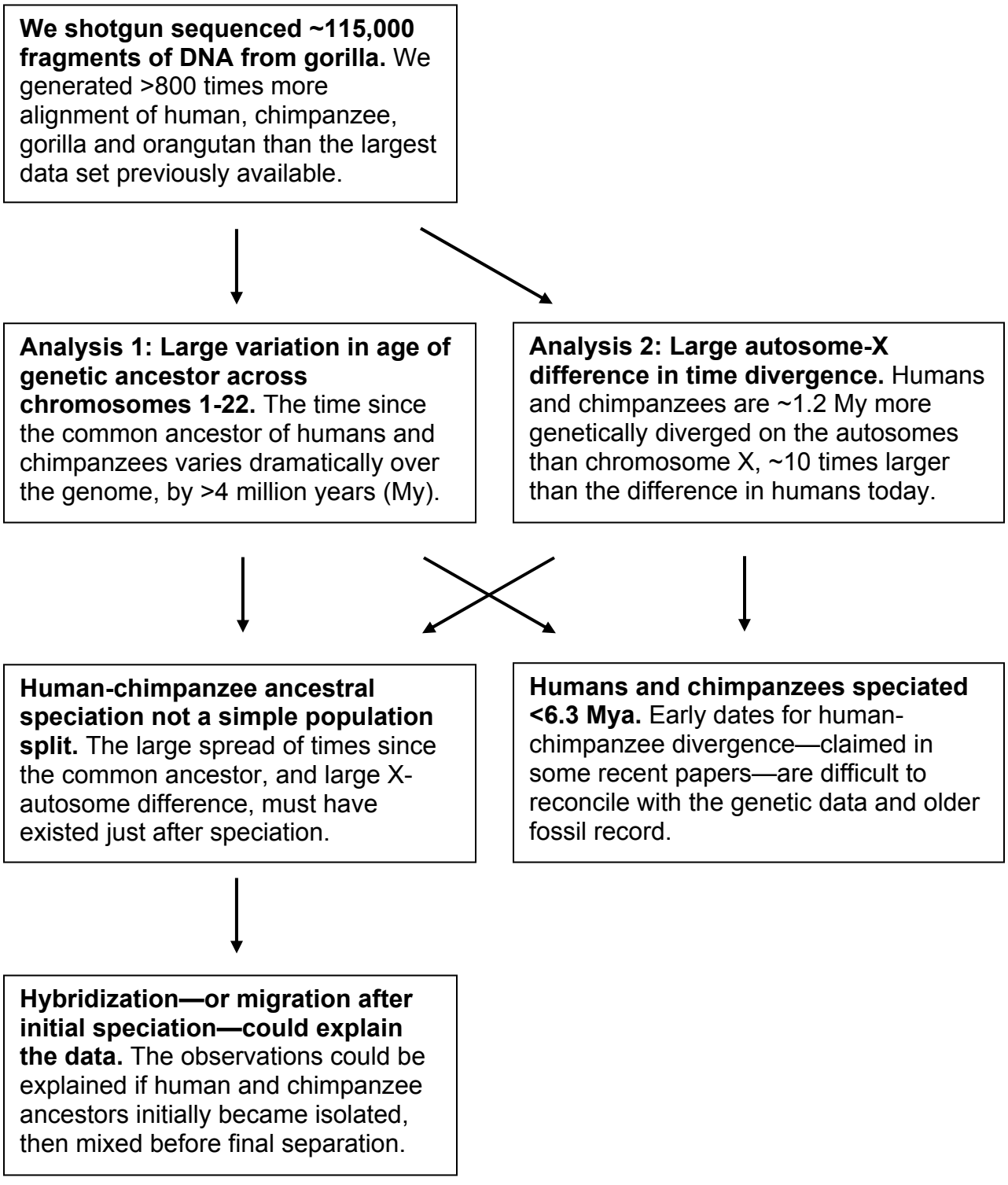
Supplementary Information

"Genetic evidence for complex speciation of humans and chimpanzees"
Patterson N, Richter DJ, Gnerre S, Lander ES and Reich D; *Nature* 2006

Note on January 23 2008 update to Supplementary Information: In the original Supplementary Information, we left out the correct Supp. Table 8, accidentally using the same image for both Supp. Tables 8 and 9. The only changes here are the substitution of the correct Supp. Table 8 and the addition of this cover note.)

Diagram of experimental work and main findings



Supplementary Methods

DNA sequence data (expansion of Methods): We sequenced 117,862 randomly distributed ("shotgun") fragments of DNA: 115,152 from a western lowland gorilla (*Gorilla gorilla,* individual NG05251 in the Coriell catalog: locus.umdnj.edu/primates/species_summ.html) and 2,710 from a black-handed spider monkey (*Ateles geoffryi*, individual NG05352). All these data are publicly available at the NCBI trace archive (www.ncbi.nlm.nih.gov/Traces/trace.cgi); to access them, see the instructions on "Accessing raw data and sequence alignments", below. We combined this with data from public sequencing projects (www.ncbi.nlm.nih.gov/Traces/trace.cgi) (Supp. Table 1). All reads were aligned to the NCBI Build 34 human genome assembly using either Arachne (Jaffe et al. 2003) or BLASTZ (Schwartz et al. 2003) (Supp. Table 1). Reads were only included in analysis if both ends of the clone from which they came aligned uniquely to the human reference sequence, in opposite orientation, and spaced by a distance consistent with the clone insert size. We supplemented the shotgun data with bacterial artificial chromosome sequencing across contiguous regions of chromosome 7 centered on the *CFTR* gene ("target 1" in the Comparative Vertebrate Sequencing project) (Hwang et al. 2004), as well as from the X chromosome centered on *ZXDA* and *ZXDB* ("target 46") (www.nisc.nih.gov/open_page.cgi?path=/projects/comp_seq.html).

Sequence alignments (expansion of Methods): To identify regions of the genome with sequence from all species of interest, we began with the species for which least sequence was available (usually gorilla), focusing only on reads that aligned to the human genome project reference sequence (NCBI Build 34). We overlapped reads from this species with reads from all the other species producing two-way overlaps (e.g. gorilla-orangutan, gorilla-macaque, gorilla-chimpanzee). These were then combined to produce 3- and finally 4-way overlaps. Overlaps were only used if they included ≥100bp of alignment from all species. For shotgun data we obtained local alignments using the Multiple Alignment Program (Huang 1994) with parameters gap_size=5, gap_open=4, gap_extend=3, match=1, mismatch=-2. For the contiguous data we used the Threaded Block Set Aligner (Blanchette et al. 2004) because it is optimized for larger segments of sequence. All alignments are available online and at our laboratory website (genepath.med.harvard.edu/~reich).

Eliminating misaligned reads (expansion of Methods): To minimize errors due to misalignment, we applied several filters, only including alignments that passed all these filters (Supp. Table 2). (1) We eliminated alignments where the observed number of within-species differences in a cluster was greater than expected from a Poisson distribution conservatively assuming a within-species polymorphism rate of 0.002 differences per base pair ($P<0.001$). (2) We eliminated alignments with unusually high shotgun coverage from any one species, that is, where the number of reads in the cluster divided by the length of the cluster in kilobases was >20 and 3 standard deviations above the mean. (4) We eliminated alignments where there was extreme asymmetry in the inferred branch lengths ($P<0.0001$ by a binomial test) (Supp. Fig. 1)—for example, where many more mutations had been observed on the chimpanzee than the human lineage since human-chimpanzee divergence. This also involved eliminating alignments where the outgroup species (usually macaque) had significantly fewer divergent sites than any of the other branches ($P<0.0001$). (4) We eliminated all alignments that mapped to known segmental duplications in either the human or chimpanzee genome (Cheng et al. 2004). The robustness of some of the main statistics after applying these filters is shown in Supp. Fig. 1. In addition, Supp.

Table 3 show that application of these filters to remove misaligned reads does not change the qualitative conclusions of our analysis.

Branch length estimates in the presence of recurrent mutations (expectation maximization algorithm): When enough species are available in an alignment (e.g. HCGOM), it is possible to observe divergent sites that are not consistent with a single historical mutation (e.g. HO, CO, GO, HCO, HGO and CGO) (Table 1). Not taking into account the recurrent mutation process can produce biased estimates of branch lengths (Supp. Table 5). We used an implementation of the expectation maximization algorithm (EM) (Baum et al. 1970) (Supp. Note 3) to obtain branch length estimates corrected for the presence of recurrent mutations. Briefly, the algorithm makes the simplifying assumptions that mutation rates have been historically constant at each locus, and that branch lengths do not vary across the genome. The algorithm then searches for the combination of branch lengths, for example t_H, t_C, t_G, t_O, t_{HC}, t_{HG}, t_{CG}, t_{HCG}, and t_M in the 5-species data, and relative probabilities of recurrent mutation (probability that sites are due to a history of 1, 2 or 3 mutations) that maximizes the likelihood of the observed data. Application to 5- and 6-species alignments (Supp Table 4) shows that in the presence of recurrent mutation, the EM algorithm finds a combination of branch lengths that in detail predicts the relative rates of events not consistent with a single mutation.

Branch length estimates in the presence of recurrent mutations (simulation analysis): The main limitation of the EM analysis is its assumption that genealogical histories are the same across the genome. To explore deviations from this assumption, we implemented a computer simulation that explicitly modeled variation in genealogical histories (Kingman 1982), thus finding a combination of parameters that produced data closely matching the HCGOM shotgun data from the autosomes (Supp. Table 11). In addition to producing branch length estimates close to those from the EM algorithm (Supp. Table 5), the modeling analysis has the additional feature of allowing us to estimate the true human-chimpanzee divergence near HC, HG or CG events. Near an HG or CG event not due to a recurrent mutation, the simulations suggest that true human-chimpanzee divergence is 1.47-fold of the average. Near an HC event, they indicate it is 0.86-fold of the average.

Jackknife error estimates: To obtain error estimates on the statistics analyzed for this study, we used a "weighted jackknife" procedure (Frank et al. 1999). To partition the data set into non-overlapping short regions of alignment for the jackknife analysis, we screened through the genome in order of the sequence, accumulating at least 50 divergent sites into each segment, before moving on to the next when a gap >10,000 bases was observed. For the contiguous data, we required at least 500 sites per segment and gaps of >100 bp. For each statistic for which we were interested in obtaining an error estimate, we calculated the statistic over the entire data set except for each segment, dropping out each segment in turn. By studying the variability in the statistic for each of these iterations, we obtained an error estimate.

Supplementary Table 1: Sources for shotgun sequence data

Species	Latin Name	Aligned Reads	Passing Reads	Data Source	Genome Aligner
Chimpanzee	*Pan troglodytes*	22,818,044	21,945,720	Broad Inst., Washington University	Arachne
Gorilla	*Gorilla gorilla*	115,152	108,331	This study	Arachne
Orangutan	*Pongo pygmaeus*	6,577,402	6,302,185	Baylor, Washington University	Arachne
Macaque	*Macaca mulatta*	15,968,510	13,810,571	Baylor, Venter Inst., Washington Univ.	Arachne
New World Monkey	*Ateles geoffroyi*	2,710	2,250	This study	BLASTZ

Supplementary Table 2: Filtering out poor alignments

	Number of alignments after application of each filter					Characteristics of final data set		
	Prior to filtering	High intraspecific polymorphism	Read pile-ups	Tree asymmetries	Segmental duplications	Bases per alignment (±1 s.d.)	Aligned bases (autosomes*)	Aligned bases (X chromosome)
HCGOM	33,016	31,706	31,644	31,637	30,839	300 ± 133	8,899,720	372,354
HCGM	51,966	50,083	50,038	50,021	48,689	376 ± 163	17,552,410	747,260
HCOM*	40,473	37,189	36,813	36,749	35,560	915 ± 593	17,525,320	15,160,801
HCMN	474	463	463	463	457	297 ± 115	133,691	not used

* The HCOM alignment only uses data from chromosomes 7 and X.

Supplementary Table 3: The main conclusions are robust to data filters

	$\tau^{HC}_{X\text{-}chrom}/\tau^{HC}_{genome}$	$\tau^{HC}_{near\ HC\ events}/\tau^{HC}_{genome}$
main data set	0.836 ± 0.022	0.862 ± 0.009
just CpGs	0.831 ± 0.049	0.850 ± 0.037
no requirement for sequence conservation flanking a site	0.857 ± 0.024	0.854 ± 0.008
≥2 bases conserved either side of each site	0.830 ± 0.024	0.852 ± 0.010
≥5 bases conserved either side of each site	0.818 ± 0.027	0.876 ± 0.016

Note: For the last three rows we do not filter out read misalignments.

