	**0.8%**	**4.71**	**D(CC1,EC5,WC2,Human)**	**8.5%**	**0.7%**	**12.72**
D(CC7,EC4,WC1,Human)	**3.3%**	**0.7%**	**4.76**	**D(CC4,EC5,WC2,Human)**	**9.2%**	**0.7%**	**12.76**
D(CC5,EC1,WC1,Human)	**5.1%**	**0.9%**	**5.78**	**D(CC5,EC3,WC2,Human)**	**8.3%**	**0.6%**	**13.37**
D(CC2,EC2,WC1,Human)	**4.9%**	**0.9%**	**5.81**	**D(CC5,EC6,WC2,Human)**	**8.3%**	**0.6%**	**13.40**
D(CC3,EC6,WC1,Human)	**3.4%**	**0.6%**	**5.97**	**D(CC1,EC6,WC2,Human)**	**8.7%**	**0.6%**	**13.76**
D(CC3,EC3,WC1,Human)	**3.6%**	**0.6%**	**6.06**	**D(CC5,EC7,WC2,Human)**	**10.2%**	**0.7%**	**14.11**
D(CC3,EC4,WC1,Human)	**4.1%**	**0.6%**	**6.64**	**D(CC1,EC7,WC2,Human)**	**10.2%**	**0.7%**	**14.31**
D(CC3,EC7,WC1,Human)	**4.7%**	**0.7%**	**6.95**	**D(CC5,EC4,WC2,Human)**	**9.1%**	**0.6%**	**14.44**
D(CC2,EC5,WC1,Human)	**5.2%**	**0.7%**	**6.97**	**D(CC1,EC3,WC2,Human)**	**9.7%**	**0.6%**	**15.13**
D(CC1,EC1,WC1,Human)	**6.5%**	**0.9%**	**7.26**	**D(CC1,EC4,WC2,Human)**	**9.8%**	**0.6%**	**15.50**
D(CC4,EC1,WC1,Human)	**7.0%**	**1.0%**	**7.27**	**D(CC4,EC7,WC2,Human)**	**11.4%**	**0.7%**	**15.56**
D(CC6,EC2,WC1,Human)	**5.3%**	**0.7%**	**7.32**	**D(CC4,EC3,WC2,Human)**	**10.7%**	**0.7%**	**15.69**
D(CC2,EC7,WC1,Human)	**7.2%**	**0.8%**	**9.36**	**D(CC4,EC6,WC2,Human)**	**11.0%**	**0.7%**	**16.54**
D(CC6,EC7,WC1,Human)	**6.7%**	**0.7%**	**9.49**	**D(CC4,EC4,WC2,Human)**	**11.0%**	**0.7%**	**16.79**
D(CC6,EC5,WC1,Human)	**6.2%**	**0.6%**	**9.61**				

Table S10.11: D-statistics for comparison of eastern- and central chimpanzees to western chimpanzees. D-value and jackknife standard error estimates are given in per cent. Significant results are shown in bold. Comparisons of individuals sequenced on the same lanes are highlighted green.

In a next step we tested how central individuals differ in their relationship to western and eastern individuals. We see that individual CC3 and CC7 show a significantly closer relationship to Clint than other central chimpanzee individuals. Individual CC6 exhibits also significant similarity in some pairwise tests, but does not show this difference consistent over all comparisons (see Table S10.12). The tests with other western individuals shows a similar trend, but less individual comparisons are significant.

Comparison	D-value %	std. err. %	Z-score	Comparison	D-value %	std. err %.	Z-score
D(CC3,CC4,Clint,Human)	**-8.4%**	**0.5%**	**-18.01**	D(CC2,CC3,WC1,Human)	2.3%	0.6%	3.79
D(CC3,CC5,Clint,Human)	**-5.1%**	**0.5%**	**-10.79**	D(CC4,CC5,WC1,Human)	2.5%	0.6%	4.18
D(CC2,CC4,Clint,Human)	**-4.4%**	**0.5%**	**-8.29**	D(CC1,CC2,WC1,Human)	2.8%	0.7%	4.21
D(CC3,CC6,Clint,Human)	**-3.3%**	**0.5%**	**-6.46**	**D(CC1,CC6,WC1,Human)**	**2.8%**	**0.6%**	**4.58**
D(CC1,CC4,Clint,Human)	-1.9%	0.5%	-4.15	**D(CC6,CC7,WC1,Human)**	**3.9%**	**0.7%**	**5.37**
D(CC2,CC5,Clint,Human)	-1.7%	0.5%	-3.21	**D(CC4,CC6,WC1,Human)**	**5.2%**	**0.6%**	**7.97**
D(CC3,CC7,Clint,Human)	-0.2%	0.6%	-0.41	**D(CC5,CC7,WC1,Human)**	**5.8%**	**0.7%**	**8.80**
D(CC1,CC5,Clint,Human)	0.8%	0.5%	1.53	**D(CC1,CC7,WC1,Human)**	**6.4%**	**0.7%**	**9.40**
D(CC2,CC6,Clint,Human)	1.2%	0.6%	1.84	**D(CC1,CC3,WC1,Human)**	**6.0%**	**0.6%**	**10.22**
D(CC1,CC2,Clint,Human)	1.9%	0.5%	3.56	**D(CC4,CC7,WC1,Human)**	**8.1%**	**0.7%**	**11.96**
D(CC5,CC6,Clint,Human)	2.2%	0.5%	4.34	**D(CC3,CC4,WC2,Human)**	**-7.9%**	**0.6%**	**-12.59**
D(CC2,CC7,Clint,Human)	**3.5%**	**0.6%**	**5.47**	**D(CC3,CC5,WC2,Human)**	**-5.2%**	**0.6%**	**-8.75**
D(CC4,CC5,Clint,Human)	**2.7%**	**0.5%**	**5.60**	**D(CC2,CC4,WC2,Human)**	**-4.2%**	**0.7%**	**-5.87**
D(CC6,CC7,Clint,Human)	**3.3%**	**0.6%**	**5.65**	D(CC3,CC6,WC2,Human)	-2.8%	0.7%	-4.28
D(CC1,CC6,Clint,Human)	**3.4%**	**0.5%**	**6.65**	D(CC2,CC5,WC2,Human)	-2.4%	0.7%	-3.32
D(CC2,CC3,Clint,Human)	**3.5%**	**0.5%**	**6.71**	D(CC1,CC4,WC2,Human)	-1.3%	0.6%	-2.17
D(CC5,CC7,Clint,Human)	**5.6%**	**0.5%**	**10.42**	D(CC3,CC7,WC2,Human)	0.2%	0.7%	0.31
D(CC4,CC6,Clint,Human)	**5.5%**	**0.5%**	**10.47**	D(CC2,CC6,WC2,Human)	0.7%	0.8%	0.87
D(CC1,CC7,Clint,Human)	**5.8%**	**0.6%**	**10.55**	D(CC1,CC5,WC2,Human)	1.0%	0.6%	1.66
D(CC1,CC3,Clint,Human)	**6.2%**	**0.5%**	**12.51**	D(CC2,CC7,WC2,Human)	2.0%	0.8%	2.42
D(CC4,CC7,Clint,Human)	**8.1%**	**0.5%**	**15.85**	D(CC4,CC5,WC2,Human)	2.3%	0.6%	3.73
D(CC3,CC4,WC1,Human)	**-8.2%**	**0.6%**	**-14.43**	D(CC2,CC3,WC2,Human)	2.7%	0.7%	3.74
D(CC3,CC5,WC1,Human)	**-4.9%**	**0.6%**	**-8.52**	D(CC1,CC2,WC2,Human)	2.8%	0.7%	4.11
D(CC2,CC4,WC1,Human)	**-5.0%**	**0.7%**	**-7.42**	**D(CC5,CC6,WC2,Human)**	**2.9%**	**0.7%**	**4.44**
D(CC3,CC6,WC1,Human)	**-3.0%**	**0.6%**	**-4.68**	**D(CC6,CC7,WC2,Human)**	**3.6%**	**0.8%**	**4.57**
D(CC2,CC5,WC1,Human)	-2.2%	0.6%	-3.38	**D(CC1,CC6,WC2,Human)**	**4.7%**	**0.7%**	**7.14**
D(CC1,CC4,WC1,Human)	-1.8%	0.6%	-3.03	**D(CC4,CC6,WC2,Human)**	**5.4%**	**0.7%**	**7.88**
D(CC2,CC6,WC1,Human)	-0.1%	0.7%	-0.12	**D(CC5,CC7,WC2,Human)**	**6.0%**	**0.7%**	**8.60**
D(CC3,CC7,WC1,Human)	0.2%	0.7%	0.24	**D(CC1,CC7,WC2,Human)**	**6.9%**	**0.7%**	**9.53**
D(CC1,CC5,WC1,Human)	0.6%	0.6%	0.95	**D(CC1,CC3,WC2,Human)**	**6.2%**	**0.6%**	**9.93**
D(CC2,CC7,WC1,Human)	2.5%	0.8%	3.17	**D(CC4,CC7,WC2,Human)**	**7.7%**	**0.7%**	**10.56**
D(CC5,CC6,WC1,Human)	2.0%	0.6%	3.38				

Table S10.12: D-statistics for comparison of central chimpanzees to Clint. D-value and jackknife standard error estimates are given in per cent. Significant comparisons are shown in bold.

The closer relationship of CC3 and CC7 to Clint among all other central chimpanzees may be caused by a closer genetic relationship of CC3 to eastern chimpanzees. We therefore tested all pairwise combinations of central chimpanzees against all eastern individuals (Table S10.13). We see that CC3, CC7, CC6 and CC2 show consistently more shared derived positions with eastern individuals than CC1, CC4 and CC5. This difference between the two groups of central chimpanzees is significant except for one comparison involving data of different read length.

Comparison	D-value	std. err.	Z-score	Comparison	D-value	std. err.	Z-score
D(CC3,CC4,EC1,Human)	**-7.3%**	**0.8%**	**-9.07**	D(CC6,CC7,EC4,Human)	2.0%	0.6%	3.27
D(CC3,CC5,EC1,Human)	**-6.2%**	**0.7%**	**-8.40**	**D(CC2,CC3,EC4,Human)**	**3.0%**	**0.6%**	**5.30**
D(CC2,CC4,EC1,Human)	**-6.8%**	**1.0%**	**-7.09**	**D(CC1,CC2,EC4,Human)**	**3.3%**	**0.6%**	**5.42**
D(CC2,CC5,EC1,Human)	**-5.7%**	**0.9%**	**-6.38**	**D(CC5,CC6,EC4,Human)**	**4.0%**	**0.5%**	**7.51**
D(CC3,CC6,EC1,Human)	-1.4%	0.8%	-1.67	**D(CC1,CC6,EC4,Human)**	**4.5%**	**0.5%**	**8.81**
D(CC3,CC7,EC1,Human)	-0.9%	0.9%	-1.03	**D(CC5,CC7,EC4,Human)**	**5.8%**	**0.6%**	**9.74**

