Supporting Information

Nam et al. 10.1073/pnas.1419306112

SI Text

Background Selection. Background selection (BGS) reduces diversity locally in a genome by a process in which deleterious mutations are continuously pruned from the population, effectively reducing the number of genes from which future generations can be sampled. The strength of BGS is determined by the rate at which deleterious mutations enter the population, U, the recombination rate, R, and the strength of selection, s. If π_0 denotes the neutral diversity in the genome and if π is the diversity in a locus experiencing BGS, then the reduction in diversity is given by

$$\frac{\pi}{\pi_0} = \exp\left(-\frac{U}{s+R}\right)$$

(equation 6.24 in ref. 1).

We will assume a constant per-nucleotide deleterious mutation rate and recombination. Then, the rates U and R are both functions of the locus lengths, $U = uL$ and $R = rL$, where u is the per-nucleotide deleterious rate and r denotes the per-nucleotide pair recombination rate.

Comparing diversity inside with outside the regions of low diversity, we are interested in the relative reduction, which can be caused by changes to the selection rate, mutation rate, or recombination rate as factors of f_U, f_S, and f_R, respectively. The relative reduction can, thus, be expressed as

$$\frac{\exp\left(-\frac{f_U \times U}{f_s \times s + f_R \times R}\right)}{\exp\left(-\frac{U}{s+R}\right)}.$$

Although all of the parameters in these equations are unknown, we do have some knowledge of their general order of magnitude or can choose conservative values to increase the relative reduction in diversity explainable by BGS.

BGS is strongest when selection is weak, but when s is very small, we expect the evolution to be nearly neutral. After s approaches $1/Ne$, we do not expect any BGS effects at all. Because Ne is on the order of 10,000–100,000, a lower limit on s is 10^{-4}–10^{-5}. We consider both $s = 10^{-5}$ and $s = 10^{-4}$ and allow the selection inside the low-diversity regions to be one-tenth of the outside to make BGS stronger there.

We do not have recombination maps for most of the species considered, but from the human map, we know that the mean recombination rate on the X chromosome in the regions not showing reduced diversity is around 1.5 cm/Mb and that difference between the recombination rate inside and outside the reduced regions is less than a factor of two; therefore, we use $f_R = 0.5$.

We have very little information about what the deleterious mutation rate is, but we can use the mean human mutation rate, estimated to be 1.2×10^{-8} per generation (2), to explore various possibilities. If we use 1.2×10^{-8}, it would amount to assuming that 100% of mutations are under weak negative selection. By multiplying it with a number d between zero and one, we can interpret this number as the fraction of the loci that we believe are under selection. In Fig. S6, the columns correspond to different choices of d from 1% to 10% combined with different choices for s.

The rows in Fig. S6 correspond to different choices of f_U varying from 1 to 10 (therefore, the combination of $d = 0.1$ and $f_U = 10$ at the bottom right of Fig. S6 amounts to assuming that all sites within the low=diversity regions are under selection) combined with different choices for f_s, with $f_s = 1.0$ for no difference in the selection strength inside and outside of the low-diversity regions and $f_s = 0.1$ setting the selection inside the regions an order of magnitude lower. Fig. S6*A* shows the reduction in diversity compared with a neutral π_0 and the relative diversity of inside to outside of the low-diversity regions. In Fig. S6*B*, the dashed red line indicates 20%.

In the most extreme cases, we see a reduction in diversity of about 20% of the diversity within the low-diversity regions compared with outside but only in the cases where 100% of the nucleotides within regions are under selection. In the cases where 50% of the nucleotides are under selection within the regions, compared with 5% outside, the regions still retain about 50% of the diversity seen outside the regions.

Simulation of Sweeps. To assess the potential effect of hard and soft sweeps on diversity on the X chromosome, we performed a large number of simulations of a Wright–Fisher model exploring combinations of selection coefficients (s), effective population sizes (N), and frequencies of the selected variant at the onset of selection (f). We compute the time to the most recent common ancestor (TMRCA) along two recombining sequences and use this as a proxy for nucleotide diversity.

To simulate a selective sweep, we first sample frequency trajectories of a variant selected by s. We do this using rejection sampling (rejecting trajectories where the selected variant does not go to fixation). Trajectories for hard sweeps begin at 1 and proceed to $2N \times 3/4$ by repeated binomial sampling with probability parameter

$$\frac{N_{mut}}{N_{mut} + (N - N_{mut}) \times (1 - s)},$$

where N_{mut} is the number of selected variants in the previous generation. Trajectories for soft sweeps begin with an initial frequency f of the selected variant and are prepended with a trajectory from 1 to $f \times 2N \times 3/4$ representing variant frequency before the onset of the selection.

For each trajectory, we then consider a sample of two sequences representing 10 cm in length (equivalent to 10 Mb assuming a recombination rate of 1 cm/Mb). Because the effect of a sweep on flanking diversity is expected to be symmetric, we put the selected variant at the 5′-end position. To compute the TMRCA along the two sequences, we simulate backward the coalescence with recombination in discrete generations, allowing multiple mergers but only one recombination event per lineage in each generation.

Given a recombination event, the sequence downstream of the recombination point will become unlinked from the sweep with a probability equal to the frequency of chromosomes in the population not linked to the variant. A recombination event may similarly cause unlinked sequence fragments to again become linked to the selected variant with a probability equal to the frequency of chromosomes carrying the variant. Lineages that carry the selected variant can only share an immediate ancestor with other lineages that also carry this variant.

The simulation proceeds until all sequence segments separated by recombination events have found a most recent common

ancestor. For each combination of parameters s, N, and f, we perform 1,000 simulations, and the mean TMRCA along the 10 Mb is computed in bins of 10 kb (Fig. S7).