Supplementary Table 4: The expectation maximization (EM) algorithm in detail predicts the rate of recurrent mutations

Contiguous HCGOMN alignment (autosomes)

Type of site	Observed divergent sites	Maximum likelihood estimate of sites due to recurrent mutations
H	2,664	14
C	2,667	14
G	3,611	18
O	7,955	71
M	19,199	413
HC	677	27
HG	60	23
CG	86	23
HCG	4,074	93
HCGO	7,095	1,115
N	50,867	221
HM	59	57
CM	54	57
GM	74	78
OM	446	398
HO	23	24
CO	22	24
GO	51	38
HCM	15	19
HGM	6	5
CGM	7	5
HOM	6	7
COM	3	6
GOM	33	41
HCO	32	39
HGO	16	25
CGO	23	25
HCOM	174	202
HGOM	145	149
CGOM	131	149
HCGM	557	534

Shotgun HCGOM alignment (autosomes)

Type of site	Observed divergent sites	Maximum likelihood estimate of sites due to recurrent mutations
H	28,504	331
C	28,495	331
G	38,677	473
HC	8,561	571
HG	1,302	436
CG	1,430	437
HCG	41,928	2,819
O	82,670	1,933
M	244,270	1,453
HO	412	462
CO	397	459
GO	764	777
HCO	1,347	1,286
HGO	989	912
CGO	872	910

The divergent sites categories in gray are the only sort that can be observed without recurrent mutation. The rates of the other sites can be predicted with surprisingly good accuracy, however, in a model where for a proportion of sites, mutations recur. For the 6-species alignment from contiguous chromosome 7 data (including a new-world monkey), we show a best-fitting model where 94.5% of bases experiencing a mutation since the MRCA had a single-hit, 5.1% a double-hit, and 0.4% a triple-hit. The 'expected' values are then a good match to the observed, making it likely that multiple-hit mutations are largely explaining the sites not consistent with a single historical mutation and that our inferences about the branch-lengths are roughly accurate. We present 6-species data, even though it is not the main focus of our analysis, because it illustrates more fully how the method predicts the rates of these sites. We also present data from our main 5-species shotgun alignment, where the best fit is 96.6% of mutations single-hit, 2.7% double-hit, and 0.7% triple-hit.

Supp. Table 5: Branch length estimates after correction for recurrent mutation

		Branch lengths			Corrected lengths	
		HCGOM shotgun	HCGO shotgun	HCGM shotgun	EM	Modeling
Raw counts of each type of divergent site	n_H	28,504	29,376	28,916		
	n_C	28,495	29,484	28,892		
	n_G	38,677	40,024	39,441		
	n_{HC}	8,561	9,325	9,908		
	n_{HG}	1,302	1,699	2,291		
	n_{CG}	1,430	1,842	2,302		
	n_{HCG}	41,928				
	n_O	82,670	124,598			
	n_M	244,270		286,198		
	n_{HO}	412				
	n_{CO}	397				
	n_{GO}	764				
	n_{HCO}	1,347				
	n_{HGO}	989				
	n_{CGO}	872				
Estimated branch-length as % of H+C+G+C+HG+CG	H	26.6%	26.3%	25.9%	26.9%	27.6%
	C	26.6%	26.4%	25.9%	27.0%	27.8%
	G	36.2%	35.8%	35.3%	36.6%	34.6%
	HC	8.0%	8.3%	8.9%	7.7%	8.1%
	HG	1.22%	1.52%	2.05%	0.82%	0.97%
	CG	1.34%	1.65%	2.06%	0.95%	0.96%
Inferred proportion of sites that are recurrent *	%H	1.2%	4.1%	2.6%		
	%C	1.2%	4.5%	2.5%		
	%G	1.2%	4.5%	3.1%		
	%HC	6.7%	14%	19%		
	%HG	34%	49%	62%		
	%CG	31%	46%	57%		

Note: The first three columns are all based on the same data set (HCGOM shotgun data; Table 1). To obtain HCGO and HCGM data, we just remove one of the species from the HCGOM alignment. This increases the proportion of recurrent mutations: for example, HGO events, which are not consistent with a single historical mutation, appear to be HG events in an HCGM alignment.

* To estimate recurrent mutation rates for each event class in the HCGOM data, we use the EM analysis (Supp. Note 3). For the HCGM and HCGO data, we extrapolate from the EM analysis on the HCGOM data. For example to estimate the HG recurrent mutation rate in an HCGM alignment, we note that in the HCGOM data these would appear to be either HG or HGO events. In the HCGOM data 34% of the 1,340 HG events in the HCGOM data are estimated to be recurrent, and 100% of the 989 HGO events are estimated to be recurrent; thus, the extrapolated recurrent mutation rate is 62% = ((0.34)(1340)+(1.00)(989))/(1340+989).

Supp. Table 6: Divergence in subsets of the genome compared to autosome average

		HCGOM shotgun	HCGM shotgun	HCOM shotgun *	Contiguous *
	Autosomes	8,899,720 bp	17,552,410 bp	17,525,320 bp	1,064,457 bp
	X chrom.	372,354 bp	747,260 bp	15,160,801 bp	112,785 bp
Human-chimpanzee					
$\tau^{HC}_{\text{near HC sites}}$		0.862 ± 0.009	0.882 ± 0.006		0.84 ± 0.03
$\tau^{HC}_{\text{X-chromosome}}$		0.836 ± 0.022	0.835 ± 0.016	0.834 ± 0.006	0.77 ± 0.05
$\tau^{HC}_{\text{X-chromosome}}$ (only human side of lineage)		0.841 ± 0.030	0.851 ± 0.022	0.844 ± 0.008	0.88 ± 0.04
Human-gorilla					
$\tau^{HG}_{\text{X-chromosome}}$ (only human side of lineage)		0.977 ± 0.029	0.980 ± 0.020		1.02 ± 0.05

Note: Estimates of divergence are uncorrected for recurrent mutation, but change by <0.003 after using the EM (Supp. Note 3) correction. The bottom two rows calculate divergence only on the human side of the lineage. This ensures that the reduction in human-chimpanzee divergence is not due to an artifact like mutation rate slow-down on chromosome X, as this would reduce the human-gorilla divergence as well.

* Chromosome 7 is used to represent the autosomes for the HCOM and contiguous data.

Supplementary Table 7: Synthesis of genetic and fossil records

	Human-chimpanzee speciation $\tau^{HC}_{species}$	Human-chimpanzee genome divergence τ^{HC}_{genome}	* Human-orangutan genome divergence τ^{HO}_{genome}	* Human-macaque genome divergence τ^{HM}_{genome}
Using 6.5 Mya (*Sahelanthropus*) as a minimum date for human-chimpanzee speciation	> 6.5	> 7.8	> 20.7	> 36.0
Using 18 Mya (*Proconsul*) as a maximum date for human-orangutan speciation (assuming genome divergence was <2 My earlier)	< 6.3 †	< 7.5	< 20.0	< 34.8
Using 33 Mya (*Aegyptopithecus*) as a maximum date for human-macaque speciation (assuming genome divergence was <2 My earlier)	< 6.3 †	< 7.6	< 20.1	< 35.0

Notes: All dates are in millions of years ago. To convert between species and genome divergence we use $\tau^{HC}_{species}/\tau^{HC}_{genome} < 0.835$ (Supp. Table 6). By making simplifying assumptions it is also possible to produce estimated ranges of speciation times (see Supp. Table 12).

* To convert between genetic divergences of different primates we use $\tau^{HO}_{genome}/\tau^{HC}_{genome} = 2.662$ and $\tau^{HM}_{genome}/\tau^{HC}_{genome} = 4.63$ (Supp. Table 8).

† A more realistic upper bound of <17 Mya for human-orangutan genome divergence results in $\tau^{HC}_{species} < 5.4$ Mya.

Supp. Table 8: Estimates of relative genomic divergences of the primates

(Update on January 23 2008: In the original version of the Supplementary Materials, we used the same image for both Supp. Tables 8-9. After noticing this problem in January 2008, we prepared a version with the correct version of Supp. Table 8.)

		Autosomes					X chromosome	
		HCGOM	**HCOM**	**HCMN**	**Contiguous data***		**HCGOM**	**HCOM**
		8,899,720 bp	17,525,320 bp	133,691 bp	All data	CpG dinculeotides	372,354 bp	15,160,801 bp
Genome divergence ratios	$\frac{\tau^{HG}_{genome}}{\tau^{HC}_{genome}}$	**1.248 ± 0.005** 1.344 ± 0.011			1.25 ± 0.04 1.26 ± 0.04	1.18 ± 0.10 1.18 ± 0.10	1.448 ± 0.031 1.499 ± 0.069	
	$\frac{\tau^{HO}_{genome}}{\tau^{HC}_{genome}}$	2.628 ± 0.014 2.765 ± 0.021	**2.662 ± 0.015** 2.705 ± 0.014		2.68 ± 0.08 2.62 ± 0.08	2.66 ± 0.19 2.24 ± 0.17	2.927 ± 0.083 3.218 ± 0.124	2.958 ± 0.019 3.080 ± 0.017
	$\frac{\tau^{HM}_{genome}}{\tau^{HC}_{genome}}$			4.79 ± 0.22 5.60 ± 0.22	**4.63 ± 0.13** 5.59 ± 0.15	4.86 ± 0.34 4.66 ± 0.33		

Note 1: The relative genome divergence of human and each primate (G=gorilla, M=macaque, N=new world monkey) versus human-chimpanzee. For each calculation there are two halves to the genealogical tree and we take advantage of this to perform a "rate test". On the top of each cell we present a calculation based only on the number of mutations accumulated on the human lineage since divergence. On the bottom we present a calculation based on the lineage of the other species in the comparison. The rates estimated from both sides should be equal or nearly so if mutation rate has been a good "molecular clock". Bold font indicates estimates used for calculations in this study (chosen because they use the largest data set size).

Note 2: All calculations use the EM analysis to correct for recurrent mutation. The ratio of human-gorilla to human-chimpanzee divergence is nearly unchanged if the calculation is redone without the correction. The ratio of human-orangutan to human-chimpanzee divergence changes by the largest factor: it increases by 1.9%: to 2.714, which is conservative for our analyses.

* For the contiguous data, we quote the inferences from Hwang and Green (2004). (Very similar results are obtained by application of the EM analysis to the same data.) We also quote the results from Hwang and Green specifically for CpG dinucleotides, since these are known to mutate at a more constant rate over the mammalian tree (Hwang and Green 2004), and hence can be used to estimate time divergences over the human-macaque split.

Supplementary Table 9: Mutations in the great apes adhere to a near-constant molecular clock

Node of genealogy being assessed	**HCGOM shotgun** maximum assymmetry	**Contiguous data (Hwang and Green)** maximum assymmetry
(H)(C)	1.007 ± 0.009	1.030 ± 0.037
(HC)(G)	1.087 ± 0.008	1.041 ± 0.030
(HCG)(O)	1.061 ± 0.006	1.012 ± 0.023
(HCGO)(M)	NA	1.249 ± 0.023

Note 1: To carry out a rate test for each node in the genealogy, we look for the greatest assymmetry seen among the species on either side of that note. For example, in the HCGOM shotgun data for the orangutan node, the largest assymmetry is that the human lineage is 1.061 times less diverged than the orangutan.

Note 2: These results strongly suggest that the molecular clock has been approximately constant since the divergence of the apes, but has been faster on the old world monkey lineage since ape-old world monkey divergence, consistent with previous observations (e.g. Steiper et al. 2005).