D(CC1,CC4,EC1,Human)	-0.6%	0.8%	-0.75	**D(CC1,CC7,EC4,Human)**	**6.7%**	**0.6%**	**11.08**
D(CC2,CC6,EC1,Human)	0.1%	0.9%	0.09	**D(CC4,CC6,EC4,Human)**	**6.0%**	**0.5%**	**11.22**
D(CC1,CC5,EC1,Human)	0.1%	0.8%	0.10	**D(CC4,CC7,EC4,Human)**	**7.4%**	**0.6%**	**12.14**
D(CC6,CC7,EC1,Human)	0.9%	1.0%	0.94	**D(CC1,CC3,EC4,Human)**	**6.7%**	**0.5%**	**13.20**
D(CC2,CC7,EC1,Human)	1.0%	1.1%	0.96	**D(CC3,CC4,EC5,Human)**	**-8.0%**	**0.6%**	**-14.21**
D(CC2,CC3,EC1,Human)	1.1%	0.9%	1.23	**D(CC3,CC5,EC5,Human)**	**-6.8%**	**0.5%**	**-12.45**
D(CC4,CC5,EC1,Human)	1.8%	0.8%	2.28	**D(CC2,CC4,EC5,Human)**	**-4.5%**	**0.7%**	**-6.79**
D(CC1,CC2,EC1,Human)	3.5%	0.9%	3.76	**D(CC2,CC5,EC5,Human)**	**-3.6%**	**0.6%**	**-5.74**
D(CC5,CC6,EC1,Human)	**4.8%**	**0.8%**	**5.98**	D(CC3,CC6,EC5,Human)	-2.5%	0.6%	-4.30
D(CC1,CC7,EC1,Human)	**6.2%**	**1.0%**	**6.42**	D(CC1,CC4,EC5,Human)	-0.7%	0.6%	-1.25
D(CC4,CC6,EC1,Human)	**5.8%**	**0.9%**	**6.64**	D(CC3,CC7,EC5,Human)	-0.1%	0.6%	-0.09
D(CC5,CC7,EC1,Human)	**6.2%**	**0.9%**	**6.69**	D(CC1,CC5,EC5,Human)	0.0%	0.6%	0.05
D(CC1,CC6,EC1,Human)	**5.8%**	**0.8%**	**7.10**	D(CC2,CC6,EC5,Human)	0.4%	0.7%	0.65
D(CC4,CC7,EC1,Human)	**7.1%**	**1.0%**	**7.47**	D(CC4,CC5,EC5,Human)	1.2%	0.6%	2.05
D(CC1,CC3,EC1,Human)	**7.1%**	**0.8%**	**9.05**	D(CC6,CC7,EC5,Human)	2.6%	0.7%	3.90
D(CC3,CC4,EC2,Human)	**-8.9%**	**0.6%**	**-14.76**	D(CC2,CC7,EC5,Human)	3.0%	0.7%	4.14
D(CC3,CC5,EC2,Human)	**-7.0%**	**0.6%**	**-12.08**	**D(CC2,CC3,EC5,Human)**	**3.1%**	**0.6%**	**5.24**
D(CC2,CC4,EC2,Human)	**-7.7%**	**0.7%**	**-10.26**	**D(CC1,CC2,EC5,Human)**	**3.9%**	**0.6%**	**6.24**
D(CC2,CC5,EC2,Human)	**-5.0%**	**0.7%**	**-6.91**	**D(CC5,CC6,EC5,Human)**	**4.6%**	**0.6%**	**7.58**
D(CC3,CC6,EC2,Human)	-2.4%	0.6%	-3.88	**D(CC1,CC6,EC5,Human)**	**4.7%**	**0.6%**	**7.81**
D(CC1,CC4,EC2,Human)	-2.1%	0.6%	-3.31	**D(CC5,CC7,EC5,Human)**	**6.1%**	**0.6%**	**9.44**
D(CC2,CC6,EC2,Human)	-1.6%	0.8%	-2.00	**D(CC1,CC7,EC5,Human)**	**6.5%**	**0.7%**	**9.82**
D(CC3,CC7,EC2,Human)	-0.4%	0.7%	-0.61	**D(CC4,CC6,EC5,Human)**	**6.2%**	**0.6%**	**9.91**
D(CC1,CC5,EC2,Human)	0.2%	0.6%	0.33	**D(CC4,CC7,EC5,Human)**	**7.7%**	**0.6%**	**11.85**
D(CC2,CC7,EC2,Human)	1.0%	0.8%	1.19	**D(CC1,CC3,EC5,Human)**	**7.1%**	**0.6%**	**12.78**
D(CC6,CC7,EC2,Human)	1.3%	0.8%	1.69	**D(CC3,CC4,EC6,Human)**	**-8.9%**	**0.5%**	**-17.56**
D(CC2,CC3,EC2,Human)	1.2%	0.7%	1.80	**D(CC3,CC5,EC6,Human)**	**-6.5%**	**0.5%**	**-13.22**
D(CC4,CC5,EC2,Human)	2.5%	0.7%	3.81	**D(CC2,CC4,EC6,Human)**	**-5.0%**	**0.6%**	**-8.69**
D(CC1,CC2,EC2,Human)	**5.1%**	**0.7%**	**6.85**	**D(CC3,CC6,EC6,Human)**	**-3.2%**	**0.5%**	**-6.20**
D(CC1,CC6,EC2,Human)	**5.0%**	**0.7%**	**7.33**	**D(CC2,CC5,EC6,Human)**	**-3.0%**	**0.6%**	**-5.22**
D(CC5,CC6,EC2,Human)	**5.1%**	**0.7%**	**7.59**	**D(CC1,CC4,EC6,Human)**	**-2.2%**	**0.5%**	**-4.44**
D(CC1,CC7,EC2,Human)	**6.0%**	**0.7%**	**8.29**	D(CC3,CC7,EC6,Human)	-1.2%	0.6%	-2.06
D(CC5,CC7,EC2,Human)	**6.6%**	**0.7%**	**8.93**	D(CC1,CC5,EC6,Human)	0.0%	0.5%	0.07
D(CC4,CC6,EC2,Human)	**6.7%**	**0.6%**	**10.34**	D(CC2,CC6,EC6,Human)	0.6%	0.6%	0.90
D(CC1,CC3,EC2,Human)	**7.0%**	**0.6%**	**11.74**	D(CC6,CC7,EC6,Human)	1.8%	0.6%	3.11
D(CC4,CC7,EC2,Human)	**8.5%**	**0.7%**	**12.33**	D(CC2,CC7,EC6,Human)	2.3%	0.6%	3.93
D(CC3,CC4,EC3,Human)	**-7.5%**	**0.5%**	**-15.06**	**D(CC4,CC5,EC6,Human)**	**2.4%**	**0.5%**	**4.72**
D(CC3,CC5,EC3,Human)	**-5.9%**	**0.5%**	**-12.15**	**D(CC1,CC2,EC6,Human)**	**3.1%**	**0.5%**	**5.67**
D(CC2,CC4,EC3,Human)	**-4.2%**	**0.6%**	**-7.19**	**D(CC2,CC3,EC6,Human)**	**3.6%**	**0.5%**	**6.92**
D(CC2,CC5,EC3,Human)	**-2.7%**	**0.6%**	**-4.80**	**D(CC5,CC6,EC6,Human)**	**4.3%**	**0.5%**	**7.90**
D(CC3,CC6,EC3,Human)	**-2.4%**	**0.5%**	**-4.51**	**D(CC1,CC6,EC6,Human)**	**4.7%**	**0.5%**	**8.59**
D(CC1,CC4,EC3,Human)	-1.7%	0.5%	-3.32	**D(CC5,CC7,EC6,Human)**	**5.8%**	**0.6%**	**9.83**
D(CC1,CC5,EC3,Human)	0.1%	0.5%	0.18	**D(CC1,CC7,EC6,Human)**	**6.1%**	**0.6%**	**10.41**
D(CC3,CC7,EC3,Human)	0.3%	0.6%	0.56	**D(CC4,CC6,EC6,Human)**	**6.4%**	**0.5%**	**12.08**
D(CC2,CC6,EC3,Human)	0.7%	0.7%	1.04	**D(CC4,CC7,EC6,Human)**	**8.1%**	**0.6%**	**13.58**
D(CC6,CC7,EC3,Human)	1.7%	0.6%	3.01	**D(CC1,CC3,EC6,Human)**	**6.9%**	**0.5%**	**14.23**
D(CC4,CC5,EC3,Human)	1.9%	0.5%	3.60	**D(CC3,CC4,EC7,Human)**	**-9.0%**	**0.6%**	**-15.58**
D(CC2,CC3,EC3,Human)	2.1%	0.6%	3.62	**D(CC3,CC5,EC7,Human)**	**-6.3%**	**0.5%**	**-11.60**
D(CC2,CC7,EC3,Human)	**3.0%**	**0.7%**	**4.48**	**D(CC2,CC4,EC7,Human)**	**-5.0%**	**0.7%**	**-7.40**
D(CC1,CC2,EC3,Human)	**3.2%**	**0.6%**	**5.40**	**D(CC3,CC6,EC7,Human)**	**-2.9%**	**0.6%**	**-4.76**
D(CC5,CC6,EC3,Human)	**3.9%**	**0.5%**	**7.50**	**D(CC2,CC5,EC7,Human)**	**-3.0%**	**0.7%**	**-4.48**
D(CC1,CC6,EC3,Human)	**4.5%**	**0.6%**	**8.22**	D(CC1,CC4,EC7,Human)	-1.4%	0.6%	-2.42
D(CC1,CC7,EC3,Human)	**5.7%**	**0.6%**	**9.66**	D(CC3,CC7,EC7,Human)	-0.3%	0.6%	-0.48
D(CC5,CC7,EC3,Human)	**6.2%**	**0.6%**	**10.56**	D(CC2,CC6,EC7,Human)	0.8%	0.7%	1.10
D(CC4,CC6,EC3,Human)	**5.6%**	**0.5%**	**10.57**	D(CC1,CC5,EC7,Human)	0.8%	0.6%	1.31
D(CC1,CC3,EC3,Human)	**6.3%**	**0.5%**	**12.14**	D(CC6,CC7,EC7,Human)	2.0%	0.7%	2.81

D(CC4,CC7,EC3,Human)	**7.4%**	**0.6%**	**12.67**	D(CC2,CC7,EC7,Human)	2.4%	0.7%	3.22
D(CC3,CC4,EC4,Human)	**-7.9%**	**0.5%**	**-15.25**	D(CC4,CC5,EC7,Human)	2.4%	0.6%	4.12
D(CC3,CC5,EC4,Human)	**-5.9%**	**0.5%**	**-11.76**	**D(CC2,CC3,EC7,Human)**	**3.3%**	**0.7%**	**4.84**
D(CC2,CC4,EC4,Human)	**-4.6%**	**0.6%**	**-8.22**	**D(CC1,CC2,EC7,Human)**	**3.4%**	**0.6%**	**5.26**
D(CC2,CC5,EC4,Human)	**-3.3%**	**0.6%**	**-5.74**	**D(CC1,CC6,EC7,Human)**	**4.2%**	**0.6%**	**6.66**
D(CC3,CC6,EC4,Human)	-2.2%	0.5%	-4.15	**D(CC5,CC6,EC7,Human)**	**4.2%**	**0.6%**	**7.11**
D(CC1,CC4,EC4,Human)	-1.5%	0.5%	-2.82	**D(CC5,CC7,EC7,Human)**	**5.8%**	**0.7%**	**8.65**
D(CC3,CC7,EC4,Human)	-0.3%	0.6%	-0.42	**D(CC1,CC7,EC7,Human)**	**6.3%**	**0.7%**	**9.08**
D(CC1,CC5,EC4,Human)	0.5%	0.5%	0.88	**D(CC4,CC6,EC7,Human)**	**6.7%**	**0.6%**	**10.99**
D(CC2,CC6,EC4,Human)	1.1%	0.6%	1.78	**D(CC4,CC7,EC7,Human)**	**8.6%**	**0.7%**	**12.23**
D(CC4,CC5,EC4,Human)	1.6%	0.5%	3.12	**D(CC1,CC3,EC7,Human)**	**6.9%**	**0.6%**	**12.51**
D(CC2,CC7,EC4,Human)	2.1%	0.7%	3.22				

Table S10.13: D-statistics for comparison of central chimpanzees to eastern chimpanzees. D-value and jackknife standard error estimates are given in per cent. Significant comparisons are shown in bold. Most of the significant results separate the chimpanzee group of CC2, CC3, CC6, CC7 from CC1, CC4, CC5. The only comparison between these groups without significant signal is shown with gray background.

The pairwise comparison of all eastern individuals to central chimpanzee does not show a significant difference between eastern individuals (Table S10.14).

Comparison	D-value	std. err.	Z-score	Comparison	D-value	std. err.	Z-score
D(EC3,EC5,CC1,Human)	-1.8%	0.6%	-3.05	D(EC1,EC6,CC4,Human)	1.0%	0.9%	1.15
D(EC4,EC5,CC1,Human)	-0.7%	0.6%	-1.06	D(EC3,EC4,CC4,Human)	0.7%	0.6%	1.20
D(EC3,EC6,CC1,Human)	-0.5%	0.5%	-0.89	D(EC1,EC5,CC4,Human)	1.3%	1.0%	1.38
D(EC6,EC7,CC1,Human)	-0.3%	0.6%	-0.49	D(EC2,EC6,CC4,Human)	1.2%	0.7%	1.81
D(EC3,EC4,CC1,Human)	-0.2%	0.5%	-0.41	D(EC2,EC7,CC4,Human)	2.1%	0.8%	2.55
D(EC3,EC7,CC1,Human)	-0.2%	0.6%	-0.26	D(EC1,EC7,CC4,Human)	2.7%	1.0%	2.66
D(EC4,EC7,CC1,Human)	-0.1%	0.6%	-0.09	D(EC1,EC4,CC4,Human)	2.7%	0.9%	3.04
D(EC4,EC6,CC1,Human)	0.1%	0.5%	0.19	D(EC1,EC3,CC4,Human)	2.9%	0.9%	3.15
D(EC1,EC2,CC1,Human)	0.3%	1.1%	0.24	D(EC2,EC4,CC4,Human)	2.2%	0.7%	3.18
D(EC5,EC7,CC1,Human)	0.7%	0.7%	0.95	D(EC2,EC3,CC4,Human)	2.8%	0.7%	3.97
D(EC1,EC5,CC1,Human)	1.0%	1.0%	1.00	D(EC3,EC5,CC5,Human)	-1.9%	0.6%	-3.03
D(EC2,EC5,CC1,Human)	0.9%	0.7%	1.30	D(EC4,EC5,CC5,Human)	-1.3%	0.6%	-2.09
D(EC2,EC7,CC1,Human)	1.7%	0.8%	2.29	D(EC2,EC5,CC5,Human)	-0.4%	0.7%	-0.48
D(EC2,EC6,CC1,Human)	1.7%	0.7%	2.49	D(EC4,EC6,CC5,Human)	-0.2%	0.5%	-0.44
D(EC1,EC6,CC1,Human)	2.3%	0.9%	2.56	D(EC6,EC7,CC5,Human)	-0.1%	0.7%	-0.09
D(EC2,EC3,CC1,Human)	1.8%	0.7%	2.58	D(EC4,EC7,CC5,Human)	0.2%	0.6%	0.29
D(EC5,EC6,CC1,Human)	1.6%	0.6%	2.70	D(EC3,EC4,CC5,Human)	0.2%	0.6%	0.32
D(EC1,EC3,CC1,Human)	2.7%	0.9%	2.97	D(EC3,EC6,CC5,Human)	0.4%	0.6%	0.65
D(EC1,EC7,CC1,Human)	3.2%	1.0%	3.18	D(EC3,EC7,CC5,Human)	0.5%	0.6%	0.80
D(EC2,EC4,CC1,Human)	2.4%	0.7%	3.38	D(EC1,EC2,CC5,Human)	1.4%	1.1%	1.27
D(EC1,EC4,CC1,Human)	3.1%	0.9%	3.44	D(EC1,EC5,CC5,Human)	1.7%	1.0%	1.68
D(EC3,EC5,CC2,Human)	-1.5%	0.7%	-2.20	D(EC2,EC3,CC5,Human)	1.2%	0.7%	1.79
D(EC2,EC5,CC2,Human)	-1.7%	0.8%	-1.99	D(EC2,EC7,CC5,Human)	1.7%	0.8%	2.21
D(EC2,EC6,CC2,Human)	-1.1%	0.8%	-1.44	D(EC2,EC6,CC5,Human)	1.6%	0.7%	2.36
D(EC2,EC3,CC2,Human)	-1.0%	0.8%	-1.31	D(EC2,EC4,CC5,Human)	1.7%	0.7%	2.42
D(EC3,EC6,CC2,Human)	-0.6%	0.6%	-1.03	D(EC5,EC7,CC5,Human)	1.8%	0.7%	2.53
D(EC2,EC4,CC2,Human)	-0.7%	0.8%	-0.92	D(EC1,EC6,CC5,Human)	2.3%	0.8%	2.75
D(EC4,EC5,CC2,Human)	-0.6%	0.7%	-0.88	D(EC1,EC4,CC5,Human)	2.4%	0.8%	2.81
D(EC4,EC6,CC2,Human)	-0.6%	0.6%	-0.87	D(EC5,EC6,CC5,Human)	1.9%	0.6%	2.99
D(EC3,EC4,CC2,Human)	0.0%	0.6%	-0.07	D(EC1,EC7,CC5,Human)	3.1%	1.0%	3.00
D(EC4,EC7,CC2,Human)	0.0%	0.7%	-0.06	D(EC1,EC3,CC5,Human)	2.8%	0.9%	3.07
D(EC1,EC4,CC2,Human)	0.3%	1.0%	0.26	D(EC3,EC5,CC6,Human)	-1.0%	0.6%	-1.60