Complete List of the Great Ape Genome Project. Javier Prado-Martinez[a], Peter H. Sudmant[b], Jeffrey M. Kidd[c,d], Heng Li[e], Joanna L. Kelley[d], Belen Lorente-Galdos[a], Krishna R. Veeramah[f], August E. Woerner[f], Timothy D. O'Connor[b], Gabriel Santpere[a], Alexander Cagan[g], Christoph Theunert[g], Ferran Casals[a], Hafid Laayouni[a], Kasper Munch[h], Asger Hobolth[h], Anders E. Halager[h], Maika Malig[b], Jessica Hernandez-Rodriguez[a], Irene Hernando-Herraez[a], Kay Prüfer[g], Marc Pybus[a], Laurel Johnstone[f], Michael Lachmann[g], Can Alkan[i], Dorina Twigg[c], Natalia Petit[a], Carl Baker[b], Fereydoun Hormozdiari[b], Marcos Fernandez-Callejo[a], Marc Dabad[a], Michael L. Wilson[j], Laurie Stevison[k], Cristina Camprubí[l], Tiago Carvalho[a], Aurora Ruiz-Herrera[l,m], Laura Vives[b], Marta Mele[a], Teresa Abello[n], Ivanela Kondova[o], Ronald E. Bontrop[o], Anne Pusey[p], Felix Lankester[q,r], John A. Kiyang[q], Richard A. Bergl[s], Elizabeth Lonsdorf[t], Simon Myers[u], Mario Ventura[v], Pascal Gagneux[w], David Comas[a], Hans Siegismund[x], Julie Blanc[y], Lidia Agueda-Calpena[y], Marta Gut[y], Lucinda Fulton[z], Sarah A. Tishkoff[aa], James C. Mullikin[bb], Richard K. Wilson[z], Ivo G. Gut[y], Mary Katherine Gonder[cc], Oliver A. Ryder[dd], Beatrice H. Hahn[ee], Arcadi Navarro[a,ff,gg], Joshua M. Akey[b], Jaume Bertranpetit[a], David Reich[e], Thomas Mailund[h], Mikkel H. Schierup[h,hh], Christina Hvilsom[x,ii], Aida M. Andrés[g], Jeffrey D. Wall[k], Carlos D. Bustamante[d], Michael F. Hammer[f], Evan E. Eichler[b,jj], and Tomas Marques-Bonet[a,gg]

[a]Institut de Biologia Evolutiva (Consejo Superior de Investigaciones Científicas-Universitat Pompeu Fabra), Parc de Recerca Biomèdica de Barcelona, Barcelona, Catalonia 08003, Spain; [b]Department of Genome Sciences, University of Washington, Seattle, WA 98195; [c]Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109; [d]Department of Genetics, Stanford University, Stanford, CA 94305; [e]Department of Genetics, Harvard Medical School, Boston, MA 02115; [f]Arizona Research Laboratories, Division of Biotechnology, University of Arizona, Tucson, AZ 85721; [g]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig 04103, Germany; [h]Bioinformatics Research Centre and [hh]Department of Bioscience, Aarhus University, Aarhus C DK-8000, Denmark; [i]Faculty of Engineering, Bilkent University, Ankara 06800, Turkey; [j]Department of Anthropology, University of Minnesota, Minneapolis, MN 55455; [k]Institute for Human Genetics, University of California, San Francisco, CA 94143; [l]Departament de Biologia Cellular, Fisiologia i Immunologia and [m]Institut de Biotecnologia i de Biomedicina, Universitat Autonoma de Barcelona, Cerdanyola del Valles, Catalonia 08193, Spain; [n]Parc Zoologic de Barcelona, Barcelona, Catalonia 08003, Spain; [o]Biomedical Primate Research Centre, 2280 GH Rijswijk, The Netherlands; [p]Department of Evolutionary Anthropology, Duke University, Durham, NC 27708; [q]Limbe Wildlife Centre, BP 878 Limbe, Cameroon; [r]Paul G. Allen School for Global Animal Health, Washington State University, WA 99164; [s]North Carolina Zoological Park, Asheboro, NC 27205; [t]Department of Psychology, Franklin and Marshall College, Lancaster, PA 17604; [u]Department of Statistics, Oxford University, Oxford OX1 3TG, United Kingdom; [v]Department of Genetics and Microbiology, University of Bari, Bari 70126, Italy; [w]Department of Cellular and Molecular Medicine, University of California at San Diego, La Jolla, CA 92093; [x]Department of Biology, Bioinformatics, University of Copenhagen, Copenhagen 2200, Denmark; [y]Centro Nacional de Analisis Genomico, Parc Cientific de Barcelona, Barcelona, Catalonia 08028, Spain; [z]Genome Sequencing Center, Washington University School of Medicine, St. Louis, MO 63108; [aa]Department of Biology and Genetics and [ee]Departments of Medicine and Microbiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104; [bb]National Institutes of Health Intramural Sequencing Center, Bethesda, MD 20892; [cc]Biological Sciences, University at Albany, State University of New York, Albany, NY 12222; [dd]Genetics Division, San Diego Zoo's Institute for Conservation Research, Escondido, CA 92027; [ff]Instituto Nacional de Bioinformatica, UPF, Barcelona, Catalonia 08003, Spain; [gg]Institucio Catalana de Recerca i Estudis Avançats, Barcelona, Catalonia 08010, Spain; [ii]Copenhagen Zoo, Frederiksberg DK 2000, Denmark; and [jj]Howard Hughes Medical Institute, Seattle, WA 98195