Supplementary Table 10: Constraints on the X:autosome ratio "*R*" in the population ancestral to humans and chimpanzees

Average genomic divergence of human and orangutan (Mya)

Human and chimpanzee speciation date (Mya)	14	14.5	15	15.5	16	16.5	17	17.5	18	18.5	19	19.5	20
4.3	0.10	0.22	0.30	0.37	0.42	0.46	0.49	0.52	0.55	0.57	0.58	0.60	0.61
4.4		0.14	0.25	0.32	0.38	0.43	0.47	0.50	0.53	0.55	0.57	0.59	0.60
4.5		0.05	0.18	0.27	0.34	0.40	0.44	0.48	0.51	0.53	0.55	0.57	0.59
4.6			0.10	0.21	0.30	0.36	0.41	0.45	0.48	0.51	0.54	0.56	0.57
4.7			0.01	0.14	0.24	0.32	0.38	0.42	0.46	0.49	0.52	0.54	0.56
4.8				0.06	0.18	0.27	0.34	0.39	0.43	0.47	0.50	0.52	0.54
4.9					0.11	0.21	0.29	0.35	0.40	0.44	0.47	0.50	0.53
5					0.02	0.15	0.24	0.31	0.37	0.41	0.45	0.48	0.51
5.1						0.07	0.18	0.26	0.33	0.38	0.42	0.46	0.49
5.2							0.11	0.21	0.29	0.34	0.39	0.43	0.46
5.3							0.03	0.15	0.24	0.30	0.36	0.40	0.44
5.4								0.08	0.18	0.26	0.32	0.37	0.41
5.5									0.12	0.21	0.28	0.34	0.38
5.6									0.04	0.15	0.23	0.30	0.35
5.7										0.08	0.18	0.26	0.32
5.8										0.00	0.12	0.21	0.28
5.9											0.05	0.15	0.23
6												0.09	0.18
6.1												0.01	0.12
6.2													0.06
6.3													

$$R = \frac{0.835 - 2.662\,\tau^{HC}{}_{species} \,/\, \tau^{HO}{}_{genome}}{1 - 2.662\,\tau^{HC}{}_{species} \,/\, \tau^{HO}{}_{genome}}$$

Notes: Constraints on the X:autosome ratio in the population ancestral to humans and chimpanzees, using the equation in Supp. Note 10 (reproduced above). If human-chimpanzee speciation occurred earlier than the *Orrorin* and *Ardipithecus* fossils (>5.8 Mya), and if human-orangutan genomic divergence occurred <20 Mya, the X:autosome ratio of the time since the common ancestor at time $\tau^{HC}_{species}$ must have been <0.29. Very similar results obtained by calibration to the macaque fossil record (not shown).

Supplementary Table 11: Parameters used in simulations of ape divergence

Parameter	Modeled	Observed data (target for model fitting)	Simulated results
$\tau^{HC}_{species}$	5.2 Mya		
$\tau^{HG}_{species}$	6.95 Mya		
$\tau^{HO}_{species}$	18.3 Mya		
$\tau^{HM}_{species}$	32.5 Mya		
mean and standard deviation for human-chimpanzee ancestral divergence	1.8 Mya ± 1.6 Mya		
pop. size ancestral to gorilla divergence	50,000		
pop. size ancestral to orangutan divergence	20,000		
pop. size ancestral to macaque divergence	20,000		
Mut. rate (per generation) ± 1 standard dev.	$1.17 \times 10^{-8} \pm 0.25 \times 10^{-8}$		
Size of clusters (bases) ± standard deviation	300 ± 133		
Number of clusters simulated	50,000		
$\tau^{HM}_{genome}/\tau^{HC}_{genome}$ *		4.63	4.60
$\tau^{HO}_{genome}/\tau^{HC}_{genome}$		2.70	2.69
$\tau^{HG}_{genome}/\tau^{HC}_{genome}$		1.25	1.24
$\frac{HG + CG}{H + C + HG + CG}$		0.046	0.044
Human-chimp divergence near 1 HC site		0.86	0.87
Human-chimp divergence near an HG or CG		1.34	1.34
Proportion of sites that cannot have been due to a single historical mutation (e.g. HO)		0.0100	0.0099

Note: We were not able to fit the data assuming a constant sized population ancestral to humans and chimpanzees. Instead, we modeled the time since the common ancestor of humans and chimpanzees as 5.2 Mya plus a random draw from a gamma function with mean = 1.8 Mya and standard deviation = 1.6 Mya.

* For $\tau^{HO}_{genome}/\tau^{HC}_{genome}$ and $\tau^{HG}_{genome}/\tau^{HC}_{genome}$ we use estimates of relative divergence on the human side of the lineage (uncorrected for recurrent mutation; HCGOM shotgun data). For $\tau^{HM}_{genome}/\tau^{HC}_{genome}$ we use the value 4.63 from Hwang and Green (2004) (Supp. Table 8). While the human-macaque relative divergence estimate is tentative because of unreliability of the molecular clock in the human-macaque comparison (Supp. Tables 8,9), errors in this ratio do not affect our key inferences about the history of the African apes.

Supp. Table 12: Speciation time estimates from a simplified model of demographic history

	Human-chimpanzee speciation $\tau^{HC}_{species}$	Human-gorilla speciation $\tau^{HG}_{species}$	Human-orangutan speciation $\tau^{HO}_{species}$	Human-macaque speciation $\tau^{HM}_{species}$
Calibration to human-chimp divergence Using 6.5 Mya (*Sahelanthropus*) as a minimum date for human-chimp speciation and 10 Mya as a maximum date (*arbitrary upper bound as there is no good fossil information)*	6.5 - 10	7.3 - 19	20 - 45	36 - 79
Calibration to human-orangutan divergence Using 13 Mya (*Sivapithecus*) as a minimum date for human-macaque speciation and 18 Mya (*Proconsul*) as a maximum date	2.9 - 6.3	3.4 - 11.3	13 - 18	21 - 34
Calibration to human-macaque divergence Using 21 Mya (Morotopithecus) as a minimum date for human-macaque speciation and 33 Mya (*Aegyptopithecus*) as a maximum date	2.7 - 6.3	3.0 - 12.0	10 - 20	21 - 33

Notes: All estimates are in millions of years ago. Any estimate of speciation times makes simplifying modeling assumptions about demographic history; these estimates must be interpreted cautiously as we have shown that the speciation process for human and chimpanzee ancestors was complex.

Model used to relate τ_{genome} and $\tau_{species}$ for African apes. To obtain best estimates for the speciation times, it is necessary to deploy a simplifying model. We use a 4-parameter model (Supp. Note 2) in which humans and chimpanzees split from a freely mixing ancestral population of size N_{HC} at time $\tau^{HC}_{species}$, and from a freely mixing population of size N_{HG} (also ancestral to gorillas) at time $t^{HG}_{species}$. With branch length estimates from our HCGOM shotgun data, we then infer $\tau^{HG}_{species}/\tau^{HC}_{species}$ = 1.13 - 1.90 and $\tau^{HC}_{genome}/\tau^{HC}_{species}$ = 1.20 - 1.75 (Supp. Note 2).

Population ancestral to orangutan and macaque divergence. We assume the average time since the common ancestor at the time of speciation was $\tau^{HM}_{genome}-\tau^{HM}_{species} = \tau^{HO}_{genome}-\tau^{HO}_{species}$ = 0.5 - 2 MY. (The average time since the common genetic ancestor in modern apes is between ~0.5 MY (humans, some chimpanzees), and ~1 million years (gorillas, some chimpanzees) (Yu et al. 2004).) To convert between genetic divergence of humans and chimpanzees, and these more distantly related primates, we use $\tau^{HO}_{genome}/\tau^{HC}_{genome}$ = 2.662 and $\tau^{HM}_{genome}/\tau^{HC}_{genome}$ = 4.63 (Supp. Table 8).

Remarks: Estimates of orangutan (20-45 Mya) and macaque (36-79 Mya) speciation based on calibration to the fossil record of human-chimpanzee divergence (first row) clonflict with the older fossil record. This supports the conclusion that human-chimpanzee gene flow (and thus true $\tau^{HC}_{species}$ speciation time) occurred <6.3 Mya. Dates of human-chimpanzee speciation of <4.3 Mya are also likely to be incompatible with the fossil record.

Supplementary Figure 1: Stability of statistics after filtering less reliable alignments

a **HG and CG site rate (scaled)**

Filter high polymorphism \ Filter out tree assymmetry	$P < 10^{-1}$	$P < 10^{-2}$	$P < 10^{-3}$	$P < 10^{-4}$	$P < 10^{-5}$	$P < 10^{-6}$	no filter
$P < 10^{-1}$	7.2	7.8	7.8	7.8	7.8	7.8	7.8
$P < 10^{-2}$	7.3	7.9	7.9	7.9	7.9	7.9	7.9
$P < 10^{-3}$	7.4	8.0	8.1	8.1	8.1	8.1	8.1
$P < 10^{-4}$	7.5	8.1	8.1	8.1	8.1	8.1	8.1
$P < 10^{-5}$	7.5	8.1	8.1	8.1	8.1	8.1	8.1
$P < 10^{-6}$	7.5	8.1	8.1	8.1	8.1	8.1	8.1
no filter	7.8	8.4	8.5	8.5	8.5	8.5	8.5

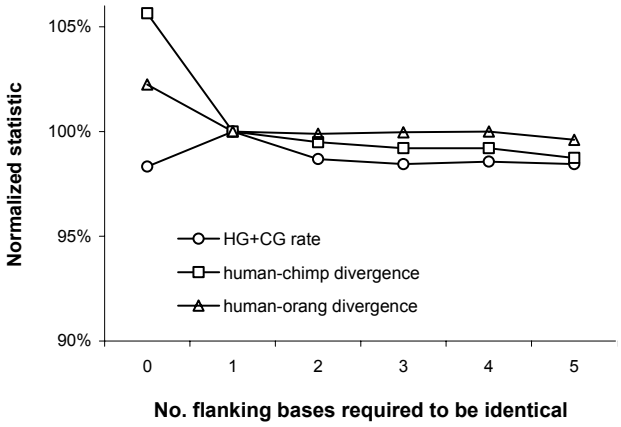
human-gorilla / human-chimpanzee divergence

Filter high polymorphism \ Filter out tree assymmetry	$P < 10^{-1}$	$P < 10^{-2}$	$P < 10^{-3}$	$P < 10^{-4}$	$P < 10^{-5}$	$P < 10^{-6}$	no filter
$P < 10^{-1}$	1.26	1.25	1.25	1.25	1.25	1.25	1.25
$P < 10^{-2}$	1.26	1.25	1.25	1.25	1.25	1.25	1.25
$P < 10^{-3}$	1.26	1.25	1.25	1.25	1.25	1.25	1.25
$P < 10^{-4}$	1.26	1.25	1.25	1.25	1.25	1.25	1.25
$P < 10^{-5}$	1.26	1.25	1.25	1.25	1.25	1.25	1.25
$P < 10^{-6}$	1.26	1.25	1.25	1.25	1.25	1.25	1.25
no filter	1.26	1.25	1.25	1.25	1.25	1.25	1.25

human-chimpanzee / human-macaque divergence

Filter high polymorphism \ Filter out tree assymmetry	$P < 10^{-1}$	$P < 10^{-2}$	$P < 10^{-3}$	$P < 10^{-4}$	$P < 10^{-5}$	$P < 10^{-6}$	no filter
$P < 10^{-1}$	0.164	0.17	0.171	0.171	0.171	0.171	0.171
$P < 10^{-2}$	0.164	0.171	0.172	0.172	0.172	0.172	0.172
$P < 10^{-3}$	0.165	0.171	0.172	0.172	0.172	0.172	0.172
$P < 10^{-4}$	0.165	0.171	0.172	0.172	0.172	0.172	0.172
$P < 10^{-5}$	0.165	0.171	0.172	0.172	0.172	0.172	0.172
$P < 10^{-6}$	0.165	0.171	0.172	0.172	0.172	0.172	0.172
no filter	0.165	0.172	0.173	0.173	0.173	0.173	0.173

b



We examined the behavior of three different statistics—the HG and CG rate, the human-gorilla/human-chimpanzee divergence ratio, and the human-chimpanzee/human-macaque divergence ratio—to explore the stability of our inferences. (a) These statistics achieved stable values if we eliminated alignments with high intraspecific polymorphism (P>0.001) and asymmetry in the genealogy (P>0.0001) (Supp. Methods), and thus we chose these thresholds for including alignments in analysis. (b) Stable values of the same three statistics are achieved when we required at least 1 flanking base to match on either side, and so we use this threshold for filtering out less reliable divergent sites from our data.

Supplementary Note 1
Definitions of speciation time and hybridization

Speciation Time:

We define the time since two modern populations speciated ($\tau_{species}$) as the amount of time that has passed since their ancestors last exchanged genes.