D(EC1,EC5,CC2,Human)	0.3%	1.1%	0.31	D(EC4,EC5,CC6,Human)	-1.0%	0.6%	-1.56
D(EC2,EC7,CC2,Human)	0.5%	0.9%	0.60	D(EC4,EC7,CC6,Human)	-0.9%	0.7%	-1.38
D(EC1,EC7,CC2,Human)	0.9%	1.2%	0.75	D(EC4,EC6,CC6,Human)	-0.5%	0.6%	-0.87
D(EC3,EC7,CC2,Human)	0.6%	0.7%	0.81	D(EC6,EC7,CC6,Human)	-0.5%	0.7%	-0.79
D(EC1,EC6,CC2,Human)	1.1%	1.0%	1.06	D(EC3,EC6,CC6,Human)	-0.3%	0.5%	-0.55
D(EC6,EC7,CC2,Human)	0.8%	0.7%	1.11	D(EC1,EC5,CC6,Human)	-0.4%	1.1%	-0.36
D(EC1,EC3,CC2,Human)	1.2%	1.0%	1.16	D(EC1,EC3,CC6,Human)	-0.2%	0.9%	-0.23
D(EC1,EC2,CC2,Human)	2.5%	1.3%	1.98	D(EC1,EC7,CC6,Human)	0.1%	1.1%	0.08
D(EC5,EC6,CC2,Human)	1.7%	0.7%	2.55	D(EC2,EC3,CC6,Human)	0.4%	0.7%	0.51
D(EC5,EC7,CC2,Human)	2.7%	0.8%	3.52	D(EC2,EC7,CC6,Human)	0.5%	0.8%	0.58
D(EC3,EC5,CC3,Human)	-1.2%	0.6%	-1.99	D(EC5,EC6,CC6,Human)	0.4%	0.6%	0.69
D(EC4,EC5,CC3,Human)	-0.9%	0.6%	-1.65	D(EC2,EC5,CC6,Human)	0.6%	0.8%	0.69
D(EC6,EC7,CC3,Human)	-0.1%	0.6%	-0.09	D(EC2,EC6,CC6,Human)	0.6%	0.7%	0.83
D(EC4,EC7,CC3,Human)	0.1%	0.6%	0.20	D(EC3,EC7,CC6,Human)	0.6%	0.7%	0.94
D(EC4,EC6,CC3,Human)	0.2%	0.5%	0.34	D(EC1,EC6,CC6,Human)	0.9%	0.9%	0.95
D(EC3,EC4,CC3,Human)	0.3%	0.5%	0.60	D(EC1,EC2,CC6,Human)	1.3%	1.1%	1.15
D(EC2,EC3,CC3,Human)	0.7%	0.6%	1.16	D(EC3,EC4,CC6,Human)	0.8%	0.5%	1.54
D(EC1,EC5,CC3,Human)	1.3%	1.0%	1.31	D(EC1,EC4,CC6,Human)	1.6%	0.9%	1.82
D(EC3,EC6,CC3,Human)	0.7%	0.5%	1.40	D(EC2,EC4,CC6,Human)	1.5%	0.7%	2.09
D(EC2,EC5,CC3,Human)	1.1%	0.7%	1.61	D(EC5,EC7,CC6,Human)	1.7%	0.8%	2.20
D(EC2,EC7,CC3,Human)	1.3%	0.7%	1.72	D(EC4,EC5,CC7,Human)	-0.9%	0.7%	-1.33
D(EC2,EC6,CC3,Human)	1.3%	0.6%	2.19	D(EC3,EC6,CC7,Human)	-0.7%	0.6%	-1.11
D(EC1,EC2,CC3,Human)	2.4%	1.0%	2.33	D(EC3,EC5,CC7,Human)	-0.6%	0.7%	-0.94
D(EC2,EC4,CC3,Human)	1.6%	0.6%	2.47	D(EC4,EC6,CC7,Human)	0.1%	0.6%	0.22
D(EC3,EC7,CC3,Human)	1.5%	0.6%	2.51	D(EC4,EC7,CC7,Human)	0.2%	0.7%	0.26
D(EC5,EC6,CC3,Human)	1.7%	0.6%	3.01	D(EC5,EC6,CC7,Human)	0.4%	0.7%	0.57
D(EC5,EC7,CC3,Human)	2.1%	0.7%	3.14	D(EC6,EC7,CC7,Human)	0.4%	0.7%	0.63
D(EC1,EC3,CC3,Human)	2.7%	0.9%	3.14	D(EC3,EC4,CC7,Human)	0.4%	0.6%	0.72
D(EC1,EC7,CC3,Human)	3.3%	0.9%	3.48	D(EC5,EC7,CC7,Human)	0.9%	0.8%	1.13
D(EC1,EC6,CC3,Human)	3.3%	0.8%	3.91	D(EC2,EC5,CC7,Human)	1.0%	0.8%	1.14
D(EC1,EC4,CC3,Human)	3.5%	0.8%	4.28	D(EC3,EC7,CC7,Human)	1.2%	0.7%	1.71
D(EC4,EC5,CC4,Human)	-1.8%	0.7%	-2.75	D(EC1,EC2,CC7,Human)	2.3%	1.3%	1.75
D(EC4,EC6,CC4,Human)	-1.1%	0.5%	-2.04	D(EC2,EC6,CC7,Human)	1.3%	0.7%	1.77
D(EC3,EC5,CC4,Human)	-1.2%	0.6%	-1.96	D(EC2,EC3,CC7,Human)	1.5%	0.8%	1.78
D(EC3,EC6,CC4,Human)	-1.1%	0.6%	-1.81	D(EC1,EC6,CC7,Human)	2.1%	1.0%	2.15
D(EC4,EC7,CC4,Human)	-1.0%	0.6%	-1.64	D(EC1,EC7,CC7,Human)	2.6%	1.2%	2.17
D(EC6,EC7,CC4,Human)	-0.2%	0.7%	-0.33	D(EC2,EC4,CC7,Human)	1.9%	0.8%	2.45
D(EC5,EC6,CC4,Human)	0.3%	0.6%	0.46	D(EC1,EC5,CC7,Human)	2.7%	1.1%	2.53
D(EC3,EC7,CC4,Human)	0.4%	0.6%	0.63	D(EC1,EC3,CC7,Human)	2.9%	1.0%	2.88
D(EC1,EC2,CC4,Human)	0.9%	1.1%	0.81	D(EC2,EC7,CC7,Human)	2.7%	0.9%	3.11
D(EC5,EC7,CC4,Human)	0.8%	0.7%	1.08	D(EC1,EC4,CC7,Human)	3.3%	1.0%	3.19
D(EC2,EC5,CC4,Human)	0.9%	0.8%	1.12				

Table S10.14: D-statistics for comparison of eastern chimpanzees to central chimpanzees. D-value and jackknife standard error estimates are given in per cent.

We also compared the two Illumina-sequenced Western chimpanzees against all other individuals (Table S10.15). We find no significant differences between the two individuals.

Comparison	D-value %	std. err. %	Z-score
D(WC1,WC2,B1,Human)	0.4%	2.4%	0.16
D(WC1,WC2,B2,Human)	0.0%	1.9%	0.02
D(WC1,WC2,B3,Human)	1.3%	1.6%	0.83
D(WC1,WC2,CC1,Human)	2.0%	1.1%	1.84
D(WC1,WC2,CC2,Human)	-1.7%	1.3%	-1.34
D(WC1,WC2,CC3,Human)	-0.4%	1.1%	-0.40
D(WC1,WC2,CC4,Human)	-0.1%	1.2%	-0.06
D(WC1,WC2,CC5,Human)	-0.5%	1.2%	-0.43
D(WC1,WC2,CC6,Human)	-0.5%	1.2%	-0.43
D(WC1,WC2,CC7,Human)	-1.9%	1.3%	-1.43
D(WC1,WC2,Clint,Human)	0.0%	0.7%	0.05
D(WC1,WC2,EC1,Human)	-2.5%	1.9%	-1.34
D(WC1,WC2,EC2,Human)	-2.4%	1.4%	-1.69
D(WC1,WC2,EC3,Human)	-1.2%	1.1%	-1.05
D(WC1,WC2,EC4,Human)	-2.1%	1.1%	-1.87
D(WC1,WC2,EC5,Human)	-2.1%	1.2%	-1.81
D(WC1,WC2,EC6,Human)	-2.2%	1.1%	-1.94
D(WC1,WC2,EC7,Human)	-0.8%	1.2%	-0.62
D(WC1,WC2,Ulindi1,Human)	0.6%	1.2%	0.46
D(WC1,WC2,Ulindi2,Human)	0.1%	1.2%	0.11

Table S10.15: D-statistics for comparison of two western chimpanzee. D-value and jackknife standard error estimates are given in per cent. Both western chimpanzees were sequenced on identical lanes.

Principal Component Analysis of Chimpanzee Populations

We additionally used a PCA analysis of Clint and central and eastern individuals to gain insight in the population structure of chimpanzee. We restricted the analysis to autosomal sequence and individuals with at least 1x coverage. We excluded low coverage bases and sites repeatmasked in the human genome. Figure S10.8 shows the first and second component of the PCA, explaining together ca. 20% of the variation. The first component separates eastern and central individuals from Clint (see Figure 10.8). The second component shows a gradient through the eastern and chimpanzee individuals. Interestingly, the central individuals recapitulate the previous found grouping of CC1, CC4, CC5 versus CC2, CC3, CC6, with the latter group being placed closer to the eastern individuals.

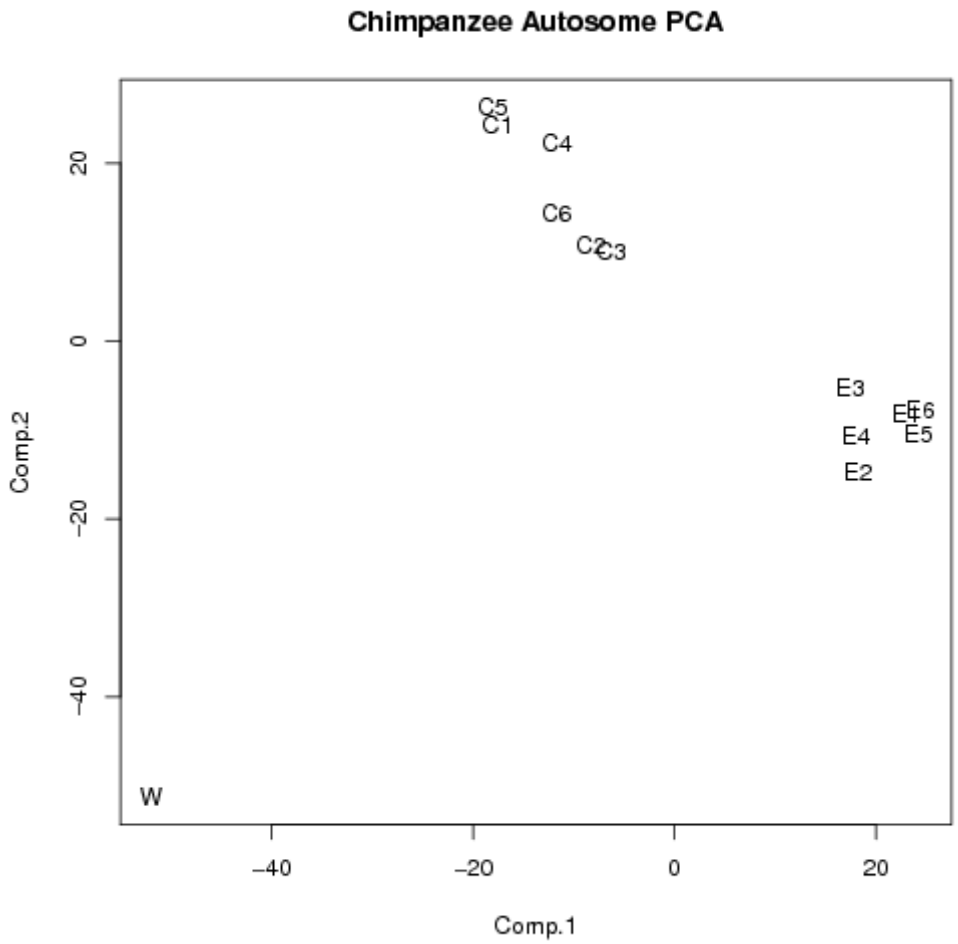


Figure S10.8: Plot of first two components for the principal component analysis of chimpanzee individuals. The plot shows central chimpanzee individuals 1 to 6 (C1,...,C6), eastern individuals 1 to 6 (E1,…,E6) and Clint (W).

Supplementary Information 11

Incomplete Lineage Sorting Regions and Balancing Selection

Kay Prüfer* and Aida Andrés*

* To whom correspondence should be addressed (pruefer@eva.mpg.de, aida_andres@eva.mpg.de)

Among the results from the incomplete lineage sorting (ILS) assignment (see SI 8) are several regions that show a high fraction of ILS as compared to the rest of the genome. Some of these regions may be enriched in incomplete lineage sorting due to long-standing balancing selection. Here, we use the results of the incomplete lineage sorting (ILS) analysis to identify regions that contain candidate targets of balancing selection. We additionally use the resequencing data of several chimpanzee and bonobo individuals and the data from the 1000 Genomes Project to test for further signatures of balancing selection: an excess of diversity within Chimpanzee and Bonobo individuals and an unusually high density of SNPs shared between bonobo and chimpanzee, bonobo and human, and chimpanzee and human. With this data, we aim to identify loci that have undergone the continuous influence of strong balancing selection since the common ancestor of humans, chimpanzees and bonobos, to present-day populations of two or more of these species. Among the regions with the highest ILS and supporting evidence is the major histocompatibility cluster (MHC) on chromosome 6, a region previously known to contain genes evolving under balancing selection [96, 97]. After careful filtering by possible technical artifacts, only MHC loci remained unusual. We conclude that high ILS regions are likely enriched for cases of balancing selection; they contain, however, also a high fraction of technical artifacts that must be carefully considered before biological conclusions can be drawn from their study.

Incomplete Lineage Sorting as a Measure for Balancing Selection

Long-standing balancing selection results in local deep genealogies because of the long-term maintenance of functionally different, selected lineages. This translates into a deep coalescent time (that is, the time where all sampled individuals share a most recent common ancestor) for the alleles evolving under balancing selection. When balancing selection is old enough to predate species-splits, the gene tree for the region defined by the sampled alleles in the species may differ from the species tree. In humans, chimpanzees and bonobos, when random individuals are drawn from these three species, some positions may show the closest relationship to be between bonobo and human or between chimpanzee and human. If the individual is heterozygous, the assembly may choose bases from both alleles to form the consensus and the grouping may differ from informative position to informative position. The incomplete lineage sorting analysis (see SI 8) scans the human, chimpanzee and bonobo genomes for regions that differ from the species tree in this manner. Here we use the posterior decoding of this analysis to identify regions that show an excess of gene

trees that differ from the established species phylogeny in regions of 50kb along the human autosomes.