1. Durrett R (2008) *Probability Models for DNA Sequence Evolution* (Springer, Berlin).

2. Kong A, et al. (2012) Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488(7412):471–475.

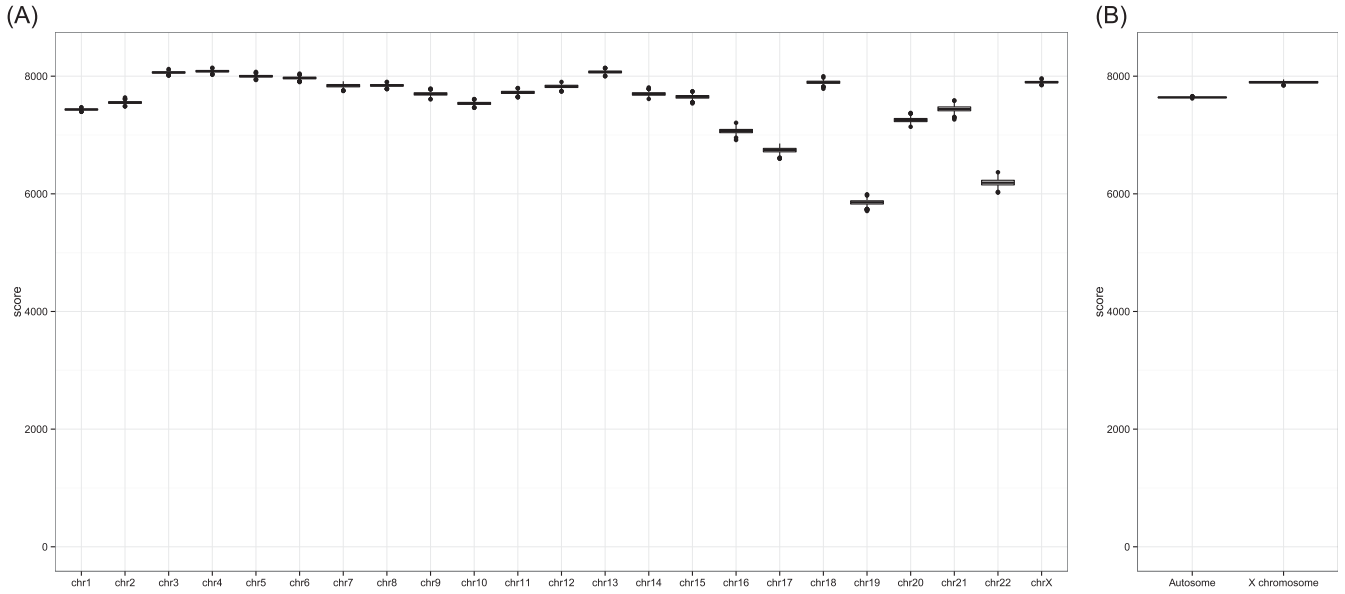

Fig. S1. The SNP quality score. (*A*) The SNP score in the variant call format files of each chromosome. (*B*) The SNP score of total autosomes and X chromosomes. The boxplots show the 95% confidence intervals calculated with 1,000 times bootstrapping replicates sampled from 1-Mb windows.

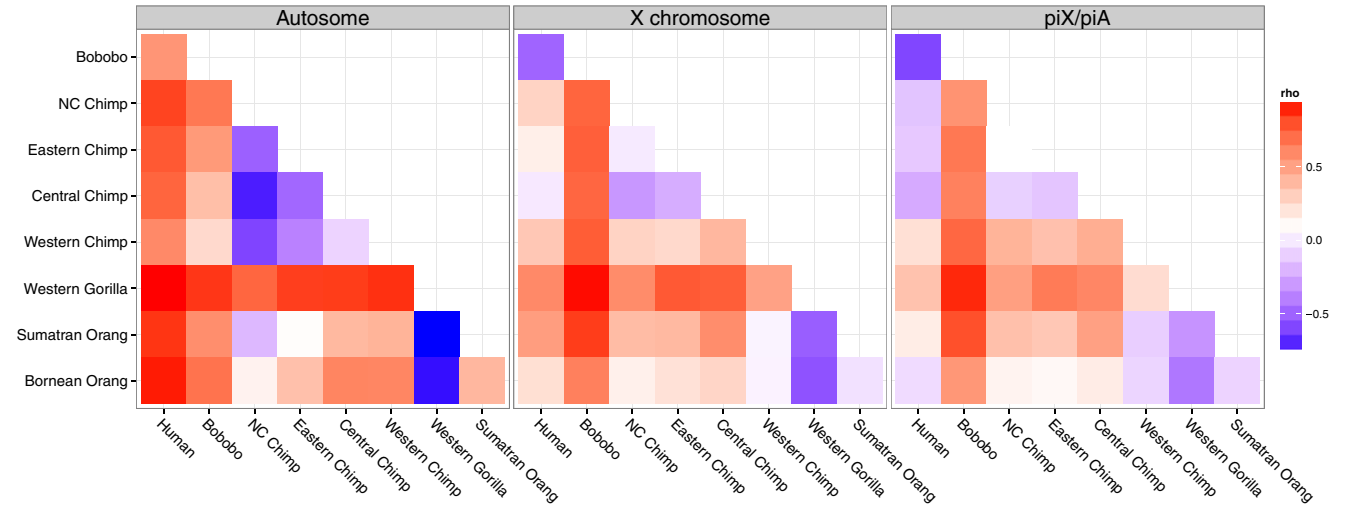

Fig. S2. Comparison of reduction in diversity between pairs of species. Heat maps of the correlation of the diversity ratio between two species and the physical distance from genes for the autosomes, the X chromosome, and the diversity ratio of X chromosomes to autosomes. The red color suggests a steeper relationship, with distance for the species indicated on the *y* axis; a blue color suggests a steeper relationship for the species on the *x* axis. NC, Nigeria–Cameroon.