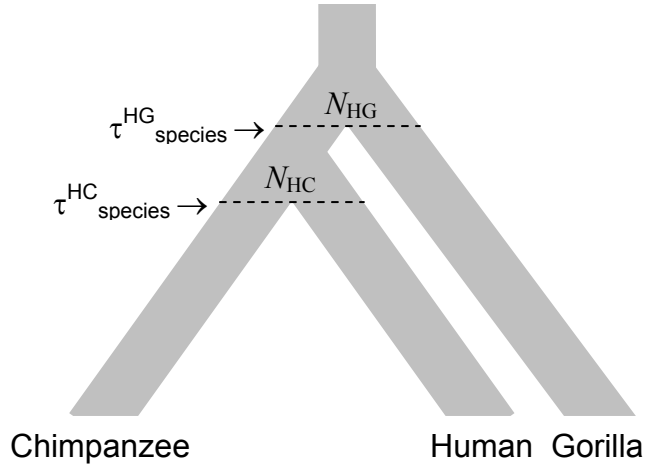
Hybridization:

We define hybridization as gene flow between two populations after an isolation barrier has formed between them.

Supplementary Note 2

Percent of genome in which humans and chimpanzees are not most closely related

We estimate the fraction p of the genome in which humans and chimpanzees are not most closely related (Fig. 1c,d). To infer this we need to assume a particular model for the demographic structure of the ancestral populations of humans, chimpanzees, and gorillas. While for most inferences in the paper, we do not use modeling assumptions for the structure of the ancestral population, to obtain a minimum estimate of p we made assumptions. For this analysis we use a model with four free parameters:



The model assumes that the human and chimpanzee lineages split from a freely mixing ancestral population of size N_{HC} at time $\tau^{HC}_{species}$, and from a population of size N_{HG} (also ancestral to gorillas) at time $\tau^{HG}_{species}$. Under this model straightforward derivations produce 3 equations:

$$t_H + t_{HG} = \tau^{HC}_{genome} = \tau^{HC}_{species} + 2N_{HC} + (2N_{HG} - 2N_{HC})e^{-\left(\tau^{HG}_{genome} - \tau^{HC}_{genome}\right)/2N_{HC}}$$

$$t_H + t_{HC} = \tau^{HG}_{genome} = \tau^{HG}_{species} + 2N_{HG}$$

$$t_{HG} = \left(2N_{HG}/3\right)e^{-\left(\tau^{HG}_{genome} - \tau^{HC}_{genome}\right)/2N_{HC}}$$

We solve these equations under constraints from the data:

(a) We use the relative values of t_H:t_{HC}:t_{HG} from the EM analysis on the HCGOM shotgun data (Supp. Table 5 gives the best estimates)

(b) We force the $\tau^{HC}_{species}/\tau^{HC}_{genome}$ ratio to be <0.835, based on the upper bound on human-chimpanzee speciation time from the X chromosome presented in Supp. Table 6.

(c) We force $\tau^{HC}_{species}/\tau^{HC}_{genome}$ to be >0.57, based on the constraints $\tau^{HC}_{species}$ >4.3 Mya (assuming that human-chimpanzee speciation had occurred by the time of *Australopithecus*) and τ^{HC}_{genome} < 7.6 Mya (from calibrations described in Supp. Note 9). Thus, $\tau^{HC}_{species}/\tau^{HC}_{genome} > 4.3/7.6 = 0.57$.

We do not have enough data to provide a unique solution to the equations above (there are four unknowns in the model, $\tau^{HC}_{species}$, $\tau^{HG}_{species}$, N_{HG} and N_{HC}, but only three independent measurements, t_H, t_{HC} and t_{HG}). We therefore take the approach of presenting *allowed ranges*, solving the equations for values of $\tau^{HC}_{species}/\tau^{HC}_{genome}$ over the allowed range of 0.57-0.835.

We also need to take into account uncertainty in our estimates of parameters due to limited data set size. To obtain 95% credible intervals for each quantity, we carried out 1,000 bootstrap replications (over the unlinked genomic segments in the HCGOM data; Supplementary Methods), randomly sampling the segments (with replacement), and thus obtaining 1,000 data set of the same size as our actual data.

In the summary table below, we present combinations of model parameters of high interest. We present the values of these parameters that are consistent with $0.57 < \tau^{HC}_{species}/\tau^{HC}_{genome} < 0.835$, and the 95% credible intervals obtained by the bootstrap analysis.

	Lower bound		Upper bound
$\tau^{HC}_{species}/\tau^{HC}_{genome}$	0.57	to	0.835
$\tau^{HG}_{species}/\tau^{HG}_{genome}$	0.728	to	0.834
$\tau^{HG}_{species}/\tau^{HC}_{species}$	1.13	to	1.90
N_{HG}/N_{HC}	0.37	to	3.96
% of genome in which humans and chimpanzees not most closely related	18%	to	29%

The analysis shows that the data can not be explained unless humans and gorillas, or chimpanzees and gorillas, are most closely related in at least 18% of the genome. One caveat is that the models we explored are limited; with only one size change allowed in the ancestral population. In Supp. Note 11, we explore a broader range of models.

Supplementary Note 3

Branch length estimates in the presence of recurrent mutations using the EM algorithm

Overview:

We implemented the Expectation-Maximization algorithm (EM) to obtain branch length estimates corrected for the presence of recurrent mutations. Briefly, the algorithm makes the simplifying assumptions that mutation rates have been historically constant at each locus, and that branch lengths do not vary across the genome. For example in the HCGOM data the algorithm searches for the combination of branch lengths (in this case $t_H, t_C, t_G, t_O, t_M, t_{HC}, t_{HG}, t_{CG}$ and t_{HCG}, and relative probabilities of recurrent mutations (probability that sites are due to a history of $1, 2$ or 3 mutations) that maximizes the likelihood of the observed data (including the rates of divergent sites not consistent with a single historical mutation).
One parameter that emerges from the analysis is an estimate, based on the data, of the proportion of divergent sites that are due to recurrent mutation. The fact that some sites are more mutable than others, and that this can sometimes be predicted by DNA sequence context, does not need to be explicitly addressed in our approach as hypermutable sites simply increase the estimated proportion of the total sites due to recurrent mutation. This proportion can then be corrected for in our analysis, without explicitly addressing which particular sites have recurrent mutations. Hypermutable sites thus do not introduce bias into EM estimates of branch length.

Details of the algorithm:

Writing this more generally, so that we can apply the methods to alignments such as HCOM, HCGOM or HCGOMN, we have species $S_1, S_2, \ldots S_n$ and we are given polymorphism data (aligned sequence) and look at all positions where there are no more that 2 alleles (biallelic data).

We have branch lengths: $(l_1, l_2, \ldots, l_k)$ which are the expected numbers of mutations of each category averaged base by base across the genome. We wish to estimate the branch lengths l_i from the data. Our approach takes account of the fact that certain sites are hyper-mutable and thus that the probabilities of mutations vary greatly.

Normalizing the branch lengths so that they sum to 1:

$$\sum_{s=1}^{k} l_s = 1 \tag{1}$$

Consider mutation events ('hits') on an edge that change the allele value. We will consider all *hit patterns* $h(i) = (C(i,1), C(i,2) \ldots, C(i,k))$ such that the sum

$$w(i) = \sum_{s=1}^{k} C(i,s) \leq L$$

where L is some small bound (in the paper we use 3). That is we ignore the possibility of more than 3 mutations at a site.

Introduce probabilities $x_0, x_1, x_2, \ldots x_L$, where x_k is the probability of having k mutations at a site. Suppose $w(i) = S$. We want to define $P(i)$, the probability

of hit pattern i, given that $w(i) = S$. The following is simple though complicated to express. Consider the polynomial $F(u_1, u_2, \ldots u_k)$ where

$$F = F(u_1, u_2, \ldots, u_k) = (l_1 u_1 + \ldots + l_k u_k)^S$$

Then $F(1, 1, \ldots, 1) = 1$. $P(i)$ is the coefficient of $\prod_{s=1}^{L} u_s^{C(i,s)}$ in F.
If we define

$$Z(i) = \frac{P(i)}{\prod_{s=1}^{L} l_s^{C(i,s)}} \tag{2}$$

then $Z(i)$ is a non-negative integer independent of $l_1, \ldots l_k$. It is in fact a multinomial coefficient.

We observe difference patterns $D_1, \ldots, D_J$. For technical reasons we need to allow a difference pattern of 0 (no difference) which is not really an observable in a useful way. We simply input this externally as a large count, our results being insensitive to the value used. In an obvious manner hit patterns map to difference vectors. That is there is a fixed map M with $M(i) = j$. Set $S(j)$ to be the set of pre-images of j so that

$$S(j) = \{i | M(i) = j\}$$

Now we take the probability of a difference pattern j to be

$$R(j) = \sum_{i \in S(j)} x_{w(i)} P(i) \tag{3}$$

since $\sum_{t=0}^{k} x_t = 1$ it is easy to check that

$$\sum_j R(j) = 1$$

If pattern j is observed $N(j)$ times then the log-likelihood $\mathcal{L}$ is just

$$\mathcal{L} = \sum_j N(j) \log R(j)$$

We will maximize this subject to the constraints:

$$\sum_s l_s = 1 \tag{4}$$

$$\sum_t x_t = 1 \tag{5}$$

$e^{\mathcal{L}}$ is a polynomial with positive coefficients and so EM theory applies. Here are the details. We first need to define the expected number of times hit-vector i was 'used'. Call this $X(i)$ Then it is easy to see that:

$$X(i) = \frac{x_{w(i)} P(i) N(j)}{R(j)}$$

where $M(i) = j$. There is an auxiliary function of *new* variables $\hat{l}_s$, $\hat{x}_t$:

$$\mathcal{Q}(\hat{l}_1, \ldots, \hat{x}_1, \ldots \hat{x}_k) = \sum_i X(i) \log \left(\hat{x}_{w(i)} \hat{P}(i) \right) \tag{6}$$

Baum theory (Baum et al. 1970) shows that

$$\mathcal{Q}(\hat{\mathbf{l}}, \hat{\mathbf{x}}) - \mathcal{Q}(\mathbf{l}, \mathbf{x}) \leq \mathcal{L}(\hat{\mathbf{l}}, \hat{\mathbf{x}}) - \mathcal{L}(\mathbf{l}, \mathbf{x}) \tag{7}$$

Thus increasing $\mathcal{Q}$ will increase $\mathcal{L}$. Maximizing $\mathcal{Q}$ is easy. Set

$$\Gamma(s) = \sum_i X(i)C(s,i) \tag{8}$$

and similarly for the weights:

$$G(t) = \sum_{i:w(i)=t} X(i) \tag{9}$$

Now we simply set our reestimates to be:

$$\hat{l_s} = \frac{\Gamma(s)}{\sum_s \Gamma(s)} \tag{10}$$

$$\hat{x_t} = \frac{G(t)}{\sum_t G(t)} \tag{11}$$

The reestimations can be done jointly as the variables $\mathbf{l}$, $\mathbf{x}$ separate in equation (6).

Supplementary Note 4

Human-chimpanzee divergence at short distances from HC events, and HG or CG events

We start with aligned sequence of humans (H), chimpanzees (C), gorillas (G), macaques (M), and sometimes orangutans (O). Our goal is to ask whether, near regions of human-gorilla (HG), chimpanzee-gorilla (CG), or human-chimpanzee (HC) clustering, the genetic divergence between humans and chimpanzees is unusually high or low.

As described in the main paper, we classify all divergent sites (events) into classes, so for example in an HCGOM alignment an *HG-event* means that H and G were one base, and C, O, M were another. We are interested in computing divergence ratios 'close' to events in a given class. For instance we might be interested in human-chimpanzee divergence near HG or CG events. For a given analysis we divide events into classes:

A The primary event class: Here HG and CG.

B Events contributing to the numerator of the divergence. Here we will measure human-chimpanzee divergence by counting $H + C + HG + CG$ events.

C Events contributing to only the denominator of the divergence. Here we will measure human-macaque divergence by counting $H + HC + HG + HCG + M$ events.

D Other event types ignored for this analysis.