We considered two measures on the ILS assignment in 50kb windows to identify candidate regions:

1) %CH+BH: measures the number of bases assigned to the states CH or BH (see SI 8 for the full description of these states) over all assigned bases, and represents windows where chimpanzees and bonobo are often closer to human than to their sister species;

2) %CH+BH+BC2: additionally includes the state BC2, thus calculating the percentage of all deep genealogies in a block;

We test the measures by running our analysis first on chromosome 6. The short arm of chromosome 6 contains the cluster of Major Histocompatibility Complex (MHC) genes, a region well-known to evolve under long-standing balancing selection. The top regions, with the highest value for the two ILS measures, contain MHC genes (data only shown for %CH+BH: see Table S11.1). This suggests that, as expected, the ILS measures identify, among others, regions with the type of balancing selection that affects MHC evolution. Among the top candidates also appears a region containing *PRIM2* (primase, DNA, polypeptide 2), a gene that was recently reported to carry a high number of shared SNPs between human and chimpanzee, suggesting that it may evolve under long-term balancing selection [98]. Of all three measures, we observe %CH+BH to be the most sensitive to detect targets of long-standing balancing selection; over a cutoff of 0.78 the measure yields solely regions overlapping the MHC and the region overlapping the gene PRIM2 (see Table S11.1). Therefore, in the following we restrict our analysis to the %CH+BH measure. For this, we define a minimum number of ILS-assigned bases of 5 kb (to exclude false positives due to a low number of assigned bases) and choose a cut-off expected to identify only regions with signatures as strong, or stronger, than the MHC loci (%CH+BH > 0.78).

When we apply these cutoffs for all regions on all autosomes, we identify 16 regions based on their %BH+CH. Figure S11.1 shows a scatter plot for the %CH+BH measure for chromosome 6 and the autosomes.

Start	End	BC1	BC2	BH	CH	%BH+CH	Genes
32600000	32650000	0	0	629	3578	100%	HLA-DRB5, HLA-DRB6
32550000	32600000	0	0	0	1406	100%	HLA-DRB5
32700000	32750000	0	89	6676	0	99%	HLA-DQA1, HLA-DQB1
57350000	57400000	0	250	361	5525	96%	PRIM2
31350000	31400000	457	2533	6342	4766	79%	HLA-B
78650000	78700000	98	2944	6675	3469	77%	-
31100000	31150000	494	0	129	1326	75%	PBMUCL1, HCG22
132700000	132750000	5213	673	11920	4619	74%	MOXD1
125500000	125550000	5461	234	2973	11502	72%	TPD52L1
141100000	141150000	5625	0	8006	6266	72%	-

Table S11.1: Top 10 regions with the largest values for the %BH+CH measure on chromosome 6. All coordinates are given relative to human reference genome (hg18).

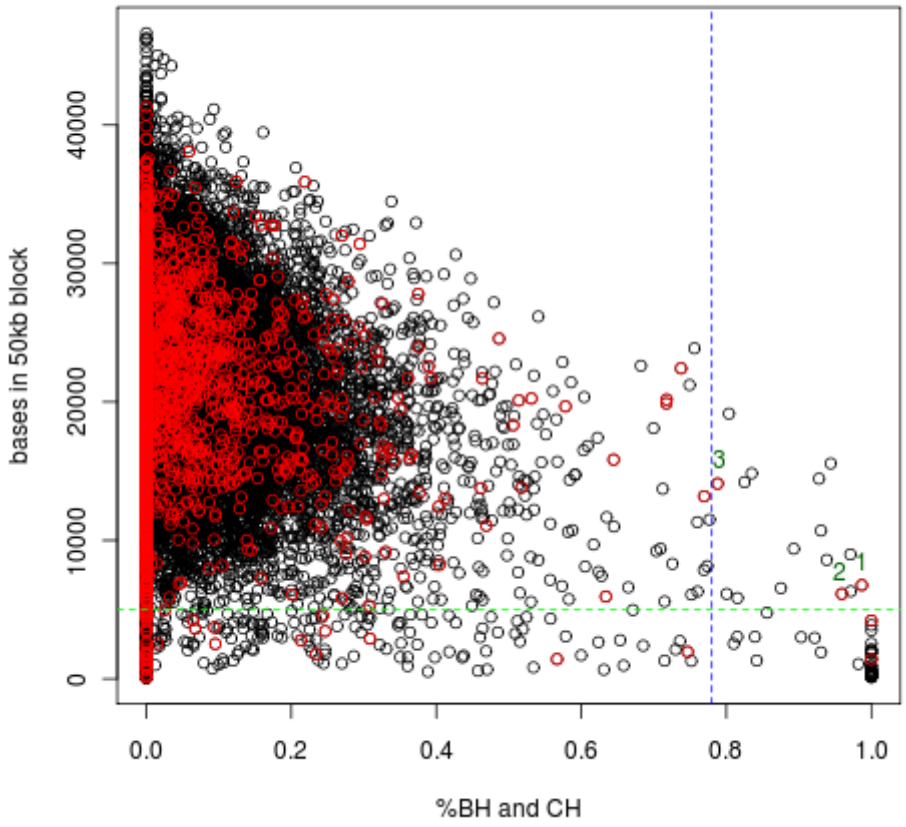


Figure S11.1: ILS-assigned bases versus %BH+CH for chromosome 6 (red) and all autosomes (black). The blue dashed line corresponds to the chosen cutoff of 0.78, the green dashed line shows the cutoff on the minimum number of ILS-assigned bases. Three regions on chromosome 6 (labeled 1, 2, 3) fall within these cutoffs and contain genes: 1) HLA-DQA1, HLA-DQB1; 2) PRIM2; 3) HLA-B.

Balancing Selection Candidates Exhibit High Diversity in Chimpanzee

Under balancing selection, the ILS measure used here mostly reflects the likelihood of selection maintaining more than one allele in the population during the time of speciation between humans and the common ancestor of chimpanzees and bonobos. But if balancing selection remains active until present times (or until recent evolutionary times), it may also affect the patterns of polymorphism in present-day populations. It will, for example, result in an increase of diversity in the genomic regions linked to the selected site.

In order to further restrict our candidate regions to those with a higher likelihood of being true targets of selection, we investigate whether the candidate regions contain high diversity within chimpanzees and bonobos. The test uses the Illumina data for all chimpanzee and bonobo individuals and the reads from Ulindi and Clint, the bonobo and chimpanzee individuals sequenced for the reference genomes. Processing is carried out as described in SI 5, with the exception that individual SNP calling is not applied to identify SNPs within Ulindi's genome but a random read is sampled as for all other individuals. We additionally filter all known duplications in Ulindi and Clint according to the WSSD analysis results of aligning reads to hg17 (SI 4).

Using this data we calculate the average number of differences per site in all pairwise comparisons

between individuals, divided by the number of comparisons in the 50kb regions along the human genome. We observe a significant enrichment for high diversity in chimpanzee in candidate regions identified by the ILS measures (see Table S11.2).

Note that although particular characteristics of the region could in principle influence both ILS and extant diversity, the measure investigates, in neutral regions, non-overlapping periods of time and is expected to be independent. However, technical artifacts, such as undetected overcollapsed duplications in all or some of the assemblies, can lead to signals for all measures.

Measure	average for %CH+BH>0.78	average for %CH+BH<=0.78	p-value (wilcoxon rank test, one sided)
diversity Bonobo	0.001452	0.000835	0.05567
diversity Chimpanzee	0.004063	0.002093	0.009046
shared SNPs bonobo-chimpanzee	0.001238	0.000058	0.0953
shared SNPs bonobo-human	0.033220	0.002523	0.005846
shared SNPs chimpanzee-human	0.042120	0.008506	0.002816
recombination rate (SRR)	0.5122	1.0090	0.0001926
coverage Bonobo	8.208*	8.850	0.00631
coverage Chimpanzee	1.442*	1.503	0.0003326

Table S11.2: Comparison of diversity, shared SNPs, recombination rate and coverage between ILS balancing selection candidates (ILS outliers) and all other regions. Given is the average for each measure. P-values are calculated by a one-sided Wilcoxon rank test on the measures per block between ILS-outliers and all others. Blue values (for diversity and shared SNPs) indicate a test for direction outliers > others; red values (for coverage and recombination) give the p-value for the one-sided test for outliers < others. *Values heavily affected by outlier windows.

Measure	Expected	Observed	p-value (χ^2)
diversity Bonobo	0.16	3	0.1101
diversity Chimpanzee	0.16	12	0.0006854
shared SNPs bonobo-chimpanzee	0.16	5	0.03311
shared SNPs bonobo-human	0.16	5	0.03311
shared SNPs chimpanzee-human	0.16	5	0.03311
recombination rate (SRR)	0.16	0	0.6892
coverage Bonobo	0.16	1	0.4354
coverage Chimpanzee	0.16	1	0.4354

Table S11.3: Observed number of top 1% regions according to various measures among the ILS candidate regions. Expected is calculated as 0.01 x ILS-candidates for all candidates with values for the tested measure. The p-value column gives the results of a χ^2-test on expected and observed values.

Balancing Selection Candidates are Enriched for Shared SNPs

Under neutrality, and in the absence of recurrent mutation, we do not expect to observe shared polymorphisms between humans and the Pan species. When speciation time and coalescence time of neutral regions can overlap (as between chimpanzee and bonobo), then neutral shared SNPs may exist. However, when balancing selection has maintained a given polymorphism since the time from the common ancestor of two species until present times (or evolutionarily recent times) in the two species, the selected variant is expected to be still polymorphic in the two species. The action of balancing selection is thus expected to bias

the density of shared SNPs upwards in all pairs of comparisons among the species considered here.

We use all bonobo and chimpanzee individuals to identify positions where bonobos and chimpanzees share a SNP with identical alleles. We divide the number of identified shared SNP positions by the number of sites where at least two bonobo and two chimpanzee individuals have coverage. We also compare SNPs within bonobo and within chimpanzee to SNPs in human, as recently published by the 1000 Genome Project [99]. We use the SNPs reported for the low-coverage whole-genome sequenced humans of all populations and normalize the number of shared SNPs by the sites with coverage for at least two individuals in bonobo and in chimpanzee. All SNP positions at CpG sites (defined to be CpG in human, orangutan or rhesus macaque genome in the HCBOR alignment (see SI 2)) are excluded to minimize the number of recurrent mutations contributing to the shared SNPs.

When we test for a shift in the distribution of the rate of shared SNPs between ILS balancing selection candidates and all other regions, we observe a (significant) higher density of shared SNPs in ILS-based candidate regions for all comparisons with human (see Table S11.2). The overlap of the candidates with the top 1% of regions with the highest shared SNPs measure is significantly higher than expected for all three measures (see Table S11.3), showing that ILS-based candidate regions are enriched for shared SNPs.

Correlation with Coverage

An overcollapse of duplicated regions could cause false high diversity and a false high fraction of shared polymorphism between species. We excluded duplicated regions according to the WSSD analysis of the bonobo and chimpanzee genomes (see SI 4). However, some short duplicated regions may remain undetected in this analysis. In order to test whether duplications may be the driving cause for the high diversity and high fraction of shared SNPs in the ILS-based candidate regions, we count the number of reads covering each position with at least one read coverage in the sequence data of chimpanzees (including Clint) and bonobo (including Ulindi). While some candidate regions are outliers according to this measure, the general trend shows depletion in coverage for the candidate regions (see Table S11.2). We find no significant overlap between the top 1% regions with the highest coverage and the candidate regions (Table S11.3). However, overcollapsed regions that are smaller than the 50 kb window used in this analysis may still contribute false positives to the results. Unfortunately, only the detailed inspection of individual outliers can control for the confounding effects of such small structural variants.

Interestingly, the previously discussed candidate of balancing selection *PRIM2* [98] yields very high values for the average coverage in bonobo and chimpanzee, indicating that the excess of shared polymorphism observed between chimpanzee and human might be an artifact due to this region likely being duplicated.

Correlation with Recombination

Recombination rate is a well-known covariate of mutation rate. Also, background selection is modulated by recombination rate, leading to a lower local Ne with lower recombination rate. Both of these factors could in

principle influence our results, so we also tested whether ILS outliers are enriched for regions with high recombination rate. Since no recombination map of chimpanzee or bonobo is currently available, we use a recombination map of the human genome as a proxy for chimpanzee and bonobo recombination rates. Hotspot locations are known to differ between human and chimpanzee [70, 71] but recent results indicate that the large-scale recombination rates are conserved between the species (Peter Donnelly, personal communication and [72]). Therefore, we average the human recombination rate over 1 Mb windows around the 50 kb block under consideration. We find a significant trend towards lower recombination rate for our candidate regions (Table S11.2). Accordingly, the overlap with the top 1% regions with the highest recombination rate is not significant (Table S11.3).

A general trend towards low recombination rate is not expected to mimic the signatures of balancing selection, since in functional regions a high amount of ILS and high diversity would tend to be observed in highly recombining regions (due to a weaker effect of background selection) [100, 101]. Instead, the observed trend might come from a slight bias in our method, which like many other methods, has higher power to detect outliers in regions with low recombination (long blocks of ILS are more easily detected with low recombination because they tend to contain more SNPs).

Candidates of Balancing Selection

Table S11.4 shows the combined list of candidate regions and their estimates of diversity, number of shared SNPs, coverage, and recombination rate. Table S11.5 shows the percentage of all blocks of 50 kb in the genome with more extreme (higher) values than the candidate region. We identify 2 regions that are in the 5% tail of the empirical distribution for high diversity and number of shared SNPs. These regions are not enriched for high coverage in neither chimpanzee nor bonobo, and are generally low in average recombination rate.

The first candidate region is located on chromosome 15 and overlaps the open reading frame *C15orf42*. Upon further investigation, we detected a 6 kb-long region (chr15:43029000-43035000) that contains most of the shared SNPs in this locus and shows elevated diversity levels in chimpanzee and bonobo as well as high coverage of reads from Ulindi (see Fig. S11.2). When we align this region to the chimpanzee and bonobo genomes (using BLAT [1], standard parameters), we detect a close second best hit in these genomes that is absent from human. The region thus seems to be duplicated in bonobo and chimpanzee, but not in human. Since of our sequence data was aligned to the human genome, these duplications would thus be overcollapsed and lead to a erroneous signal of shared SNPs and high diversity. We thus exclude this region as a likely false positive.

The second candidate region is on chromosome 6, contains the genes *HLA-DQA1* and *HLA-DQB1*, and is part of the MHC locus. Genes in the MHC locus remain some of the prime examples of balancing selection in humans [96, 97]. Table S11.6 shows that the fraction of ILS is elevated for the entire MHC region.

We identify several other regions that show some of the signatures considered here (top 10% in at

least two measures) while giving no signal for high average coverage or recombination. Upon closer individual inspection, we find it likely that these candidate regions are false positives due to different types of issues, including undetected duplications, misalignments, and read-merging artifacts.

In summary, we identified one candidate, in the MHC region, as showing the signatures of balancing selection targeted here. Upon close individual inspection, all of the other candidates were filtered out due to possible technical issues. Although genes with the specific signatures of balancing selection have been detected in the human genome [102], some of which have been further confirmed (e.g. [103, 104]), the MHC region may constitute a rare case in which balancing selection has acted over a sufficiently long time on the same pair of alleles and in different lineages to cause a signal of incomplete lineage sorting between the human, chimpanzee and bonobo genomes, while increasing diversity and trans-species polymorphism in extant populations. Actually, other aspects of the MHC signal, such as the high levels of trans-species polymorphism and the high diversity linked to high LD, have been shown to be unusual in the human genome [105, 106]. Therefore, in many aspects the MHC might represent an unusual target of balancing selection in primates.