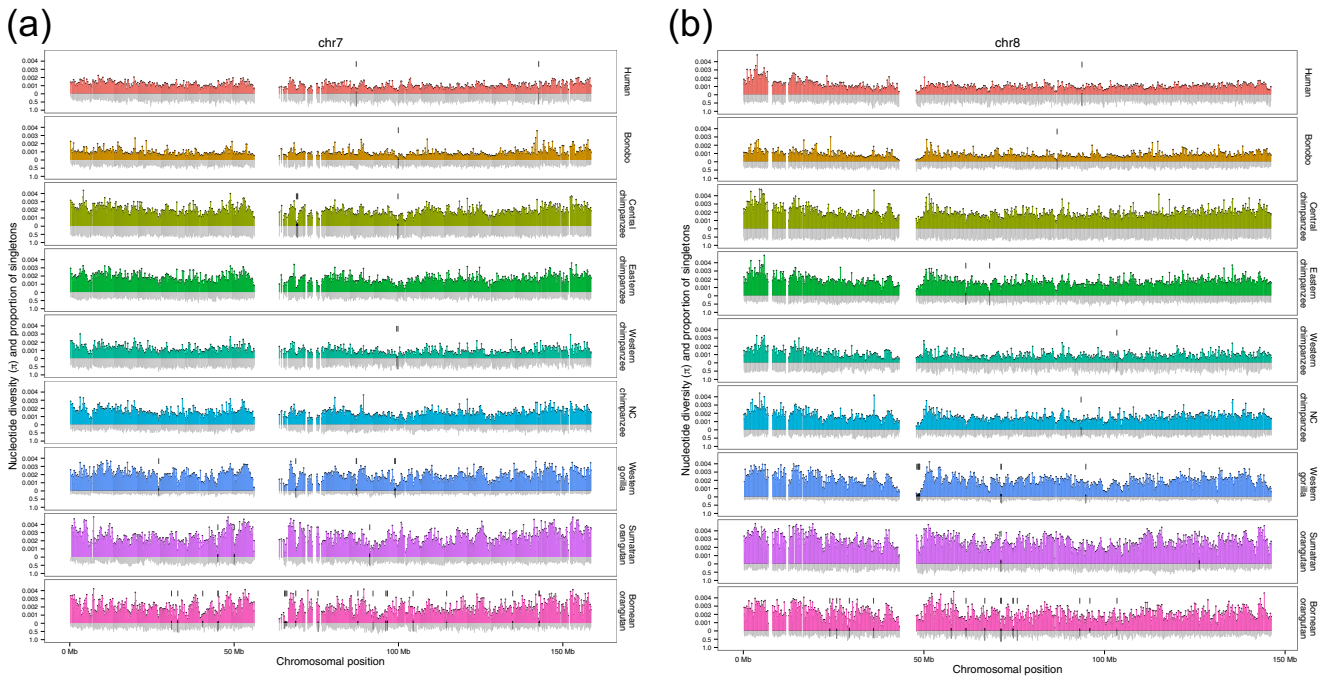

Fig. S3. The diversity pattern and site frequency spectrum of an autosome. The π and the proportions of singletons along (*A*) chromosome 7 and (*B*) chromosome 8, which have comparable chromosome size with X chromosomes. Windows with reduced diversity are identified from the lower π than 20% of the mean of each chromosome and denoted as black bars. NC, Nigeria–Cameroon.