Note that the classes can overlap. Our main interest is to estimate the mean divergence rate (ratio of B to C) very near events in class A. For each distance d (in bases) we simply count the number $b(d)$ of events in class B at a distance d from a primary event in class A. We make a similar count $c(d)$ for events in class C. Set:

$$\begin{aligned} n(d) &= b(d) + c(d) \\ r(d) &= \frac{b(d)}{n(d)} \ (r(d) = 0 \ if \ n(d) = 0) \end{aligned}$$

We then pick some sensible limit L (we chose $L = 5000$) for our analysis. We are interested in behavior close to a primary event such as HG, and we do not want our inference to critically depend on events far away, and so we ignore events more than $5,000$ bases away. We will define a simple parametric model for $s(d)$, the expected value of $r(d)$, the normalized human-chimpanzee divergence near these events. We chose:

$$\begin{aligned} s(d) &= f(\alpha, \beta, \lambda) \\ &= \alpha e^{-\lambda d} + \beta(1 - e^{-\lambda d}) \end{aligned} \tag{12}$$

Now set

$$\mathcal{S} = \sum_{d=1}^{L} n(d)(r(d) - s(d))^2 \tag{13}$$

We will minimize the quadratic function $\mathcal{S}$ as a function of the parameters α, β, λ. Let $\hat{X} = \hat{\alpha}/\hat{\beta}$.

We are interested in estimating the ratio of the class B event rate to the class C event rate, near primary events. This is given by

$$\hat{R} = \frac{\hat{X}}{1 - \hat{X}} \tag{14}$$

$\mathcal{S}$ is *not* a log-likelihood, but $\hat{R}$ is an asymptotically consistent estimator. β is simply estimated as the overall ratio of the number of class B events to the total of class B and class C events. The standard error of β is very small, and so we ignore it, minimizing $\mathcal{S}$ with β fixed.

It is important to estimate the standard error of our estimate of $\hat{R}$. Our observations are far from independent, so some care is needed. We group our shotgun reads into *segments* as described in the Supplementary Methods, ensuring that any two segments are separated by at least $10,000$ bases. We regard each segment as independent, and apply the weighted jackknife (Busing et al. 1999). We choose the weight of a segment to be the total contribution of the segment to the counts $b(d)$ and $c(d)$ summed over all distances d from 1 to the upper limit L.

It is also of interest to consider the divergence ratio near a *double event* such as two HG or CG events very close together, defined as within 200 bases. Such double events are more likely to be real regions of HG or CG clustering, arising from a single historical mutation rather than a recurrent mutation. We declare a primary 'event' to have occurred at the mid-point of the two events being considered. There is a question of how to deal with the two sites that contributed to the double event. Should they be counted as having occurred near the double event? After some experimentation we decided

1. No site is counted as near a double event for which it was one of the two events close together.

2. No site contributes to more than one double event.

This last rule is to prevent clusters of HG events causing an excessive contribution to $\mathcal{S}$. The choices we have made are conservative, underestimating the spread of divergence times.

We first give the results for single primary HG or CG events.

Single Primary Events

Data	$\hat{R}$	std.err
HCGOM	1.3421	0.0216
HCGOM-O	1.2417	0.0170
HCGM	1.2372	0.0113
CFTR	1.2556	0.0519

We have 3 main data sets as described in the main paper.

1. 5-species data (HCGOM).

2. 4-species data (HCGM).

3. 5-species (HCGOM) data from contiguous sequencing around the CFTR-gene.

In addition we also analyze the 5-species data by ignoring O, giving us a 4-species data set (HCGOM-O) with more recurrent mutations than the 5-species

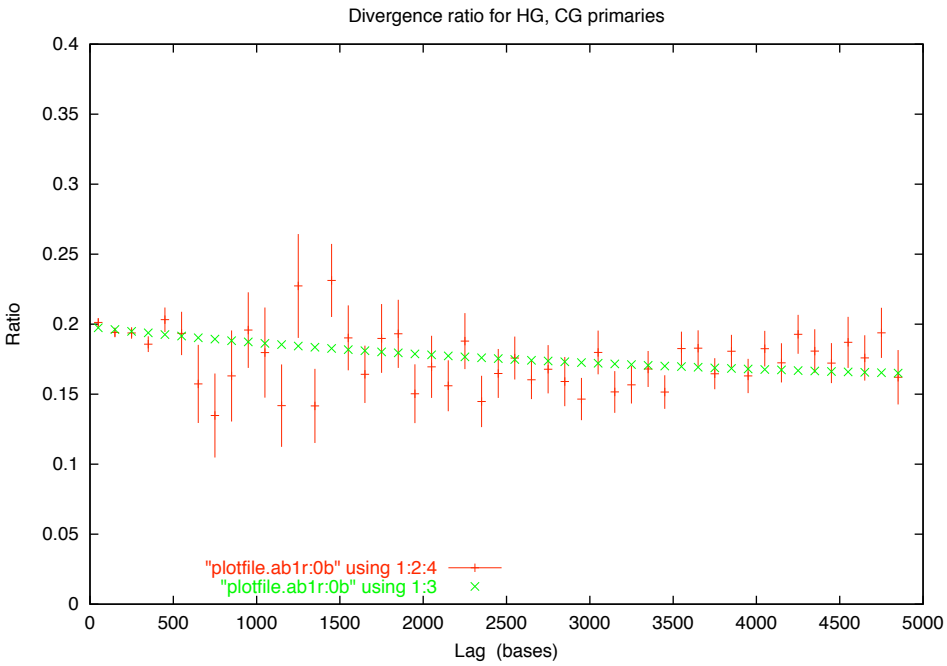


Figure 1: Fit of model for HCGOM data

data. This gives a guide to how the recurrent mutations are affecting the results for the HCGM data.

We see that the 4-species data give a divergence ratio of $1.24 \pm .01$ with the 5-species data $1.34 \pm .02$. The large difference shows the strong effect of recurrent mutations on our data.

In Figure 1 below we show a plot of the fitted s function to r for HG, CG primary events, where we also show the standard error for r at a given lag. The fit seems satisfactory.
Next we show data for the double primary HG, CG events.

Double Primary Events

Data	$\hat{R}$	std.err
HCGOM	1.4536	0.0651
HCGOM-O	1.4753	0.0586
HCGM	1.3528	0.0291
CFTR	1.3480	0.0636

Just as we expect we see a modest increase in the divergence, though the standard errors are now substantial.

It seems reasonable to conclude that near a single-mutation HG event the divergence is at least 1.34 times the genome average, and probably greater than 1.36. Near two HG events the best estimate is around 1.48. These values are all conservative minima, because they do not correct for recurrent mutation which is expected to bring all ratios closer to 1.

Next we give results for single primary HC events.

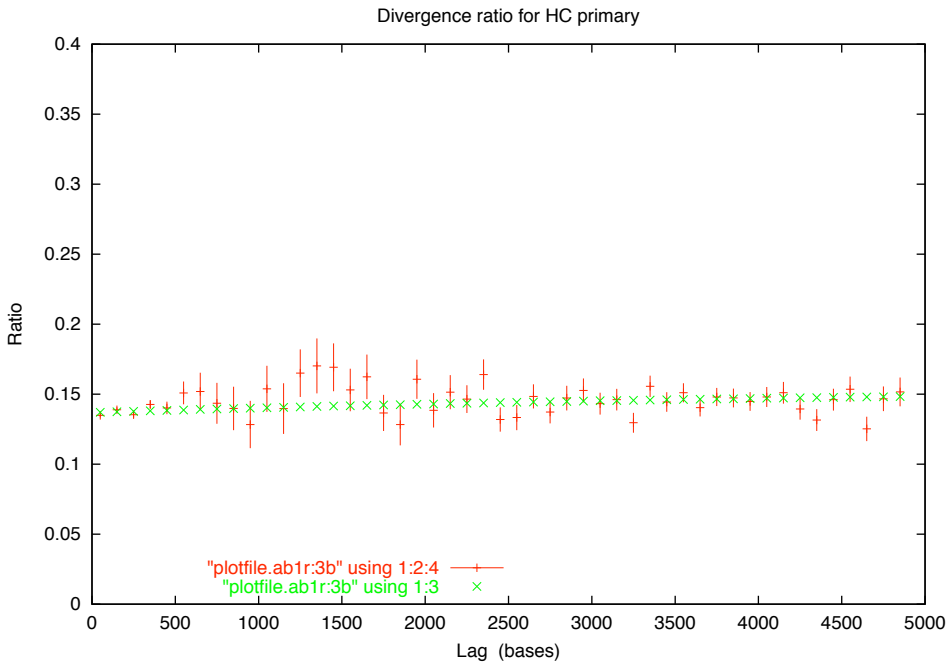


Figure 2: Fit of model for HCGOM data

Single Primary Events

Data	$\hat{R}$	std.err
HCGOM	0.8616	0.0092
HCGOM-O	0.8785	0.0086
HCGM	0.8801	0.0059
CFTR	0.8766	0.0215

Again note the excellent agreement between the $HCGOM - O$ and $HCGM$ data.

Double Primary Events

Data	$\hat{R}$	std.err
HCGOM-O	0.8620	0.0158
HCGOM	0.8653	0.0163
HCGM	0.8881	0.0088
CFTR	0.8852	0.0241

As we expect, the difference between double and single HC events is much smaller than for HG events, as the proportion of recurrent mutations is much lower. We conclude that near an HC event due to a single mutation the divergence is at most 0.88 times the genome average. Figure 2 shows a plot of the least squares fit.

Supplementary Note 5
Expectation for X:autosome divergence ratio

To estimate the X:autosome ratio in the ancestral population, we return to the family of models described in Supp. Note 2. Our strategy is to explore the full range of combinations of $\tau^{HC}_{species}$, $\tau^{HG}_{species}$, N_{HC}, and N_{HG} consistent with the autosomal data under this model. We then extrapolate the X:autosome ratio that should have existed (under this model) at time $\tau^{HC}_{species}$.

To combine chromosome X and autosome data under this model, the main principle we used is that putting together equal numbers of males and females in a population, there are 3/4 the number of X chromosomes as autosomes. Thus, the relevant population parameters on the X should be $\tau^{HC}_{species}$, $\tau^{HG}_{species}$, $0.75N_{HC}$, and $0.75N_{HG}$. From the equations in Supp. Note 2, the X:autosome ratio for the human-chimpanzee comparison should thus be:

$$\frac{\tau^{HC}{}_{X-chrom}}{\tau^{HC}{}_{genome}} = \frac{\tau^{HC}{}_{species} + 1.5N_{HC} + (1.5N_{HG} - 1.5N_{HC})e^{-\left(\tau^{HG}{}_{genome} - \tau^{HC}{}_{genome}\right)/1.5N_{HC}}}{\tau^{HC}{}_{species} + 2N_{HC} + (2N_{HG} - 2N_{HC})e^{-\left(\tau^{HG}{}_{genome} - \tau^{HC}{}_{genome}\right)/2N_{HC}}}.$$

The X:autosome ratio for the human-gorilla comparison should be:

$$\frac{\tau^{HG}{}_{X-chrom}}{\tau^{HG}{}_{genome}} = \frac{\tau^{HG}{}_{species} + 1.5N_{HG}}{\tau^{HG}{}_{species} + 2N_{HG}}.$$

We calculate both these ratios for the full range of values of $\tau^{HC}_{species}$, $\tau^{HG}_{species}$, N_{HC}, and N_{HG} consistent with the autosomal data, using the same data constraints, and methods for obtaining 95% credible intervals from bootstrap analysis, as described in Supp. Note 2.

Results:

The X:autosome time divergence ratio for the human-chimpanzee comparison is expected to be A ~ 0.918-0.943 for the full range of parameters consistent with our data (this range includes an allowance for error due to limited data set size, as described in Supp. Note 2).

The X:autosome ratio for the human-gorilla comparison is expected to be A ~ 0.932-0.958. (This includes an allowance for error due to limited data set size, as described in Supp. Note 2).

Supplementary Note 6

Inferred reduction in the HG+CG rate on the X chromosome

The rate of HG+CG events $(n_{HG}+n_{CG})/(n_H+n_{HC}+n_{HG}+n_{HCG}+n_M)$ is 4.0-fold reduced on the X as compared with the autosomes in the HCGOM shotgun data: 0.00211 ± 0.00041 vs. 0.00841 ± 0.00019 (Table 1). The error bars (obtained by jackknife analysis) show that this reduction is highly significant at 13.9 standard deviations.