The number of possible artifacts among our original set of candidates evidences the importance of careful examination of outliers in genome-wide studies. However, regions of high ILS are likely enriched for candidates of balancing selection and ILS may serve as an additional measure to identify those candidates.

Location	%CH+BH	covB	covC	divB	divC	sharedBC	sharedBH	sharedCH	Recomb
6:32700000-32750000	98.7%	7.0	1.2	0.0025	0.0087	4.01E-03	6.82E-02	6.18E-02	0.7177
15:43000000-43050000	97.2%	8.7	1.5	0.0028	0.0053	3.22E-03	3.57E-02	5.00E-02	0.6134
15:40950000-41000000	97.1%	8.6	1.5	0.0006	0.0009	3.93E-05	3.33E-02	0.00E+00	0.1848
6:57350000-57400000	95.9%	13.9	1.9	0.0071	0.0110	1.09E-02	1.28E-01	1.76E-01	0.1683
19:47700000-47750000	94.4%	8.1	1.5	0.0006	0.0034	0.00E+00	0.00E+00	0.00E+00	0.1455
19:20550000-20600000	93.9%	8.1	1.4	0.0012	0.0031	6.03E-04	0.00E+00	2.53E-02	0.4932
19:20650000-20700000	93.1%	8.7	1.4	0.0010	0.0020	4.29E-05	0.00E+00	1.87E-02	0.5021
13:85950000-86000000	92.8%	9.2	1.4	0.0004	0.0017	0.00E+00	6.06E-03	1.39E-02	0.3025
14:42950000-43000000	89.3%	8.8	1.4	0.0012	0.0020	0.00E+00	0.00E+00	1.10E-02	0.2212
1:25550000-25600000	87.5%	8.4	1.4	0.0004	0.0084	0.00E+00	0.00E+00	0.00E+00	1.2900
5:150250000-150300000	83.5%	9.5	1.5	0.0008	0.0016	0.00E+00	0.00E+00	1.06E-02	1.3416
1:88800000-88850000	82.5%	9.0	1.5	0.0006	0.0016	0.00E+00	0.00E+00	0.00E+00	0.4601
11:89050000-89100000	81.5%	5.4	1.3	0.0011	0.0041	7.39E-04	2.50E-01	2.31E-01	0.3153
2:70100000-70150000	80.4%	8.3	1.5	0.0010	0.0018	0.00E+00	0.00E+00	1.14E-02	0.6020
19:22100000-22150000	80.0%	8.1	1.4	0.0009	0.0067	2.95E-04	1.00E-02	2.23E-02	0.4209
6:31350000-31400000	78.8%	1.4	1.1	0.0010	0.0027	0.00E+00	0.00E+00	4.17E-02	0.4165

Table S11.4: Values for the ILS measure (%CH+BH), coverage (chimpanzee=covC, bonobo=covB), diversity (chimpanzee=divC, bonobo=divB), shared SNPs (bonobo-chimpanzee=sharedBC, bonobo-human=sharedBH, human-chimpanzee=sharedCH), and recombination (recomb). Green rows show the candidates within the top 5% for diversity and shared SNPs among all genome wide regions.

Location	qCHBH	qcovB	qcovC	qdivB	qdivC	qsharedBC	qsharedBH	qsharedCH	qrecomb
6:32700000-32750000	0.00%	96.96%	98.79%	0.55%	0.43%	0.25%	0.14%	0.26%	59.51%
15:43000000-43050000	0.00%	68.90%	63.02%	0.45%	0.86%	0.31%	0.34%	0.34%	67.09%
15:40950000-41000000	0.00%	71.64%	67.92%	82.49%	99.00%	31.14%	0.40%	88.58%	95.30%
6:57350000-57400000	0.01%	0.06%	0.13%	0.12%	0.31%	0.06%	0.07%	0.07%	95.92%
19:47700000-47750000	0.01%	88.61%	28.90%	85.88%	5.10%	74.87%	61.81%	88.58%	96.70%
19:20550000-20600000	0.01%	88.38%	86.73%	9.65%	8.08%	1.20%	61.81%	2.91%	76.15%
19:20650000-20700000	0.01%	67.01%	95.66%	22.89%	52.04%	27.16%	61.81%	8.89%	75.52%
13:85950000-86000000	0.02%	32.89%	96.19%	98.53%	78.25%	74.87%	20.29%	19.60%	89.48%
14:42950000-43000000	0.02%	57.64%	85.24%	8.96%	51.14%	74.87%	61.81%	30.36%	93.56%
1:25550000-25600000	0.02%	79.87%	89.57%	98.41%	0.44%	74.87%	61.81%	88.58%	27.84%
5:150250000-150300000	0.02%	11.87%	28.80%	55.36%	82.28%	74.87%	61.81%	32.13%	25.84%
1:88800000-88850000	0.03%	44.01%	67.06%	89.18%	80.48%	74.87%	61.81%	88.58%	78.48%
11:89050000-89100000	0.03%	97.97%	98.20%	10.17%	1.88%	0.98%	0.02%	0.04%	88.75%
2:70100000-70150000	0.03%	83.84%	58.73%	17.45%	71.18%	74.87%	61.81%	28.54%	67.92%
19:22100000-22150000	0.03%	89.56%	97.07%	25.13%	0.61%	2.99%	8.43%	4.85%	81.47%
6:31350000-31400000	0.04%	98.93%	99.15%	16.00%	16.48%	74.87%	61.81%	0.51%	81.80%

Table S11.5: Quantiles (sorted descending for all values) for the ILS measure (%CH+BH), coverage (chimpanzee=covC, bonobo=covB), diversity (chimpanzee=divC, bonobo=divB), shared SNPs (bonobo-chimpanzee=sharedBC, bonobo-human=sharedBH, human-chimpanzee=sharedCH), and recombination (recomb). Green rows show the candidates within the top 5% for diversity and shared SNPs among all genome wide regions.

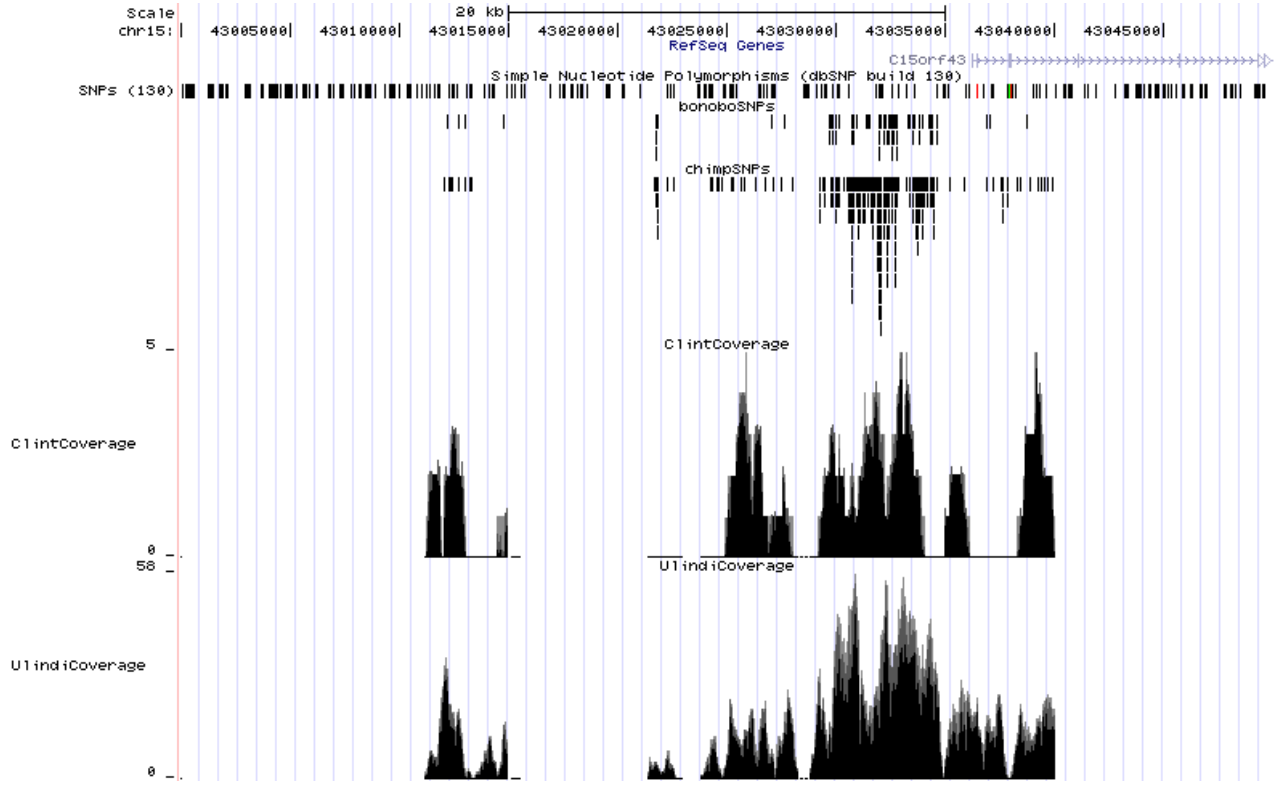


Figure S11.2: UCSC genome browser snapshot of significant candidate on chromosome 15. Shown are the density of SNPs in bonobo and chimpanzee (bonoboSNPs, chimpSNPs) and the coverage by Ulindi and Clint reads aligning to the human genome (ClintCoverage, UlindiCoverage).

ILS-state	Bases	%
BC	616835	92.8
BC2	10465	1.6
BH	24500	3.7
CH	12701	1.9

Table S11.6: Counts of bases in each of the ILS-classes in the MHC region (here defined as chr6:29750000-33200000 on hg18).

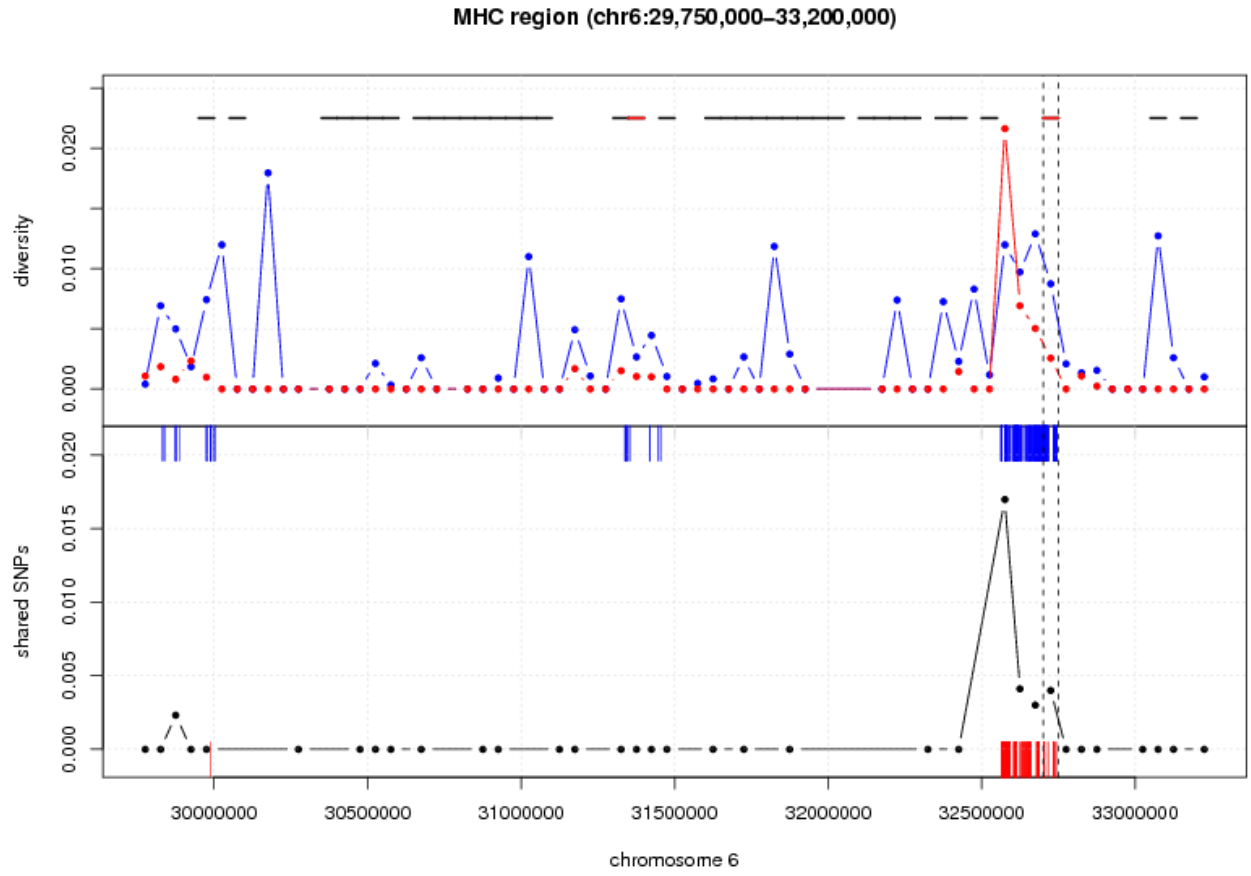


Figure S11.3: Diversity and shared SNPs in the MHC region. The top panel shows the diversity in bonobo (red) and chimpanzee (blue) in 50kb windows. Horizontal dashes in the top panel indicate the regions for which an ILS measure was calculated (black for %CH+BH < 0.78 and red for %CH+BH > 0.78). The lower panel shows the fraction of bonobo-chimpanzee shared SNPs in 50kb windows (black line), chimpanzee-human shared SNP density (blue dashes) and bonobo-human shared SNP density (red dashes). The significant 50kb window is located between the dashed black lines.

Supplementary Information 12

Protein Coding Differences between Bonobo and Chimpanzee

Kay Prüfer*, Aida Andrés, Janet Kelso and Svante Pääbo

Max-Planck Institute for Evolutionary Anthropology, Leipzig, Germany

* To whom correspondence should be addressed (pruefer@eva.mpg.de)

We use our resequencing data to identify differences between bonobo and chimpanzee that are likely fixed and overlap coding regions. We intersect the resulting list of sequence differences with the regions where bonobo is predicted to fall outside the chimpanzee variation (external regions) and regions identified as incomplete lineage sorting. In addition, we find several cases in which bonobo or chimpanzee carry premature stop mutations.

Identifying Fixed Differences and Overlap with Coding Regions

We use all Illumina-sequenced chimpanzees and bonobos in addition to the data of the reference chimpanzee (Clint) and bonobo (Ulindi) genomes. Reads were aligned and filtered as described earlier (see SI 2, 5 and 7). In short, reads were aligned against the human genome (hg18) and mapping quality filtered, bases were masked according to quality score and other criteria, and one random read was chosen per individual at each site. We then retain all sites that are covered by at least two bonobo individuals and two chimpanzee individuals. All chimpanzee and all bonobo individuals are required to show the identical base and the chimpanzee and bonobo bases must differ. A total of 2.7 million sites on all autosomes match our criteria.