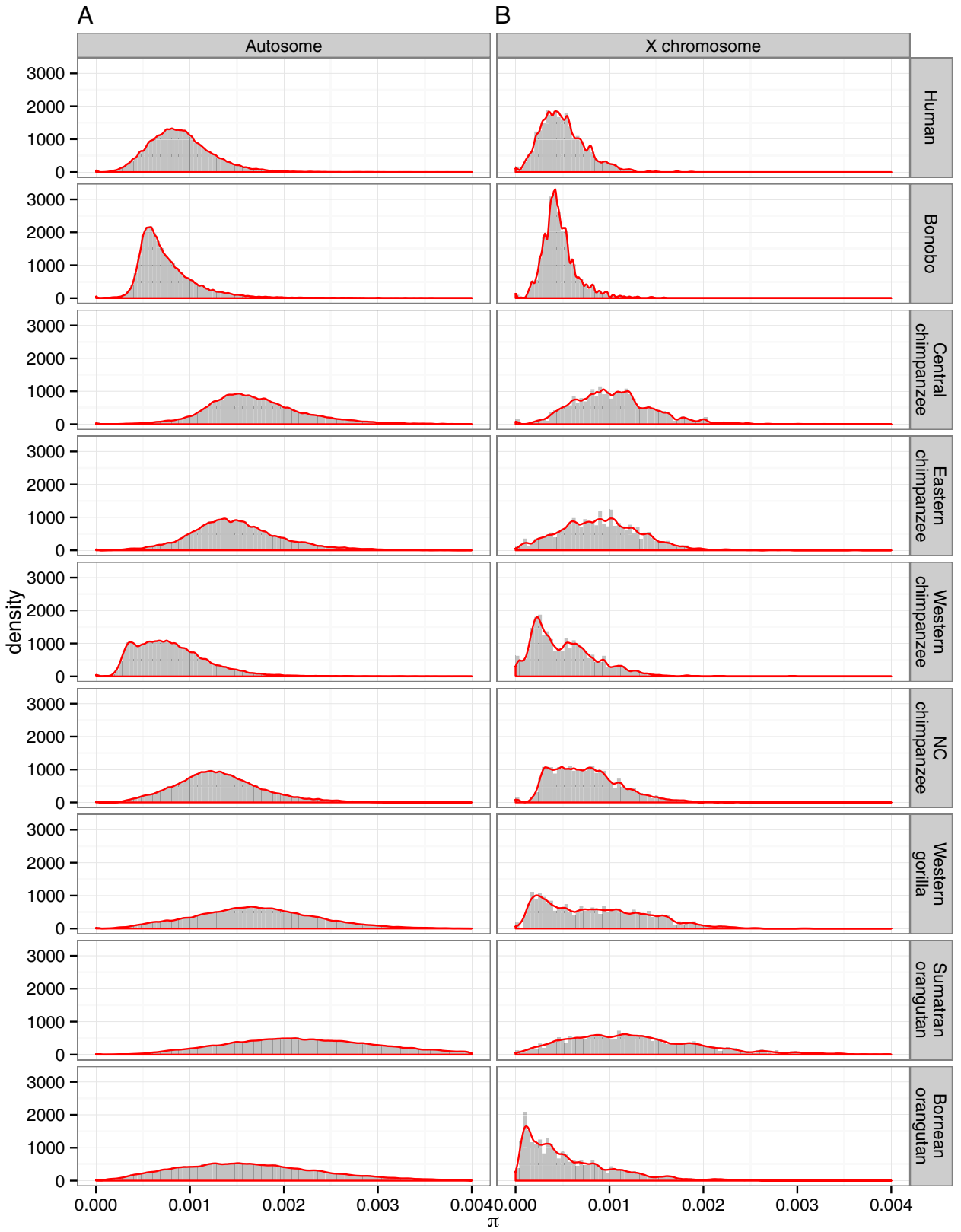

Fig. S4. The distribution of π of X chromosomes and autosomes. The histograms show the distribution of π of (*A*) autosomes and (*B*) X chromosomes. Red lines show kernel density function estimated by the Epanechnikov function, with bandwidths equal to 0.00004. NC, Nigeria–Cameroon.

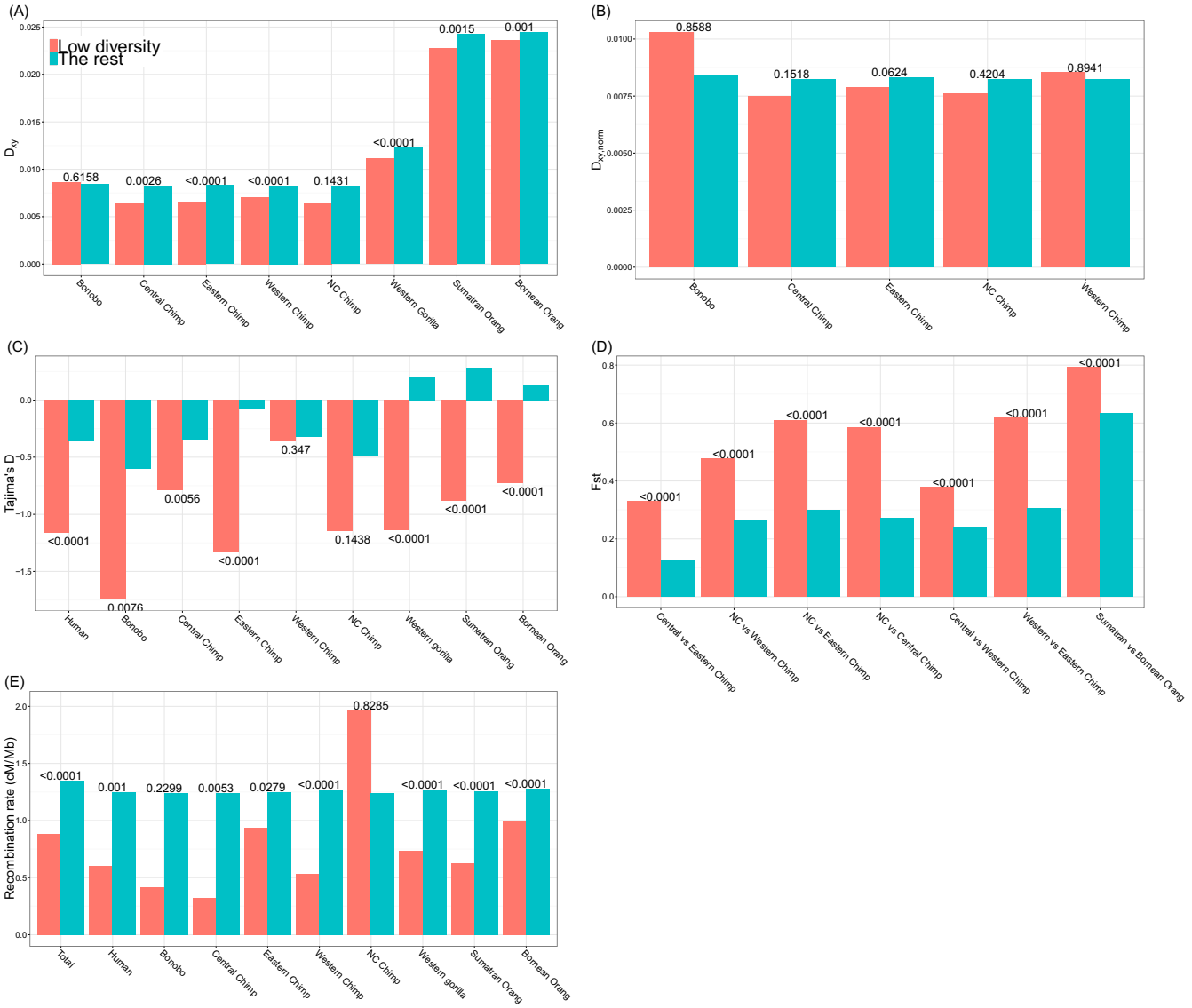

Fig. S5. Comparison between low-diversity windows and the rest of the X chromosome. (*A*) The D_{XY} values are divergence estimates between each species and human references (hg18). (*B*) The D_{XY} values of the *Pan* genus were normalized by the number of generations between each species to humans based on an assumption that a diversity-reducing process has been constantly operated since TMRCA. (*C*) The Tajima's *D* values. (*D*) The F_{ST} values were calculated from the pairs of the Nigeria–Cameroon (NC) chimpanzee, Eastern chimpanzee, Central chimpanzee, and Western chimpanzee and the pair of Sumatran orangutan and Bornean orangutan. (*E*) The human recombination rate. The lower-diversity windows are identified from π less than 20% of chromosomal average. The *P* values shown above each pair of bar plots were calculated by one-tailed bootstrapping test with 10,000 replicates.