To translate this to branch lengths, we note the best estimate for the X chromosome is that there is essentially no HG or CG clustering: $(t_{HG}+t_{CG})/(t_H+t_{HC}+t_{HG}+t_{HCG}+t_M) = 0.00013 \pm 0.00034$, using the branch length corrections from the EM analysis. This is very much less than the 0.00580 ± 0.00020 seen on the autosomes. We were not confident in using the jackknife to place an upper bound on this reduction, however, as the error bars are calculated under the normality assumption (not satisfied here since $(t_{HG}+t_{CG})/(t_H+t_{HC}+t_{HG}+t_{HCG}+t_M)$ is so close to 0 for chromosome X).

We therefore turned to a bootstrap analysis (Efron and Tibshirani 1993) to obtain a 95% credible interval for the reduction of the HG and CG rate on the X chromosome vs. autosomes. We generated 10,000 data sets by sampling (with replacement) from the 7,386 segments that comprised the autosomal data and 286 segments that comprised the X data, obtained as described in the jackknife section of Supp. Methods. For each of the 10,000 resamplings, we carried out the EM analysis separately for the autosomes and X, calculating the ratio $(t_{HG}+t_{CG})/(t_H+t_{HC}+t_{HG}+t_{HCG}+t_M)$ in each, and took the ratio. The 95% confidence interval is 0-15%, with a best estimate of zero.

The reduction in $(t_{HG}+t_{CG})/(t_H+t_{HC}+t_{HG}+t_{HCG}+t_M)$ on chromosome X is more extreme than would be expected from a demographic model that does not invoke natural selection. For example, in the model described in Supp. Note 2, the HG+CG rate is expected to be 42-50% that of the autosomes. Even in more extreme models (Supp. Note 11), we do not expect reductions in the HG+CG rate nearly as extreme as what we see in our data.

Supplementary Note 7

X-autosome difference was an order of magnitude larger in the ancestral population than in humans today

We are interested in the difference between coalescence time on chromosome X and the autosomes in the ancestral population of humans and chimpanzees and in comparing it to that in humans today.

To carry out this analysis we use quantities measured from the HCGOM shotgun data that provide information about time divergence of humans & chimpanzees on the autosomes & chromosome X.

$\tau^{HC}_{genome}/\tau^{HM}_{genome} = 0.1821 \pm 0.0009$
$\tau^{HC}_{X\text{-}chrom}/\tau^{HM}_{X\text{-}chrom} = 0.1527 \pm 0.0040$

The difference between the human and chimpanzee time divergences on chromosome X and the autosomes today, $\tau^{HC}_{genome} - \tau^{HC}_{X\text{-}chrom}$, must be the same as the difference between the two time divergences in the ancestral population at the time $\tau^{HC}_{species}$, and we use this fact in our analysis.

To compare the ancestral human-chimpanzee population to humans today, we need estimates of human heterozygosity on the autosomes and chromosome X. We denote human heterozygosity, that is, the average genetic difference between two individuals, as τ^{HH}_{genome} for the autosomes and $\tau^{HH}_{X\text{-}chrom.}$ for chromosome X. If we can use external data sets to obtain estimates of human heterozygosity normalized by human-chimpanzee divergence ($\tau^{HH}_{genome}/\tau^{HC}_{genome}$ and $\tau^{HH}_{X\text{-}chrom.}/\tau^{HC}_{X\text{-}chrom.}$), we can then write:

$$\frac{X-autosome\ difference\ in\ ancestral\ population}{X-autosome\ diff.\ in\ human\ population\ today} \approx \frac{\left(\frac{\tau^{HC}_{genome}}{\tau^{HM}_{genome}}\right)-\left(\frac{\tau^{HC}_{X-chrom}}{\tau^{HM}_{X-chrom}}\right)}{\left(\frac{\tau^{HH}_{genome}}{\tau^{HM}_{genome}}\right)-\left(\frac{\tau^{HH}_{X-chrom}}{\tau^{HM}_{X-chrom}}\right)} =$$

$$\frac{\tau^{HC}_{genome}/\tau^{HM}_{genome} - \tau^{HC}_{X-chrom}/\tau^{HM}_{X-chrom}}{\left(\tau^{HH}_{genome}/\tau^{HC}_{genome}\right)\left(\tau^{HC}_{genome}/\tau^{HM}_{genome}\right)-\left(\tau^{HH}_{X-chrom}/\tau^{HC}_{X-chrom}\right)\left(\tau^{HC}_{X-chrom}/\tau^{HM}_{X-chrom}\right)}$$

We asked two colleagues to independently provide us with estimates of human-heterozygosity normalized by human-chimpanzee divergence ($\tau^{HH}_{genome}/\tau^{HC}_{genome}$ and $\tau^{HH}_{X\text{-}chrom.}/\tau^{HC}_{X\text{-}chrom.}$):

The first analysis was carried out by Michael Zody. To obtain estimates of human heterozygosity, he compared the published human genome reference sequence (build34) to a shotgun sequencing data set based on African American samples. To obtain estimates of human-chimpanzee divergence he compared the reference genome of humans to that of chimpanzees. The resulting estimates (the same as in Taylor et al. 2006), suggest that $\tau^{HH}_{genome}/\tau^{HC}_{genome} = 0.0801$ and $\tau^{HH}_{X\text{-}chrom}/\tau^{HC}_{X\text{-}chrom.} = 0.0766$ (see table below). All our analyses focus on data sets with CpG dinucleotides removed.

The second analysis was carried out independently by James Mullikin. He compared the same DNA sequencing data sets, but used an independent methodology for identifying polymorphic and

divergent sites. This analysis suggested that $\tau^{HH}_{genome}/\tau^{HC}_{genome} = 0.0716$ and $\tau^{HH}_{X\text{-}chrom}/\tau^{HC}_{X\text{-}chrom.} = 0.0682$. We present results from both analyses, noting that slightly different results are obtained because of the different criteria used for identifying within-human polymorphisms, and human-chimpanzee divergent sites.

Analysis #1 (M. Zody data)	**Autosomes**	**X**
Human-human diversity per base pair	0.000719 ± 0.000001	0.000532 ± 0.000003
human-chimpanzee divergence per base pair	0.008979 ± 0.000001	0.006945 ± 0.000006
Human diversity/divergence ratio	0.08007 ± 0.00011	0.07661 ± 0.00044

Note: This data set was compiled by Michael Zody and used in the Taylor et al. 2005 paper. Human diversity is calculated by comparing an African American shotgun data set to the public genome reference sequence.

Analysis #2 (J. Mullikin data)	**Autosomes**	**X**
Human-human diversity per base pair	0.0007958 ± 0.0000007	0.0005599 ± 0.0000030
human-chimpanzee divergence per base pair	0.0111174 ± 0.0000022	0.0082096 ± 0.0000090
Human diversity/divergence ratio	0.071578 ± 0.000065	0.068205 ± 0.000368

Note: This data set was compiled by James Mullikin, using different techniques for identifying divergent sites but the same DNA sequence comparisons.

These results indicate that the ancestral population of humans and chimpanzee just at the time of speciation $\tau^{HC}_{species}$ must have had an extremely different structure than humans do today. Specifically, the difference in the average time since the common ancestor on chromosome X and the autosomes in humans, is 9-10% of that in the ancestral population (see table below).

	Difference between X and autosome divergence time in humans today, as a fraction of ancestral population
Analysis #1 (M. Zody data)	10% ± 2%
Analysis #2 (J. Mullikin data)	9% ± 2%

A possible criticism of these analyses is that they assume a constant molecular clock since the divergence of humans and chimpanzees. However, "rate tests" applied to our data suggest that the molecular clock seems to have been approximately constant over the time period relevant to this analysis (Supp. Table 8,9). Moreover, our calculations suggest that small deviations from a constant molecular clock will not qualitatively affect our inference that the X-autosome difference in the ancestral population was an order of magnitude greater than in humans today.

Supplementary Note 8

Calculation of α, the ratio of male:female mutation rate since human-chimpanzee divergence

To estimate the relative mutation rate in males vs. females we use the equation from Taylor et al. (2005):

$$\alpha = \frac{3(X/A) - 4}{2 - 3(X/A)}$$

Here, X/A refers to the relative divergence per base pair on the X chromosome vs. the autosomes, assuming that the lower divergence on the X chromosome reflects entirely a lower mutation rate (and not lower time divergence). Human-chimpanzee divergence has been estimated based on the chimpanzee genome sequence to be 75.0% lower on the X chromosome than the autosomes at non-CpG dinucleotides (Taylor et al. 2005). Using this value in the equation suggests that the male to female mutation rate ratio has been high since human-chimpanzee divergence, around $\alpha \sim 7.0$ (the result from Taylor et al. (2005)).

A caveat noted in Taylor et al. (2005) is that the equation assumes that time divergence on the X chromosome and the autosomes is the same, but this needs to be corrected if there has been a different time divergence on the X chromosome and the autosomes. Corrections based on guesses about the difference between the X chromosome and the autosome in the ancestral population of humans and chimpanzees suggested that α might be as low as ~6 if the diversity in the ancestral population was as high as ~4 times that in humans today (Taylor et al. 2005).

However, the difference in time divergence between the X and the autosomes is much more extreme even than the highest value modeled in Taylor et al. (2005). Using our result that time divergence on the X chromosome is 0.835 lower on average than on the autosomes (Supp. Table 6), we obtain $X/A \sim 0.750/0.835 = 0.899$. Using this in the equation yields a much lower estimate of $\alpha = 1.9$ (95% CI: 1.7-2.1), consistent with earlier estimates of $\alpha \sim$ 1.9-2.1 (Lander et al. 2001; Rat Genome Sequencing Consortium 2004).

Supplementary Note 9

Upper bound on human-chimpanzee genome divergence time

We are interested in obtaining an upper bound on τ^{HC}_{genome}. This is done straightforwardly, following the example of previous researchers (Glazko and Nei 2003; Pilbeam and Young 2004). We do this by calibrating to the fossil record of human-orangutan, and human-macaque divergence.

Our primary calibration is to human-orangutan divergence. We use the *Proconsul* fossil, which is usually interpreted as showing that the human and orangutan lineages had speciated by ~18 Mya (MacLatchy et al. 2000). It seems unlikely that genome divergence occurred more than 2 My prior to speciation (otherwise the difference between these dates would have been more than in the ancestral population of humans and chimpanzees, and twice what we see in the most diverse apes today (Yu et al. 2004)). Thus, we obtain an upper bound on human-orangutan genome divergence of τ^{HO}_{genome} <20 Mya. Our secondary calibration is to human-macaque speciation. Using *Aegypitpithecus* fossils to place an upper bound on macaque divergence of ~33 Mya (Steiper et al. 2004), we obtain τ^{HM}_{genome} <35 Mya.

To convert from these upper bounds on divergence from distantly related primates, to an upper bound on chimpanzee divergence, we use algebra:

$$\tau^{HC}_{genome} = \tau^{HO}_{genome}/(\tau^{HO}_{genome}/\tau^{HC}_{genome})$$
$$\tau^{HC}_{genome} = \tau^{HM}_{genome}/(\tau^{HM}_{genome}/\tau^{HC}_{genome})$$

Since we have measured $\tau^{HO}_{genome}/\tau^{HC}_{genome}$=2.662 ± 0.015 and $\tau^{HM}_{genome}/\tau^{HC}_{genome}$ = 4.63 ± 0.13 (Supp. Table 8), we can convert from upper bounds on orangutan and macaque genome divergence, to upper bounds on human-chimpanzee divergence. The results turn out to be very similar: τ^{HC}_{genome}<7.5 Mya for the orangutan calibration, and τ^{HC}_{genome}<7.6 Mya for the macaque calibration.

An important caveat to these analyses is the molecular clock assumption. We assumed that the accumulation of mutations has occurred at a constant rate over time, and thus the relative genomic divergences are good estimates of time divergences. To detect failures in the molecular clock, we carried out a "rate test" searching for a difference in the mutation rate on the human side of the genealogy, or orangutan/macaque side, since the species' divergence (Sarich 1983). For the orangutan comparison there is no evidence of a major change in the molecular clock (Supp. Tables 8,9), increasing our confidence in the time divergence estimates. For the macaque comparison, there is clear evidence of a failure (Supp. Tables 8,9), with more mutations accumulating on the macaque lineage since divergence, than on the human (Steiper et al. 2004).