Next, we intersect the list of potentially fixed differences between chimpanzee and bonobo with the CCDS gene annotation for the human genome (downloaded from the UCSC genome browser for release hg18). We find a total of 10813 sites overlapping annotated coding regions; 4403 sites cause amino acid substitutions (non-synonymous) while 6410 sites do not alter the amino acid (synonymous).

Due to the difference in individuals available for chimpanzee (16+Clint) and for bonobo (3+Ulindi), sites may more often be wrongly classified as fixed (while they are truly SNPs) in bonobo than chimpanzee. On average, a site is covered by 6.8 different chimpanzee individuals and 2.3 different bonobo individuals cover each site. (We sample one high-quality base per individual, so that these numbers correspond to the number of compared chromosomes to call SNPs in each ape.) This difference in coverage is reflected in the number of differences assigned to each lineage using human as outgroup. A total of 3327 sites (1345 non-synonymous, 1982 synonymous) are different in chimpanzee while bonobo and human show the same variant. On the other hand, 7439 sites (3042 non-synonymous, 4397 synonymous) are different in bonobo as compared to chimpanzee and human.

Incomplete Lineage Sorting and Coding Sequence Changes

We further our analysis by identifying amino acid changes that may have been under incomplete lineage sorting (ILS). We identify such sites as bases where either human and chimpanzee, but not bonobo, share a derived base (identified by the variant in orangutan and rhesus macaque), or where human and bonobo, but not chimpanzee, share a derived base. We find a total of 457 sites matching these criteria. A total of 282 sites (82 non-synonymous, 200 synonymous) show a grouping of chimpanzee-human (CH state), while a further 175 sites (66 non-synonymous, 109 synonymous) group bonobo-human (BH state). The difference in total counts is most likely caused by the difference in depth of coverage between chimpanzee and bonobo to call fixed differences. Some of these sites may have been created by recurrent mutation (reverting either chimpanzee or bonobo back to the ancestral state).

In order to improve our power to detect true differences caused by ILS, we overlap all sites with the annotation from the ILS analysis (SI 8). A total of 97 sites (39 non-synonymous, 58 synonymous) overlap regions assigned the ILS states CH or BH. Given that 3.2% of the genome is assigned as an ILS region, this overlap is significantly higher than expected at random (binomial test, p-value < 2.2e-16), even if we are conservatively ignoring the fact that the fraction of ILS-assigned bases is lower for coding regions, and that some bases may be filtered in the ILS HMM analysis but not this test. This high congruency is expected since Ulindi and Clint data are included in this test and the earlier ILS analysis.

Potential ILS-sites outside of ILS-regions of the HMM analysis may be caused by recurrent mutations. In order to test for the presence of recurrent mutations, we further classify ILS-sites as CpG sites and non-CpG sites. CpG sites are known to mutate faster than other sites in the genome and may thus lead to a higher rate of double mutation that mimics the appearance of ILS. In agreement with this expectation, we find that potential ILS-sites outside of ILS regions are more often CpG sites than inside of ILS regions. Of 97 sites inside ILS regions, 46 (47%) overlap a CpG site; of 360 sites outside ILS regions, 303 (84%) overlap CpG sites.

Of all overlapping substitutions that indicate ILS, 42 overlap regions assigned to the state BH by the coalescent HMM analysis and 55 overlap regions with assigned state CH. As expected, the inferred state based on the type of substitution agrees well with the assigned state by the coalescent HMM. 40/42 substitutions indicated BH grouping and fall in BH regions (i.e. 2 of 42 substitutions were compatible with a CH state, but were found within a BH region), and 52/55 substitutions indicate CH grouping and fall in CH regions (i.e. 3 BH-sites were found in CH regions).

For the remaining analysis we only consider substitutions whose state is matched by the coalescent HMM and that cause an amino acid change. We find 18 BH ILS amino-acid changes and 18 CH ILS amino-acid changes (see Table S12.1 and S12.2).

Chromosome	Position	C AA	B AA	H AA	O AA	M AA	#C	#B	CCDS id	Gene Name	AA position
chr1	11830010	P	Q	P	Q	Q	4	2	CCDS139.1	NPPA	66
chr1	40753734	K	E	K	E	E	5	2	CCDS453.1	DEM1	311
chr1	47274129	S	L	S	L	L	10	4	CCDS544.1	CYP4X1	186
chr1	149604081	I	V	I	V	V	7	3	CCDS995.1	SELENBP1	401
chr11	7021323	S	N	S	N	N	10	3	CCDS7776.1	NLRP14	497
chr12	111838906	R	S	R	S	S	7	2	CCDS41838.1	OAS1	288
chr12	111838917	T	K	T	R	R	8	2	CCDS41838.1	OAS1	292
chr12	124127104	I	V	I	V	V	8	3	CCDS9263.1	AACS	118
chr17	73634248	T	A	T	A	A	4	2	CCDS32748.1	TMC6	45
chr19	62340145	G	E	G	E	E	8	2	CCDS33125.1	ZIM3	50
chr22	25214138	L	V	L	L	L	7	2	CCDS42995.1	SRRD	132
chr22	25214204	V	I	V	I	I	11	4	CCDS42995.1	SRRD	154
chr3	114176343	I	V	I	V	V	5	2	CCDS2970.1	CD200R1	18
chr3	186735518	K	R	K	R	R	13	2	CCDS3272.1	LIPH	49
chr5	235391	C	R	C	R	R	2	2	CCDS34124.1	PLEKHG4B	1257
chr5	150927359	V	M	V	M	M	7	4	CCDS4317.1	FAT2	443
chr6	165623513	R	Q	R	Q	Q	7	2	CCDS5288.1	C6orf118	385
chr8	24402668	H	D	H	D	D	6	2	CCDS6045.1	ADAM7	400

Table S12.1: Amino-acid substitutions between chimpanzee and bonobo overlapping regions of ILS with state CH. Columns C AA, B AA, H AA, O AA and M AA give the amino acid for this position in chimpanzee, bonobo, human, orangutan and rhesus macaque, respectively. Columns #C and #B give the number of supporting individuals in chimpanzee and bonobo.

Chromosome	Position	C AA	B AA	H AA	O AA	M AA	#C	#B	CCDS id	Gene Name	AA Position
chr1	18680486	G	S	S	G	G	3	2	CCDS185.2	KLHDC7A	142
chr1	220989911	K	R	R	K	K	11	2	CCDS1535.2	FAM177B	122
chr1	246071664	Q	L	L	Q	Q	5	2	CCDS31098.1	OR11L1	53
chr10	99330672	S	N	N	S	S	4	2	CCDS7466.1	ANKRD2	203
chr13	109916377	I	V	V	I	I	5	3	CCDS41907.1	COL4A2	669
chr14	94000826	T	A	A	T	T	7	2	CCDS41983.1	SERPINA9	259
chr15	72375217	E	K	K	E	E	7	2	CCDS42058.1	CCDC33	389
chr19	42059637	S	N	N	S	S	7	3	CCDS12497.1	ZNF345	22
chr19	49431029	H	Q	Q	H	H	15	2	CCDS12636.1	ZNF227	202
chr19	62696107	V	I	I	V	V	3	2	CCDS42637.1	ZNF419	92
chr22	18188875	A	T	T	A	A	2	2	CCDS13768.1	GNB1L	2
chr3	188436887	E	A	A	E	E	5	2	CCDS33908.1	MASP1	489
chr5	1269981	R	Q	Q	R	R	4	3	CCDS34130.1	SLC6A19	365
chr5	81649837	E	K	K	E	E	5	3	CCDS34196.1	LOC92270	213
chr5	82436538	N	S	S	N	N	6	2	CCDS4059.1	XRCC4	15
chr6	132915591	K	E	E	K	K	9	3	CCDS5154.1	TAAR8	23
chr8	24226937	D	N	N	D	D	6	2	CCDS34865.1	ADAM28	159
chr8	139697473	P	S	S	P	P	6	2	CCDS6376.1	COL22A1	1293

Table S12.2: Amino-acid substitutions between chimpanzee and bonobo overlapping regions of ILS with state BH. Columns C AA, B AA, H AA, O AA and M AA give the amino acid for this position in chimpanzee, bonobo, human, orangutan and rhesus macaque, respectively. Columns #C and #B give the number of supporting inviduals in chimpanzee and bonobo.

Fixed Changes in External Regions

In SI 12, we identified a number of regions where bonobo falls outside of the variation of chimpanzee. The regions were scored according to genetic width and corrected for the confounding factor background selection. The high-ranking regions may encompass candidates for selective sweeps leading to a coalescence of all chimpanzee lineages after the split from bonobo. Here, we use the top 100 candidate regions (top external regions) to find the candidate substitutions that may underlie a selective sweep.

In a first step, we identify all fixed differences between chimpanzee and bonobo in the top external regions (without restriction to coding sequence). We find a total of 7702 differences in the regions. Using the human variant as an outgroup, we sort the differences to the chimpanzee and bonobo lineage. 2610 changes were assigned to the chimpanzee lineage and 5045 to the bonobo lineage. As described earlier, due to the lower number of bonobo individuals, bonobo polymorphisms may more often be classified as fixed according to our criteria and likely lead to the higher number of assigned changes to the bonobo lineage. In order to test whether the fraction of fixed differences in the external regions is higher on the chimpanzee lineage than expected, we also assign all differences genome wide. We find that on average 33.1% of all classifiable differences are assigned to the chimpanzee lineage. The fraction of chimpanzee changes in the top external regions give a higher percentage of 34.1%. The top external regions are significantly higher in the fraction of chimpanzee lineages changes than the genome-average (binomial test, one-sided, p-value=0.038). When we use all external regions (i.e. do not restrict to the top 100 candidate regions), this difference is even more pronounced (35.3% assigned to chimpanzee lineage, binomial test, one-sided: p-value=1.467e-11).

The difference in assigned changes between top external regions and genome average may be caused by a difference in coverage by chimpanzee individuals — lower coverage in external regions could lead to an inclusion, as fixed differences, of high-frequency derived variants on the chimpanzee lineage; the higher coverage genome-wide would allow their detection as polymorphic. However, we find that on average 6.9 chimpanzee individuals and 2.3 bonobo individuals cover fixed positions in the top external regions, very similar to the values observed genome wide (6.8 and 2.3, respectively). Similarly, bonobo lineage changes are not depleted in regions due to either higher bonobo or lower chimpanzee coverage (7.3 and 2.3 for chimpanzee and bonobo in regions and 7.2 and 2.3 genome-wide). We conclude that the top external regions are enriched for fixed chimpanzee variants as compared to bonobo. This enrichment is expected given that the regions are classified as external. It could be explained by three factors: 1) Neutral variation in chimpanzee coalescent times that is not reflected on the bonobo lineage, 2) chimpanzee selective sweeps, and 3) a change in background selection between bonobo and chimpanzee, reducing the effective population size of chimpanzee and leading to a more recent coalescence of the chimpanzee lineages. Given that the top external regions are outliers in the genetic-length distribution, we expect an enrichment for sweeps and change in background selection.

We further our analysis by intersecting the list of fixed differences with coding regions for the top external regions. We find a total of 22 differences falling within coding regions (13 non-synonymous, 9 synonymous). Of the 13 non-synonymous changes, 7 are assigned to the bonobo lineage and 6 to the

chimpanzee lineage. Upon manual inspection we exclude 2 of the 6 chimpanzee specific changes since the whole genome-alignment of human, chimpanzee and bonobo does not show the lineage-specific substitution. The remaining 4 chimpanzee specific changes (see Table 12.3) are candidates for mutations on which positive selection may have acted.

Chromosome	Position	C AA	B AA	H AA	O AA	M AA	#C	#B	CCDS id	Gene Name	AA Position
chr16	65878444	S	N	N	N	N	3	2	CCDS32466.1	PLEKHG4	989
chr2	203862798	H	L	L	L	L	5	2	CCDS2357.1	CYP20A1	346
chr4	48747587	R	H	H	H	H	8	2	CCDS3486.1	FLJ21511	662
chr8	100274407	I	V	V	V	V	8	2	CCDS6283.1	VPS13B	821

Table S12.3: Chimpanzee-specific amino-acid substitutions falling inside the top 100 external regions. Columns C AA, B AA, H AA, O AA and M AA give the amino acid for this position in chimpanzee, bonobo, human, orangutan and rhesus macaque, respectively. Columns #C and #B give the number of supporting inviduals in chimpanzee and bonobo.

Stop Mutations

In a last step, we used the fixed differences to scan for premature stop mutations on the chimpanzee and bonobo lineages. We require that all outgroups (human, orangutan, rhesus macaque) are aligned and that only chimpanzee and bonobo differ. We find 21 bonobo-specific stop mutations and 9 chimpanzee-specific stop mutations. These changes may lead to non-functional gene products and may contribute to phenotypic differences between chimpanzee and bonobo. In addition, we found one additional mutation in each lineage that changes the stop codon to an amino-acid, extending the ORF of the protein. The chimpanzee mutation leads to an extension by 4 amino-acids ("RCNN" in ZNF510) and the bonobo mutation leads to an extension by 1 amino-acid ("C" in BCAR3). Table 12.4 summarizes our findings.