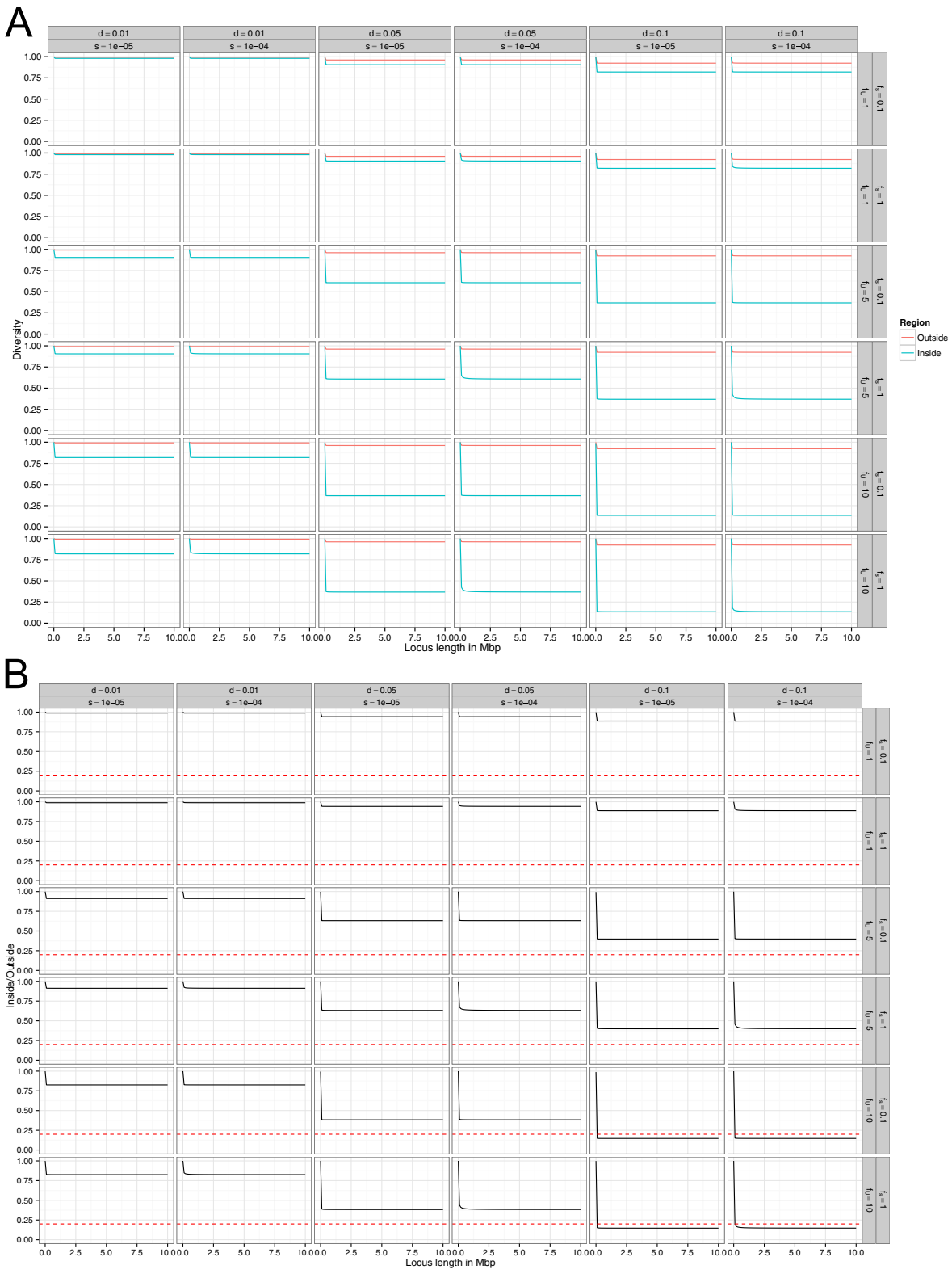

Fig. S6. The reduction in diversity caused by BGS. (*A*) The level of diversity compared with neutral diversity inside and outside the regions of interest and (*B*) the relative diversity inside and outside the regions of reduced diversity for various combinations of selection strengths and fractions of nucleotides under selection.