To save the information from the macaque fossil calibration, we therefore turned to the observation from Hwang and Green (2004) that in CpG dinucleotides, mutations have accumulated at a relatively constant rate of over time over the primate lineage (Hwang and Green 2004). Using only sites from CpG dinucleotides in the Hwang and Green data, we obtained $\tau^{HM}_{genome}/\tau^{HC}_{genome}$ = 4.76 ± 0.24 (Supp. Table 8). This produces an upper bound of τ^{HC}_{genome} <7.4 Mya.

We conclude that independent calibrations to the records of orangutan and macaque fossil divergence consistently show that τ^{HC}_{genome} <7.6 Mya, the result we use in our main analysis.

Supplementary Note 10

Empirical estimates of the X:autosome ratio "*R*" in five populations past and present

(i) *R* in the population ancestral to human and chimpanzee speciation (time $\tau^{HC}_{species}$)

To calculate the ratio of the autosome to X chromosome divergence in the population prior to human and chimpanzee speciation, we note from Supp. Table 6 that $\tau^{HC}_{X\text{-}chrom}/\tau^{HC}_{genome} = 0.835 \pm 0.016$, and thus:

$$R = \frac{\tau^{HC}_{X-chrom} - \tau^{HC}_{species}}{\tau^{HC}_{genome} - \tau^{HC}_{species}} = \frac{0.835 - \tau^{HC}_{species} / \tau^{HC}_{genome}}{1 - \tau^{HC}_{species} / \tau^{HC}_{genome}}$$

Under the demographic model in Supp. Note 2, $0.57 < \tau^{HC}_{species}/\tau^{HC}_{genome}$. Using this in the equation implies that $R<62\%$.

Fossil data provide a more substantial lower bound on $\tau^{HC}_{species}/\tau^{HC}_{genome}$. If we operate under the null hypothesis that human or chimpanzee ancestors did not originate by hybridization, then it is unlikely that human-chimpanzee ancestral gene flow occurred more recently than the hominin *Orrorin* and *Ardipithecus* fossils, which date to ~5.8 Mya (Senut et al. 2001; WoldeGabriel et al. 2001). Thus, we assume $\tau^{HC}_{species}>5.8$ Mya. From Supp. Note 9, we also have $\tau^{HC}_{species}<7.6$ Mya. We thus obtain $0.76 = 5.8/7.6 = <\tau^{HC}_{species}/\tau^{HC}_{genome}$. Using this in the equation implies that $R<29\%$.

We note that even if we use the largest value of $\tau^{HC}_{X\text{-}chrom}/\tau^{HC}_{genome}$ consistent with our data (0.861, which we obtain by adding 0.835 to 1.65 times the standard deviation of 0.016 and inputting it into the equation above), we still obtain estimates of *R* for the population ancestral to humans, chimpanzees and gorillas of <0.42, much less than the expectation for a freely mixing population.

Supp. Table 10 presents these results graphically. This analysis implies that we must either accept that $R<29\%$, or that human or chimpanzee ancestors originated by hybridization.

(ii) *R* in the population ancestral to human-chimpanzee-gorilla speciation (time $\tau^{HG}_{species}$)

The analog to the equation above for the ancestral population of humans, chimpanzees, and gorillas is $R = (0.980-\tau^{HG}_{species}/\tau^{HG}_{genome})/(1-\tau^{HG}_{species}/\tau^{HG}_{genome})$, since $\tau^{HG}_{X\text{-}chrom}/\tau^{HG}_{genome} = 0.980 \pm 0.020$ (Supp. Table 6). We do not have a good minimum for $\tau^{HG}_{species}$ based on the fossil record, so instead we use the results in Supp. Note 2 to find a range of allowed values of $\tau^{HG}_{species}/\tau^{HG}_{genome}$.

The table in Supp. Note 2 suggests that $0.728 < \tau^{HG}_{species}/\tau^{HG}_{genome} < 0.834$ for the full range of models consistent with the data. Using this in the equation above, we obtain $0.88 < R < 0.93$. By contrast, this ratio is projected to be less than 0.29 for the ancestral population of humans and chimpanzees (see above). This emphasizes the stark contrast between the ancestral populations of humans and chimpanzees, and the ancestral population of humans, chimpanzees and gorillas.

Even if we use the smallest value of $\tau^{HG}_{X\text{-}chrom}/\tau^{HG}_{genome}$ consistent with our data (0.947, which we obtain by subtracting 0.980 minus 1.65 times the standard deviation in Supp. Table 6) and input it into the equation above, we still obtain estimates of *R* for the population ancestral to humans, chimpanzees and gorillas (0.68 - 0.81) that are much larger than the bound of <0.29 for the ancestral population of humans and chimpanzees. (If we use the largest value of $\tau^{HG}_{X\text{-}chrom}/\tau^{HG}_{genome}$ =1 consistent with our data, we obtain an estimate of $R = 1$.)

(iii) *R* in the population ancestral to bonobos and common chimpanzees
In a parallel project, we carried out shotgun sequencing of bonobo DNA, and compared it to DNA from common chimpanzees, to learn about the historical relationships of these species. The data are currently being prepared for publication by David Reich and Jennifer Caswell. An initial analysis with respect to the X:autosome ratio is presented here.

We wish to obtain an estimate of R for the ancestors of bonobos and common chimpanzees. The analog of the equation above is $R = (\tau^{BC}_{X\text{-chrom}}/\tau^{BC}_{genome} - \tau^{BC}_{species}/\tau^{BC}_{genome})/(1-\tau^{BC}_{species}/\tau^{BC}_{genome})$.

To obtain $\tau^{BC}_{X\text{-chrom}}/\tau^{BC}_{genome} \approx (\tau^{BC}_{X\text{-chrom}}/\tau^{HM}_{X\text{-chrom}})/(\tau^{BC}_{genome}/\tau^{HM}_{genome}) = 0.89 \pm 0.07$, we combine two sources of information. We use a human-bonobo-chimpanzee (HBC) alignment of 8,052,215 bp on the autosomes, and 186,480 on chromosome X, to obtain estimates of $\tau^{BC}_{X\text{-chrom}}/\tau^{HC}_{X\text{-chrom}}$ and $\tau^{BC}_{genome}/\tau^{HC}_{genome}$. We then multiply these by $\tau^{HC}_{X\text{-chrom}}/\tau^{HM}_{X\text{-chrom}}$ and $\tau^{HC}_{genome}/\tau^{HM}_{genome}$, respectively (given in Supp. Note 7), to obtain the numerator and denominator we need.

To obtain $\tau^{BC}_{species}/\tau^{BC}_{genome} = 0.55 \pm 0.05$, we combine data from a western-western-central-human-macaque alignment (WWCHM, 5,330,435 bp), a central-central-western-human-macaque alignment (CCWHM, 5,026,834 bp), and a western-central-bonobo-human-macaque alignment (WCBHM, 598,814 bp). We use the EM analysis (Supp. Note 3) to estimate the branch lengths. To obtain an estimate of speciation time based on these results, we use a simplifying demographic model of the structure of the ancestral population, and nearly the same model for demographic history as in Supp. Note 2 (J. Caswell and D. Reich in preparation). This allows us to calculate $\tau^{BC}_{species}/\tau^{BC}_{genome} = \tau^{BC}_{species}/(\tau^{BC}_{genome} + 2N_{BC}) = 0.55 \pm 0.05$.

Substituting these results into the equation above produces an estimate of $R = 0.75 \pm 0.21$ for the population ancestral to chimpanzees and bonobos. Although the error around this estimate is large, the value is consistent with the expectation R=0.75 for a freely mixing population without selection.

(iv) *R* in humans today. At first glance it seems simple to calculate R in humans: just divide the genetic divergence on the X chromosome, by that on the autosomes. However, it is known that mutation rates are higher on average on the autosomes than on the X, and so this is inappropriate.

To address this problem, researchers have often divided human heterozygosity on the autosomes and X, by human-chimpanzee divergence in the same regions as a normalization for differences in the mutation rate. However, this normalization is not sufficient either. The normalization assumes that the time divergence of humans and chimpanzees is the same on the X and the autosomes, but our data shows that it is very much less (Supp. Table 6).

In what follows we therefore compare X chromosome and autosome heterozygosity by using human-macaque divergence as an alternative normalization for differences in the mutation rate. We use the following equation to convert between the estimates of human heterozygosity normalized by human-chimpanzee divergence, to human-heterozygosity normalized by macaque divergence:

$$R = \frac{\tau^{HH}_{X-chrom}}{\tau^{HH}_{genome}} \approx \frac{\tau^{HH}_{X-chrom}/\tau^{HM}_{X-chrom}}{\tau^{HH}_{genome}/\tau^{HM}_{genome}} = \frac{\left(\tau^{HH}_{X-chrom}/\tau^{HC}_{X-chrom}\right)\left(\tau^{HC}_{X-chrom}/\tau^{HM}_{X-chrom}\right)}{\left(\tau^{HH}_{genome}/\tau^{HC}_{genome}\right)\left(\tau^{HC}_{genome}/\tau^{HM}_{genome}\right)}$$

To estimate $\tau^{HH}_{X\text{-}chrom}/\tau^{HC}_{X\text{-}chrom}$ and $\tau^{HH}_{genome}/\tau^{HC}_{genome}$ we return to the human diversity data described in Supp. Note 7. In addition to the calculations based on comparison of the reference genome sequence to African American shotgun data (with the same data analyzed separately by M. Zody and J. Mullikin), we also included analyses of a third and fourth heterozygosity data set.

The first new analysis was an assessment of human heterozygosity in European Americans, comparing the HuAA and HuBB libraries generated by the Celera human genome sequencing project (Venter et al. 2001). (This analysis was carried out by J. Mullikin.)

The second new analysis was an assessment of western chimpanzee heterozygosity, carried out by Michael Zody who compared the reference chimpanzee genome sequence (from "Clint", a western chimpanzee) to one of two other western chimpanzees that were studied at ~0.1x shotgun coverage (Yvonne and Karlien) as part of the chimpanzee sequencing project (Mikkelsen et al. 2005).

The results are quoted in the table below, along with estimates of *R* obtained by inputting these measurements of population diversity into the equation above (we also use the estimates $\tau^{HC}_{X\text{-}chrom}/\tau^{HM}_{X\text{-}chrom} = 0.1527 \pm 0.0040$, and $\tau^{HC}_{genome}/\tau^{HM}_{genome} = 0.1821 \pm 0.0009$ from our data; Supp. Table 8). A 95% credible interval is calculated as +/- 1.96 standard deviations around the mean:

	Source	Autosomes	X	95% credible interval for R
African Americans	M. Zody	0.05817 ± 0.00008	0.05736 ± 0.00053	78 - 87%
African Americans	J. Mullikin	0.07158 ± 0.00007	0.06821 ± 0.00037	76 - 84%
European Americans	J. Mullikin	0.0608 ± 0.00008	0.04517 ± 0.00051	59 - 66%
Western chimpanzees	M. Zody	0.0605 ± 0.0008	0.0476 ± 0.0033	56 - 76%

We note that our analyses suggest that the X:autosome ratio is substantially lower for a comparison of two European Americans, than for a comparison of an African American to the public reference sequence. The difference may reflect real differences between the demographic histories African American and European American populations. To be conservative, we quote the inclusive range of 59-87% in the text for *R* for humans. We emphasize that all the estimates of "R" we have obtained for modern populations are higher than the upper bound of R<29% we have inferred for the ancestral population of humans and chimpanzees.

(v) X:autosome ratio in the chimpanzee population today. Using unpublished data from the chimpanzee genome sequencing project (M. Zody, H. Ji and D. Reich), we used the same procedure discussed in the previous section (iv) to estimate *R* for the modern western chimpanzee population. The analysis is summarized in the table above, leading to a 95% credible interval of R = 56-76% for the modern western chimpanzee population.