Chromosome	Position	C AA	B AA	H AA	O AA	M AA	#C	#B	CCDS id	Gene Name	AA Position	protein length
chr11	5099116	*	W	W	W	W	8	3	CCDS31372.1	OR52A4	90	305
chr12	111839855	*	W	W	W	W	4	2	CCDS41838.1	OAS1	335	401
chr14	19552470	*	C	C	C	C	11	4	CCDS32027.1	OR4K14	241	311
chr2	79108468	*	E	E	E	E	12	4	CCDS1962.1	REG3G	121	176
chr20	1499701	*	Y	Y	Y	Y	4	2	CCDS13019.1	SIRPB1	278	399
chr5	137303727	*	R	R	R	*	14	3	CCDS43367.1	PKD2L2	612	614
chr6	50044404	*	W	W	W	W	3	2	CCDS43472.1	DEFB113	65	83
chr6	117235054	*	Y	Y	Y	Y	9	2	CCDS5112.1	GPRC6A	169	927
chr6	132915543	*	Q	Q	Q	Q	6	3	CCDS5154.1	TAAR8	7	343
chr9	98560883	R	*	*	*	*	12	3	CCDS35074.1	ZNF510	684	684
chr1	6502122	R	*	R	R	R	6	2	CCDS41240.1	PLEKHG5	13	1084
chr1	93800386	*	C	*	*	*	7	3	CCDS745.1	BCAR3	826	826
chr10	15147651	Y	*	Y	Y	Y	8	2	CCDS31152.1	OLAH	155	266
chr11	57715352	R	*	R	R	R	5	2	CCDS31544.1	OR9Q2	272	315
chr11	123129950	Q	*	Q	Q	Q	9	2	CCDS31695.1	OR6X1	163	313

chr11	123399827	K	*	K	K	K	2	2	CCDS31703.1	OR10G9	300	312
chr12	99180179	K	*	K	K	K	7	2	CCDS9075.1	DEPDC4	232	295
chr14	19365478	E	*	E	E	E	5	2	CCDS32022.1	OR4N2	11	308
chr15	40433905	Q	*	Q	Q	Q	5	2	CCDS42027.1	CAPN3	4	729
chr15	63036548	Q	*	Q	Q	E	5	2	CCDS10197.2	ANKDD1A	513	523
chr17	31206263	R	*	R	R	R	9	3	CCDS11299.1	C17orf66	544	571
chr18	18989900	Y	*	Y	Y	Y	10	2	CCDS42418.1	CABLES1	12	369
chr19	9313494	Q	*	Q	Q	Q	7	2	CCDS12211.1	ZNF559	123	539
chr19	56348300	K	*	K	K	K	11	3	CCDS42601.1	SIGLEC7	371	375
chr19	62697265	Q	*	Q	Q	Q	8	2	CCDS42637.1	ZNF419	478	479
chr2	70897480	Q	*	Q	Q	Q	6	3	CCDS1910.1	CLEC4F	181	590
chr2	234292761	Q	*	Q	Q	Q	6	2	CCDS33405.1	UGT1A4	186	535
chr21	39480859	L	*	L	L	L	10	2	CCDS13662.1	BRWD1	2309	2321
chr3	49175630	W	*	W	W	W	2	2	CCDS2790.1	CCDC71	339	468
chr3	95262793	Q	*	Q	Q	Q	7	4	CCDS2926.1	DHFRL1	85	188
chr4	68777661	Y	*	Y	Y	Y	8	2	CCDS3521.1	TMPRSS11B	285	417
chr7	143264177	K	*	K	K	K	4	2	CCDS43666.1	OR2F2	307	318

Table S12.4: Premature-stop mutations in bonobo and chimpanzee. Columns C AA, B AA, H AA, O AA and M AA give the amino acid for this position in chimpanzee, bonobo, human, orangutan and rhesus macaque, respectively. Columns #C and #B give the number of supporting inviduals in chimpanzee and bonobo. Green background indicates chimpanzee lineage mutations, blue background bonobo-lineage mutations.

TAAR8 Contains Stop Mutations and Evidence for ILS

The sequence of the trace amine associated receptor 8 (TAAR8) contains an amino-acid exchange that is compatible with bonobo-human ILS. It additionally contains a premature stop codon at amino-acid position seven in chimpanzee. We extracted the coding region, as annotated in the human genome, using a whole-genome alignment of human, chimpanzee, bonobo, gorilla (gorgor3), orangutan and rhesus macaque (prepared with the same process and identical parameters as given in SI 3). We find that all great apes contain premature stop codons as compared to human (see Figure S12.1 and Table S12.5).

TAAR8 is an G-coupled protein receptor of unknown function [107, 108]. It is predicted to contain 4 extracellular, 4 intracellular and 7 transmembrane domains (according to SwissProt[109] entry Q969N4 (TAAR8_HUMAN)) and has been found to be expressed in amygdala and kidney [107]. The bonobo open reading frame of 256 amino-acids shortens the product by two transmembrane, one intracellular (length 41 amino-acids) and one extracellular domain (length 1 amino-acids). The last intracellular domain in the predicted bonobo gene-product is shortened by two bases (of 42) and may be affected by six amino-acid changes caused by a frameshift mutation. Other Great Apes lost at least one more intracellular domain.

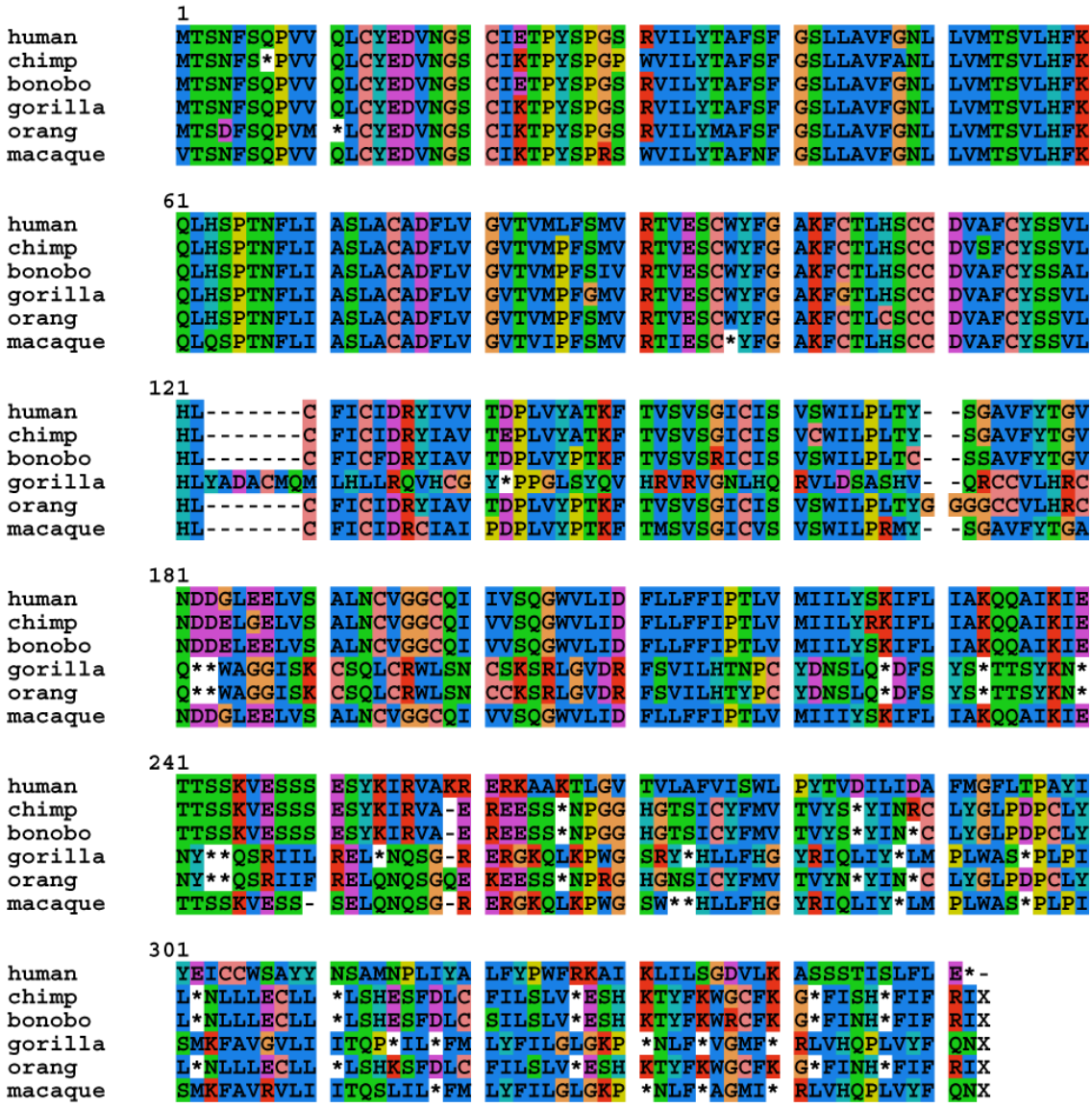


Figure S12.1: Protein alignment for the TAAR8 sequence of human, chimpanzee, bonobo, gorilla, orangutan and rhesus macaque.

Great Ape	ORF length in AAs
Human	342
Bonobo	256
Chimpanzee	7
Gorilla	180
Orangutan	11
Rhesus Macaque	97

Table S12.5: TAAR8 open reading frame length in amino-acids for all Great Apes.

In order to assess the level of protein conservation of the TAAR8 ORF in the different species, we calculated the ratio of the proportions of non-synonymous and synonymous changes in each lineage (dN/dS), using a ML approach implemented in the program *codeml* from the package PAML [110]. We tested two alignments: 1) the complete alignment to the human TAAR8 sequence and 2) the alignment up to the frameshift that leads to the premature stop in chimpanzee and bonobo (amino-acid position 251 in the human protein). For both alignments, we adjust all primate sequences to match the human open reading frame. For this, we delete insertions and mask stop mutations and deletions with the character "N". Note, therefore, that the dN/dS values of gorilla, orangutan and macaque do not necessarily reflect the true coding evolution in these species, since frameshift mutations modify the ORF in these species. These estimates provide, nevertheless, a comparison with the human protein-coding sequence.

Using a model that allows free variation of dN/dS among branches, we obtained the estimated values shown in Figure S12.2. When considering the complete TAAR8 gene, several lineages have a high dN/dS ratio, with only the human branch, internal branches, and the lineages of orangutan and macaque showing dN/dS<1 and being consistent with purifying selection. These results are also obtained when restricting the alignment to the bonobo ORF. The presence of a signal of purifying selection on the macaque and orangutan lineage hints towards an independent and recent pseudogenization of TAAR8 on these species. However, further study is needed to confirm the absence of polymorphism in these stop mutations and exclude the possibility of sequencing or assembly error. Similarly, the approach of calculating dN/dS ratio on the very short lineages of bonobo, chimpanzee and several internal branches does not allow us to draw final conclusions about the evolution of TAAR8. Future research, utilizing polymorphism data in a large number of bonobos and chimpanzees, has the potential to elucidate the current selective forces in these two apes in more detail.

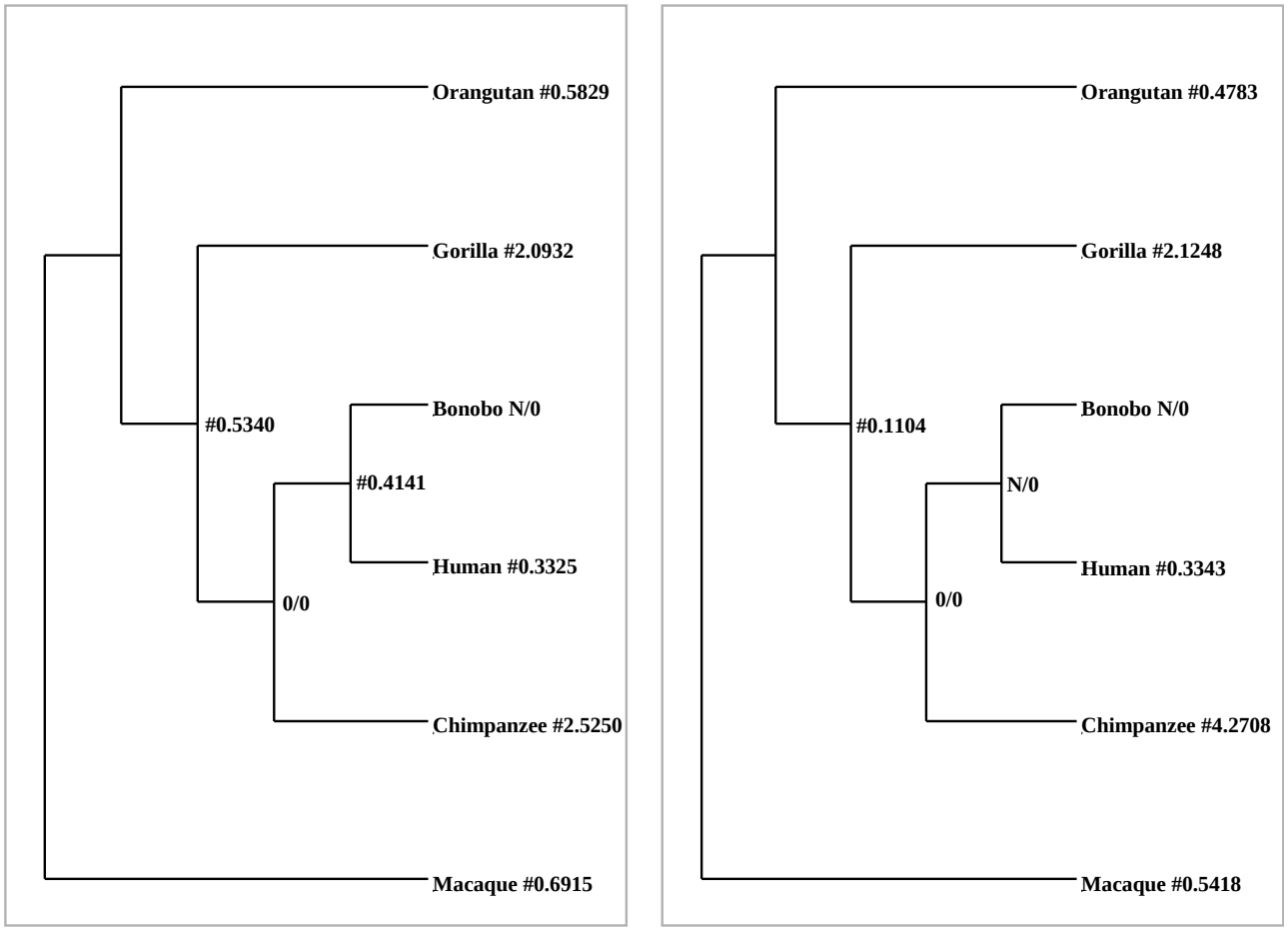


Figure S12.2: Rates of dN/dS for TAAR8, estimated for a full alignment to the human ORF (left) or to the truncated bonobo ORF (right). Value N/0 denotes a dS of zero while dN>0. Value 0/0 denotes dN of zero and dS of zero. A likelihood ratio test against a model with a constant dN/dS ratio on all branches is not significant for the full alignment (left, p=0.13; χ^2 test with d.f.=8), but significant for the truncated alignment (right, p=0.019; χ^2 test with d.f.=8).