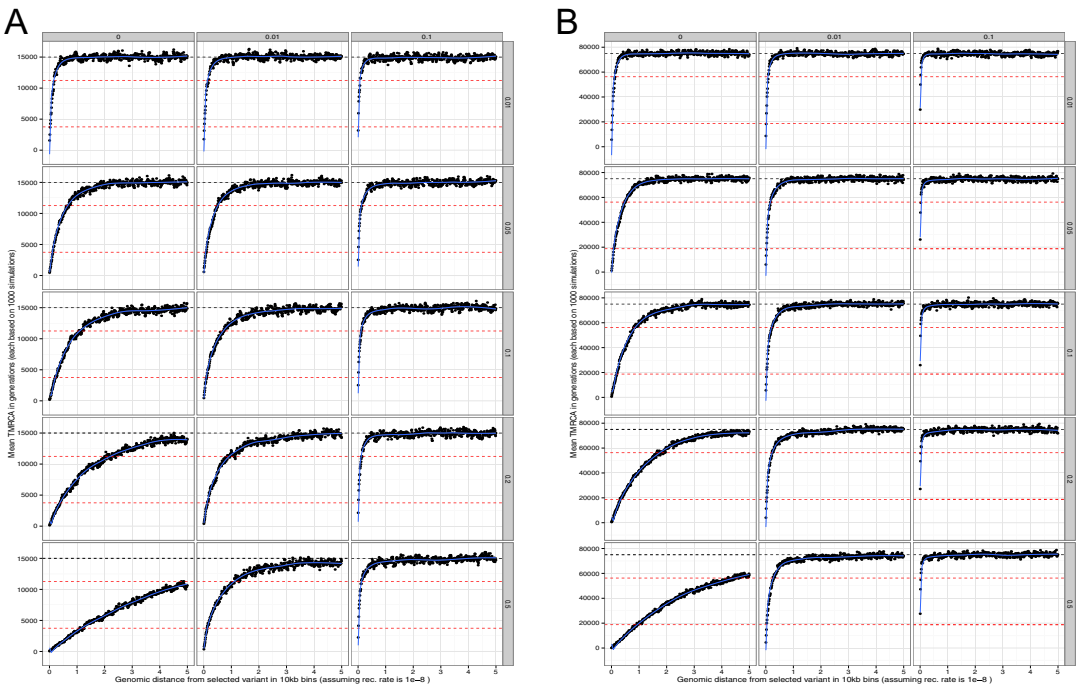

Fig. S7. Simulation results for strong sweep. Time to TMRCA as a function of genetic distance to selected variant is shown for combinations of *s* (right) and frequency of variant before onset of selection (top). Note that a frequency of zero represents hard sweeps. Each subplot is based on 1,000 independent simulations. Horizontal black dashed lines show expected TMRCA without selection. Red dashed lines show 75% and 25% reductions in TMRCA. The *Ne* is set to (*A*) 10,000 and (*B*) 50,000.

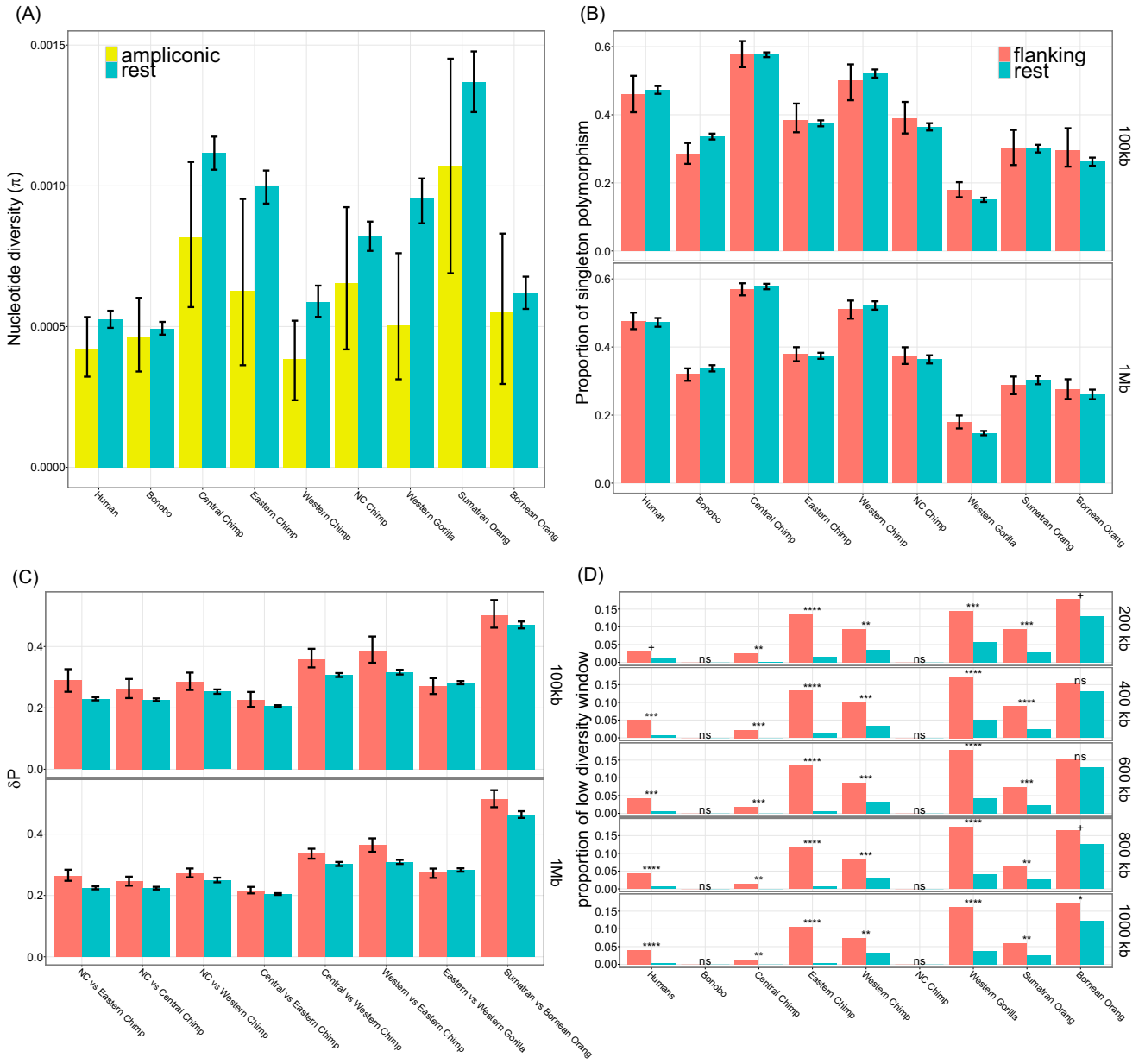

Fig. S8. The signature of selective sweeps on the flanking sequences of ampliconic regions. The X chromosome was divided into three regions: ampliconic regions (yellow), flanking regions (100 kb or 1 Mb in size; red), and the rest of X chromosomes (cyan). The bar plots show (*A*) π in the ampliconic regions and the rest, (*B*) the proportion of singleton polymorphisms in the flanking regions and the rest, and (*C*) the population differentiation estimated from the flanking regions and the rest. The error bars indicate 95% confidence intervals calculated from 1,000 bootstrapping iterations resampled from 1-Mb windows. (*D*) This bar plot shows the proportion of low-diversity windows identified from π less than 20% of X-chromosomal average in the flanking regions (200 kb to 1 Mb in size) and the rest. The bonobo and the Nigeria–Cameroon (NC) chimpanzee do not have a low-diversity window in both of two categories, because these two species have only one low-diversity window, which is entirely ampliconic. The significance levels shown above each pair of bar plots were calculated by one-tailed Fisher's exact tests (ns denotes $P \geq 0.10$). *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; ****$P < 0.0001$; $^{+}P < 0.10$.