Summary of empirical estimates of "R" for 5 populations past and present	95% credible interval
i) Population prior to humans and chimpanzees speciation (time $\tau^{HC}_{species}$)	0 - 29%
ii) Population prior to gorilla speciation (time $\tau^{HG}_{species}$)	68 - 100%
iii) Population prior to bonobo and chimpanzee speciation (time $\tau^{BC}_{species}$)	33 -100%
iv) Population of humans today	59 - 87%
v) Population of western chimpanzees today	56 - 76%

Supplementary Note 11

We could not find a demography explaining the low X chromosome divergence between humans & chimpanzees

We explored a wide range of neutral demographic models, searching for scenarios that could explain a reduction in the $t_{HG}+t_{CG}$ branch length to less than 0-15% the value on the autosomes, and a reduction of R to <0.29. The motivation for this is to explore more demographic scenarios than those described in Supp. Note 2.

We could not find a model for the demographic structure of the ancestral population that could by itself explain the evidence of low X chromosome divergence. This suggests that natural selection explains the low X chromosome divergence.

(i) We explored a wide range of demographic histories but could not explain the patterns on the X chromosome. To explore whether there are demographic histories that can explain our data without invoking natural selection, we first remark that in Supp. Note 2 we considered ancestral populations that were constant in size, and found that there was no model in this class that could explain our data. To generalize these results, we now consider two periods: the period prior to human-gorilla speciation $t > \tau^{HG}_{species}$, and between human-chimpanzee and human-gorilla speciation $\tau^{HC}_{species} < t < \tau^{HG}_{species}$.

We first consider the time prior to human-gorilla speciation $t > \tau^{HG}_{species}$. We recall from Supp. Note 10 that in the population ancestral to humans, chimpanzees, and gorillas, the average time since the common ancestor on the X chromosome was likely >87% of that on the autosomes (greater than the expectation of 75% for a freely mixing, constant-sized population). This means that for those sections of the genome where the human and chimpanzee lineages trace back to $\tau^{HG}_{species}$ without sharing a common ancestor, the rate of HG and CG events is expected to be reduced by only ~87% at most compared with the autosomes. Thus, most of the reduction in the X divergence comparing humans and chimpanzees, must be due to demographic events in the period $\tau^{HC}_{species} < t < \tau^{HG}_{species}$.

Second, we consider the period between speciations $\tau^{HC}_{species} < t < \tau^{HG}_{species}$. We begin by considering a freely mixing population that was changing in size. If there was male-female symmetry in this population—if males and females had equal distributions of offspring—the probability that humans and chimpanzees share a common ancestor during any generation is 4/3 higher than the autosomes. Coalescent theory (Kingman et al. 1982) then shows we can rescale the population to an "effective" size of N_e for this period so that the probability of humans and chimpanzees not sharing an ancestor for the X and the autosomes ($p_{X\text{-chrom}}$ and $p_{autosomes}$) between $\tau^{HC}_{species} < t < \tau^{HG}_{species}$ is:

$$p_{autosomes} = e^{-\left(\tau^{HG}_{species} - \tau^{HC}_{species}\right)/2N_e}$$
$$p_{X-chrom} = e^{-(4/3)\left(\tau^{HG}_{species} - \tau^{HC}_{species}\right)/2N_e} .$$

Algebra then shows that $(p_{X\text{-chrom}}) = (p_{autosomes})^{4/3}$. To use these results to make a prediction about the reduction in the $t_{HG}+t_{CG}$ rate on the autosomes, we recall from Supp. Note 2 that a 95% credible interval for $p_{autosomes}$ is 18-29%. It follows that a 95% credible interval for $p_{X\text{-chrom}}/p_{autosomes}$ is 56-

66%, and thus we are confident that the proportion of the X chromosome where human and chimpanzees trace their lineage back to $\tau^{HG}_{species}$ without coalescing, should be at least 56% of the proportion on the autosomes. Since the X chromosome rate of HG and CG events in these sections should be at least 68% the rate of the autosomes (summary table in Supp. Note 10), we thus obtain 56% × 68% = 38% for a minimum $t_{HG}+t_{CG}$ rate on the X versus the autosomes. One caveat to this analysis is that the minimum of $p_{autosomes} > 18\%$ is itself obtained from a modeling analysis. However, repeating the analysis with $p_{autosomes}$ as low as 1% still gives predictions for $t_{HG}+t_{CG}$ on the X that are higher than the upper bound from our data.

More generally, we note that for any demographic scenario we have been able to construct, we have the inequality $(p_{X\text{-}chrom}) \geq (p_{autosomes})^{4/3}$ (modeling not shown). Even allowing for a scenario where the population ancestral to humans and chimpanzees was actually highly substructured, we have not been able to construct a scenario even close to matching the data. These results suggest that the low X divergence in our data is at least in part due to the influence of natural selection, and cannot be explained by neutral demographic history alone.

(ii) Sex asymmetry cannot explain the low X divergence. We next explored whether asymmetry between the sexes in terms of their distributions of offspring can explain our data.

We consider a model where children are more evenly distributed among males than females. This scenario seems improbable as female primates usually invest more energy in each child than males, but it is what is necessary to reduce the X/autosome ratio. We consider for the sake of argument an extreme example where every female with offspring has four children fathered by different males (each with only one child). This means that the male effective population size is 4 times that of the female. The effective population size on the X relative to that on the autosomes is then:

$$\frac{N_{X-chrom}}{N_{autosomes}} = \frac{0.5(N_e/4) + 0.25(N_e)}{0.5(N_e/4) + 0.5(N_e)} = 0.6$$

Using these results and the same argument as in section (i) we estimate the rate of HG and CG events vs. the autosomes. Defining p as the proportion of the genome in which humans and chimpanzees share a common ancestor more recently than gorilla speciation, we obtain $(p_{X\text{-}chrom}) = (p_{autosomes})^{1/0.6}$. Recalling from Supp. Note 2 that a 95% credible interval for $p_{autosomes}$ is 18-29%, it follows that a 95% credible interval for $p_{X\text{-}chrom}/p_{autosomes}$ is 32-44%, and thus we are confident that the proportion of the X chromosome where human and chimpanzees trace their lineage back to $\tau^{HG}_{species}$ without coalescing, should be at least 32% of the proportion on the autosomes.

Since the X chromosome rate of HG and CG events in these sections should be at least 87% the rate of the autosomes (above and Supp. Note 10), we thus obtain 32% × 68% = 22% for a minimum $t_{HG}+t_{CG}$ rate on the X versus the autosomes. In fact, our data strongly indicate that the maximum consistent with the data is <15% (Supp. Note 6).

We conclude that in the presence of sex asymmetry, or extreme demography, we cannot observe a relationship between the X chromosome and autosomes similar to what we see in our data. The reduction below expectation is so extreme that even a combination of sex asymmetry and extreme demography would have great difficulty in explaining the data. Natural selection must instead explain the low X/autosome ratio and near absence of HG and CG clustering on the X chromosome.

Accessing raw data and alignments

"Genetic evidence for complex speciation of humans and chimpanzees"
Patterson N, Richter DJ, Gnerre S, Lander ES and Reich D; *Nature* 2006

Sequence obtained for this study: We sequenced 117,862 reads of DNA: 115,152 from a western lowland gorilla (*Gorilla gorilla,* individual NG05251 in the Coriell catalog: locus.umdnj.edu/primates/species_summ.html) and 2,710 from a black-handed spider monkey (*Ateles geoffryi*, individual NG05352). All sequencing reads are publicly available at the NCBI trace archive (http://www.ncbi.nlm.nih.gov/Traces); to access them, carry out the following queries:

(1) Gorilla data (*Gorilla gorilla*):
CENTER_NAME='WIBR' and CENTER_PROJECT='G611'
CENTER_NAME='WIBR' and CENTER_PROJECT='G612'
CENTER_NAME='WIBR' and CENTER_PROJECT='G618'
CENTER_NAME='WIBR' and CENTER_PROJECT='G619'
CENTER_NAME='WIBR' and CENTER_PROJECT='G744'

(2) New world monkey data (*Ateles geoffroyi*)
CENTER_NAME='WIBR' and CENTER_PROJECT='G820'

We note that the NCBI trace archive contains slightly more reads that we report in our analyses, because not every read submitted to the Trace Archive passed standard pre-filtering steps.

Alignments: The alignments of humans, chimpanzees, gorillas, and more distantly related primates can be downloaded from our lab website (http://genepath.med.harvard.edu/~reich) or the *Nature* website. The first two data sets are packaged into "tar" files. When opened with the unix command "tar –xvf name", these expand into many files: one for each alignment. The third and fourth data sets, corresponding to alignments of contiguous sequence, are in Threaded Block Set aligner (tba) format, and packaged into "gz" files. These can be opened with the unix command "gunzip name".

(1) HCGOM shotgun data	hcgom_aligns.tar	33,016 alignments
(2) HCGM shotgun data	hcgm_aligns.tar	51,966 alignments
(3) HCGOM contiguous chr. 7	hcgom7_contig_aligns.tba.gz	1 contiguous alignment
(4) HCGOM contiguous chr. X	hcgomX_contig_aligns.tba.gz	1 contiguous alignment

Data sets: The filtered data can be downloaded from our lab website (http://genepath.med.harvard.edu/~reich) or the *Nature* website. Data are packaged into "gz" files, which can be opened with the unix command "gunzip name".

(1) HCGOM shotgun data	hcgom_shotgun.gz	498,771 divergent sites
(2) HCGM shotgun data	hcgm_shotgun.gz	858,941 divergent sites
(3) HCGOM contiguous chr. 7	hcgom7_contig.gz	69,521 divergent sites
(4) HCGOM contiguous chr. X	hcgomX_contig.gz	8,769 divergent sites

Further questions: Please contact David Reich (reich@genetics.med.harvard.edu) for any further clarifications about these data

References

Baum, L.E., Petrie, T., Soules, G., Weiss, N. A maximization technique occurring in the statistical analysis of probabilitic functions of Markov Chains. *Ann. Math. Stat.* **41,** 164-171 (1970).

Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14,** 708-715 (2004).

Cheng Z, *et al.* A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437,** 88-93 (2005).

Efron B, Tibshirani R. An introduction to the bootstrap. (Chapman and Hall, 1993).

Frank, M., Meijer, E., van der Leeden, R. Delete-m jackknife for unequal m. *Statistics and Computing* **9,** 3-8 (1999).

Glazko, G.V. & Nei, M. Estimation of divergence times for major lineages of primate species. *Mol. Biol. Evol.* **20,** 424-434 (2003).

Huang, X. On global sequence alignment. *Comp. App. Biosc.* **10,** 227-235 (1994).

Hwang, D.G. & Green, P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. USA* **101,** 13994-14001 (2004).

Jaffe D.B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13,** 91-96 (2003).

Kingman, J.F.C. The Coalescent. *Stochastic processes and their applications.* **13,** 235-248 (1982).

Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860-921 (2001).

MacLatchy, L., Gebo, D., Kityo, R., Pilbeam, D. Postcranial functional morphology of Morotopithecus bishopi, with implications for the evolution of modern ape locomotion. *J. Hum. Evol.* **38,** 1-25 (2000).

Mikkelsen, T.S. *et al.* (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69-87.

Pilbeam, D. & Young, N. Hominoid evolution: synthesizing disparate data. *Comptes Rendus Palevol.* **3,** 305-321 (2004).

Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493−521 (2004).

Sarich, V.M. Retrospective on hominoid macromolecular systematics. In *New Interpretations of Ape and Human Ancestry* (eds Ciochon, R.L. & Corruccini, R.S.) 137-150 (Plenum press, New York, 1983).

Schwartz S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13,** 103-107 (2003).

Senut, B., Pickford, M., Gommery, D., Mein, P., Cheboi, K., Coppens, Y. First hominid from the Miocene (Lukeino Formation, Kenya). *C. R. Acad. Sci. Ser. IIa* **332,** 137–144 (2001).

Steiper, M.E., Young, N.M., Sukarna, T.Y. Genomic data support the hominoid slowdown and an Early Oligocene estimate for the hominoid-cercopithecoid divergence. *Proc Natl Acad Sci USA* **101,** 17021-17026 (2004).

Taylor, J., Tyekucheva, S., Zody, M., Chiaromonte, F., Makova, K.D. Strong and Weak Male Mutation Bias at Different Sites in the Primate Genomes: Insights from the Human-Chimpanzee Comparison. *Mol Biol Evol* **23,** 565-573 (2006).

Venter, J.C. The sequence of the human genome. *Science* **291,** 1304-1351.

Yu, N., Jensen-Seaman, M.I., Chemnick, L., Ryder, O., Li, W.H. Nucleotide diversity in gorillas. *Genetics* **166,** 1375-1383 (2004).