References

1. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002. **12**(4): p. 656-64.
2. Schnable, P.S., et al., *The B73 maize genome: complexity, diversity, and dynamics.* Science, 2009. **326**(5956): p. 1112-5.
3. Consortium, C.S.a.A., *Initial sequence of the chimpanzee genome and comparison with the human genome.* Nature, 2005. **437**(7055): p. 69-87.
4. Zhang, Z., et al., *A greedy algorithm for aligning DNA sequences.* J Comput Biol, 2000. **7**(1-2): p. 203-14.
5. Myers, E.W., et al., *A whole-genome assembly of Drosophila.* Science, 2000. **287**(5461): p. 2196-204.
6. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
7. Istrail, S., et al., *Whole-genome shotgun assembly and comparison of human genome assemblies.* Proc Natl Acad Sci U S A, 2004. **101**(7): p. 1916-21.
8. Miller, J.R., et al., *Aggressive assembly of pyrosequencing reads with mates.* Bioinformatics, 2008. **24**(24): p. 2818-24.
9. Denisov, G., et al., *Consensus generation and variant detection by Celera Assembler.* Bioinformatics, 2008. **24**(8): p. 1035-40.
10. Gomez-Alvarez, V., T.K. Teal, and T.M. Schmidt, *Systematic artifacts in metagenomes from complex microbial communities.* ISME J, 2009. **3**(11): p. 1314-7.
11. Huson, D.H., K. Reinert, and E. Myers. *The greedy path-merging algorithm for sequence assembly*. 2001. ACM.
12. Myers, E.W., *Toward simplifying and accurately formulating fragment assembly.* J Comput Biol, 1995. **2**(2): p. 275-90.
13. Churchill, G.A. and M.S. Waterman, *The accuracy of DNA sequences: estimating sequence quality.* Genomics, 1992. **14**(1): p. 89-98.
14. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
15. Ning, Z., A.J. Cox, and J.C. Mullikin, *SSAHA: a fast search method for large DNA databases.* Genome Res, 2001. **11**(10): p. 1725-9.
16. Fischer, A., et al., *Evidence for a complex demographic history of chimpanzees.* Mol Biol Evol, 2004. **21**(5): p. 799-808.
17. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
18. Li, H. and R. Durbin, *Fast and accurate long-read alignment with Burrows-Wheeler transform.* Bioinformatics, 2010. **26**(5): p. 589-95.
19. Li, H., J. Ruan, and R. Durbin, *Mapping short DNA sequencing reads and calling variants using mapping quality scores.* Genome Res, 2008. **18**(11): p. 1851-8.
20. Altshuler, D., et al., *An SNP map of the human genome generated by reduced representation shotgun sequencing.* Nature, 2000. **407**(6803): p. 513-6.
21. Mullikin, J.C., et al., *An SNP map of human chromosome 22.* Nature, 2000. **407**(6803): p. 516-20.
22. Gibbs, R.A., et al., *Evolutionary and biomedical insights from the rhesus macaque genome.* Science, 2007. **316**(5822): p. 222-34.
23. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
24. Harris, R.S., *Improved Pairwise Alignment of Genomic DNA*, 2007, Pennsylvania State University.
25. Karolchik, D., et al., *The UCSC Genome Browser Database: 2008 update.* Nucleic Acids Res, 2008. **36**(Database issue): p. D773-9.
26. Kent, W.J., et al., *Evolution's cauldron: duplication, deletion, and rearrangement in the*

mouse and human genomes. Proc Natl Acad Sci U S A, 2003. **100**(20): p. 11484-9.
27. Blanchette, M., et al., *Aligning multiple genomic sequences with the threaded blockset aligner.* Genome Res, 2004. **14**(4): p. 708-15.
28. Bailey, J.A. and E.E. Eichler, *Primate segmental duplications: crucibles of evolution, diversity and disease.* Nat Rev Genet, 2006. **7**(7): p. 552-64.
29. Johnson, M.E., et al., *Recurrent duplication-driven transposition of DNA during hominoid evolution.* Proc Natl Acad Sci U S A, 2006. **103**(47): p. 17626-31.
30. She, X., et al., *The structure and evolution of centromeric transition regions within the human genome.* Nature, 2004. **430**(7002): p. 857-64.
31. Watanabe, H., et al., *DNA sequence and comparative analysis of chimpanzee chromosome 22.* Nature, 2004. **429**(6990): p. 382-8.
32. Lunter, G., C.P. Ponting, and J. Hein, *Genome-wide identification of human functional DNA using a neutral indel model.* PLoS Comput Biol, 2006. **2**(1): p. e5.
33. Meader, S., et al., *Genome assembly quality: Assessment and improvement using the neutral indel model.* Genome Res, 2010.
34. The Chimpanzee Sequencing and Analysis Consortium, *Initial sequence of the chimpanzee genome and comparison with the human genome.* Nature, 2005. **437**(7055): p. 69-87.
35. Bailey, J.A., et al., *Segmental duplications: organization and impact within the current human genome project assembly.* Genome Res, 2001. **11**(6): p. 1005-17.
36. Bailey, J.A., et al., *Recent segmental duplications in the human genome.* Science, 2002. **297**(5583): p. 1003-7.
37. Smit, A., R. Hubley, and P. Green, *RepeatMasker Open-3.0*, 1996-2004.
38. Cheng, Z., et al., *A genome-wide comparison of recent chimpanzee and human segmental duplications.* Nature, 2005. **437**(7055): p. 88-93.
39. Alkan, C., et al., *Personalized copy number and segmental duplication maps using next-generation sequencing.* Nat Genet, 2009. **41**(10): p. 1061-7.
40. Marques-Bonet, T., et al., *A burst of segmental duplications in the genome of the African great ape ancestor.* Nature, 2009. **457**(7231): p. 877-81.
41. Benson, G., *Tandem repeats finder: a program to analyze DNA sequences.* Nucleic Acids Res, 1999. **27**(2): p. 573-80.
42. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.
43. Day, N., et al., *Unsupervised segmentation of continuous genomic data.* Bioinformatics, 2007. **23**(11): p. 1424-6.
44. Fischer, A., et al., *Bonobos fall within the genomic variation of chimpanzees.* PLoS One, 2011. **6**(6): p. e21605.
45. Kircher, M., U. Stenzel, and J. Kelso, *Improved base calling for the Illumina Genome Analyzer using machine learning strategies.* Genome Biol, 2009. **10**(8): p. R83.
46. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
47. Stone, A.C., et al., *More reliable estimates of divergence times in Pan using complete mtDNA sequences and accounting for population structure.* Philos Trans R Soc Lond B Biol Sci. **365**(1556): p. 3277-88.
48. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
49. Cordaux, R. and M.A. Batzer, *The impact of retrotransposons on human genome evolution.* Nat Rev Genet, 2009. **10**(10): p. 691-703.
50. Salem, A.H., et al., *Alu elements and hominid phylogenetics.* Proc Natl Acad Sci U S A, 2003. **100**(22): p. 12787-91.
51. Hobolth, A., et al., *Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection.* Genome Res, 2011. **21**(3): p. 349-56.

52. Price, A.L., E. Eskin, and P.A. Pevzner, *Whole-genome analysis of Alu repeat elements reveals complex evolutionary history.* Genome Res, 2004. **14**(11): p. 2245-52.
53. Akagi, K., et al., *Extensive variation between inbred mouse strains due to endogenous L1 retrotransposition.* Genome Res, 2008. **18**(6): p. 869-80.
54. Wu, T.D. and C.K. Watanabe, *GMAP: a genomic mapping and alignment program for mRNA and EST sequences.* Bioinformatics, 2005. **21**(9): p. 1859-75.
55. Chimpanzee-Genome-Consortium, *Initial sequence of the chimpanzee genome and comparison with the human genome.* Nature, 2005. **437**(7055): p. 69-87.
56. Mills, R.E., et al., *Recently mobilized transposons in the human and chimpanzee genomes.* Am J Hum Genet, 2006. **78**(4): p. 671-9.
57. Xing, J., et al., *Mobile elements create structural variation: analysis of a complete human genome.* Genome Res, 2009. **19**(9): p. 1516-26.
58. Smit, A., Hubley, R., Green, P., *RepeatMasker Open-3.0.* 1996-2004.
59. Liu, G.E., et al., *Comparative analysis of Alu repeats in primate genomes.* Genome Res, 2009. **19**(5): p. 876-85.
60. Locke, D.P., et al., *Comparative and demographic analysis of orang-utan genomes.* Nature, 2011. **469**(7331): p. 529-33.
61. Levy, S., et al., *The diploid genome sequence of an individual human.* PLoS Biol, 2007. **5**(10): p. e254.
62. Hedges, D.J., et al., *Differential Alu mobilization and polymorphism among the human and chimpanzee lineages.* Genome Res, 2004. **14**(6): p. 1068-75.
63. Cordaux, R., et al., *Computational methods for the analysis of primate mobile elements.* Methods Mol Biol, 2010. **628**: p. 137-51.
64. Lee, J., et al., *Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons.* Gene, 2007. **390**(1-2): p. 18-27.
65. Mi, H., et al., *PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways.* Nucleic Acids Res, 2007. **35**(Database issue): p. D247-52.
66. Symer, D.E., et al., *Human l1 retrotransposition is associated with genetic instability in vivo.* Cell, 2002. **110**(3): p. 327-38.
67. Waterston, R.H., et al., *Initial sequencing and comparative analysis of the mouse genome.* Nature, 2002. **420**(6915): p. 520-62.
68. Han, J.S., S.T. Szak, and J.D. Boeke, *Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes.* Nature, 2004. **429**(6989): p. 268-74.
69. Kong, A., et al., *Fine-scale recombination rate differences between sexes, populations and individuals.* Nature, 2010. **467**(7319): p. 1099-103.
70. Ptak, S.E., et al., *Absence of the TAP2 human recombination hotspot in chimpanzees.* PLoS Biol, 2004. **2**(6): p. e155.
71. Winckler, W., et al., *Comparison of fine-scale recombination rates in humans and chimpanzees.* Science, 2005. **308**(5718): p. 107-11.
72. Hobolth, A., L.N. Andersen, and T. Mailund, *On computing the coalescence time density in an isolation-with-migration model with few samples.* Genetics, 2011. **187**(4): p. 1241-3.
73. Hudson, R.R., *Generating samples under a Wright-Fisher neutral model of genetic variation.* Bioinformatics, 2002. **18**(2): p. 337-8.
74. McVicker, G., et al., *Widespread genomic signatures of natural selection in hominid evolution.* PLoS Genet, 2009. **5**(5): p. e1000471.
75. Flicek, P., et al., *Ensembl 2011.* Nucleic Acids Res. **39**(Database issue): p. D800-6.
76. Prüfer, K., et al., *FUNC: a package for detecting significant associations between gene sets and ontological annotations.* BMC Bioinformatics, 2007. **8**: p. 41.
77. Dutheil, J.Y., et al., *Ancestral population genomics: the coalescent hidden Markov model approach.* Genetics, 2009. **183**(1): p. 259-74.
78. Ewing, B. and P. Green, *Base-calling of automated sequencer traces using phred. II. Error*

probabilities. Genome Res, 1998. **8**(3): p. 186-94.
79. Barreiro, L.B. and L. Quintana-Murci, *From evolutionary genetics to human immunology: how selection shapes host defence genes.* Nat Rev Genet. **11**(1): p. 17-30.
80. Samson, M., et al., *Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene.* Nature, 1996. **382**(6593): p. 722-5.
81. Haubold, B., P. Pfaffelhuber, and M. Lynch, *mlRho - a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes.* Mol Ecol, 2010. **19 Suppl 1**: p. 277-84.
82. Tavare, S., *Line-of-descent and genealogical processes, and their applications in population genetics models.* Theor Popul Biol, 1984. **26**(2): p. 119-64.
83. Hellmann, I., et al., *Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals.* Genome Res, 2008. **18**(7): p. 1020-9.
84. Murai, K.K. and E.B. Pasquale, *Can Eph receptors stimulate the mind?* Neuron, 2002. **33**(2): p. 159-62.
85. Konstantinova, I., et al., *EphA-Ephrin-A-mediated beta cell communication regulates insulin secretion from pancreatic islets.* Cell, 2007. **129**(2): p. 359-70.
86. Presgraves, D.C. and S.V. Yi, *Doubts about complex speciation between humans and chimpanzees.* Trends Ecol Evol, 2009. **24**(10): p. 533-40.
87. Hammer, M.F., et al., *Sex-biased evolutionary forces shape genomic patterns of human diversity.* PLoS Genet, 2008. **4**(9): p. e1000202.
88. Durbin, R.M., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
89. Keinan, A. and D. Reich, *Can a sex-biased human demography account for the reduced effective population size of chromosome X in non-Africans?* Mol Biol Evol, 2010. **27**(10): p. 2312-21.
90. Hammer, M.F., et al., *The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes.* Nat Genet, 2010. **42**(10): p. 830-1.
91. Lohmueller, K.E., et al., *Proportionally more deleterious genetic variation in European than in African populations.* Nature, 2008. **451**(7181): p. 994-7.
92. Green, R.E., et al., *Analysis of one million base pairs of Neanderthal DNA.* Nature, 2006. **444**(7117): p. 330-336.
93. Prüfer, K., et al., *Computational challenges in the analysis of ancient DNA.* Genome Biology, 2010. **11**(5): p. -.
94. Green, R.E., et al., *A draft sequence of the Neandertal genome.* Science, 2010. **328**(5979): p. 710-22.
95. Busing, F.M.T.A., E. Meijer, and R. Van Der Leeden, *Delete-m jackknife for unequal m.* Statistics and Computing, 1999. **9**(1): p. 3-8.
96. Hughes, A.L. and M. Nei, *Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection.* Nature, 1988. **335**(6186): p. 167-70.
97. Hughes, A.L. and M. Nei, *Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection.* Proc Natl Acad Sci U S A, 1989. **86**(3): p. 958-62.
98. Hodgkinson, A. and A. Eyre-Walker, *The genomic distribution and local context of coincident SNPs in human and chimpanzee.* Genome Biol Evol, 2010. **2**: p. 547-57.
99. Durbin, R.M., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2011. **467**(7319): p. 1061-73.
100. Hellmann, I., et al., *Why do human diversity levels vary at a megabase scale?* Genome Res, 2005. **15**(9): p. 1222-31.
101. Hobolth, A., et al., *Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model.* PLoS Genet, 2007. **3**(2): p. e7.
102. Andres, A.M., et al., *Targets of balancing selection in the human genome.* Mol Biol Evol, 2009. **26**(12): p. 2755-64.

103. Andres, A.M., et al., *Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation.* PLoS Genet, 2010. **6**(10): p. e1001157.
104. Cagliani, R., et al., *Genetic diversity at endoplasmic reticulum aminopeptidases is maintained by balancing selection and is associated with natural resistance to HIV-1 infection.* Hum Mol Genet, 2010. **19**(23): p. 4705-14.
105. Asthana, S., S. Schmidt, and S. Sunyaev, *A limited role for balancing selection.* Trends Genet, 2005. **21**(1): p. 30-2.
106. Bubb, K.L., et al., *Scan of human genome reveals no new Loci under ancient balancing selection.* Genetics, 2006. **173**(4): p. 2165-77.
107. Borowsky, B., et al., *Trace amines: identification of a family of mammalian G protein-coupled receptors.* Proc Natl Acad Sci U S A, 2001. **98**(16): p. 8966-71.
108. Lewin, A.H., *Receptors of mammalian trace amines.* Aaps J, 2006. **8**(1): p. E138-45.
109. Bairoch, A., et al., *Swiss-Prot: juggling between evolution and stability.* Brief Bioinform, 2004. **5**(1): p. 39-55.
110. Yang, Z., *PAML 4: phylogenetic analysis by maximum likelihood.* Mol Biol Evol, 2007. **24**(8): p. 1586-91.