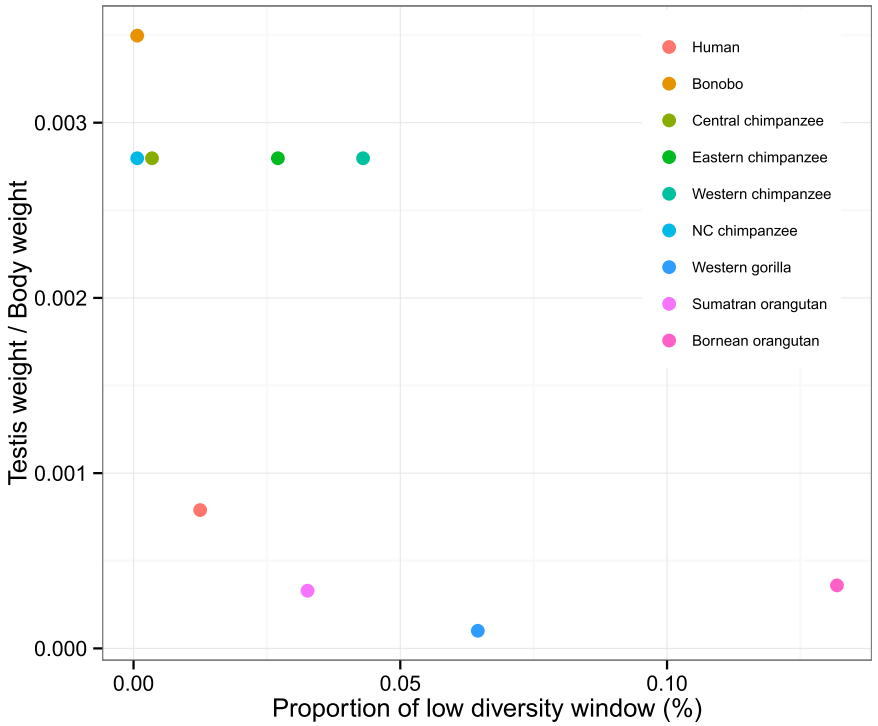

Fig. S9. Relationship between testis size and strength of sweeps. The *y* axis is the ratio of testicle to body weight, and the *x* axis is the proportion of windows that has π less than 20% of chromosomal average. NC, Nigeria–Cameroon.

Table S1. The information on taxa used in this study

Common name	Scientific name	No. of males	No. of females
Human	*Homo sapiens*	9	0
Bonobo	*Pan paniscus*	2	11
Central chimpanzee	*Pan troglodytes troglodytes*	1	3
Eastern chimpanzee	*Pan troglodytes schweinfurthii*	2	4
Western chimpanzee	*Pan troglodytes verus*	4	1
Nigeria–Cameroon chimpanzee	*Pan troglodytes ellioti*	4	6
Eastern lowland gorilla	*Gorilla beringei graueri*	2	1
Western lowland gorilla	*Gorilla gorilla gorilla*	4	23
Sumatran orangutan	*Pongo abelii*	1	4
Bornean orangutan	*Pongo pygmaeus*	1	4

Table S2. The information for each chromosome

			Number of SNPs									
Chromosome	Chromosome length	Called	Humans	Bonobo	Central chimp	Eastern chimp	Western chimp	Nigeria–Cameroon chimp	Eastern lowland gorilla	Western lowland gorilla	Sumatran orang	Bornean orang
chr1	247,249,719	160,880,685	524,659	532,077	753,781	702,539	738,892	401,273	246,742	1,076,881	914,308	653,958
chr2	242,951,149	178,023,644	590,073	615,637	903,996	810,204	855,174	426,799	258,017	1,261,636	1,077,591	753,329
chr3	199,501,827	149,518,151	507,464	501,741	759,156	669,066	696,690	349,394	231,696	1,031,974	882,837	617,080
chr4	191,273,063	140,398,337	497,502	499,566	751,745	656,108	704,374	300,581	233,459	1,068,263	962,245	673,556
chr5	180,857,866	132,273,796	445,741	457,800	671,543	598,132	631,469	291,695	197,085	956,564	830,049	585,159
chr6	170,899,992	124,881,358	434,679	441,828	608,543	558,000	594,738	341,588	192,778	885,763	765,994	546,720
chr7	158,821,424	105,796,935	371,207	375,976	532,815	489,230	516,539	287,830	163,721	763,850	633,918	453,587
chr8	146,274,826	106,897,938	388,689	376,858	565,678	513,377	534,323	256,923	177,287	793,517	714,506	501,791
chr9	140,273,252	80,518,518	297,208	295,644	434,652	384,594	413,264	214,414	121,451	575,041	511,923	338,889
chr10	135,374,737	93,075,998	332,059	332,026	480,970	428,533	450,665	195,541	142,553	691,889	620,383	418,648
chr11	134,452,384	92,310,050	315,792	313,337	447,758	405,996	426,195	220,131	116,623	648,986	546,936	381,892
chr12	132,349,534	97,193,825	320,773	332,003	466,816	426,115	450,904	262,120	143,781	661,486	561,514	365,844
chr13	114,142,980	70,684,904	242,844	241,450	366,737	333,965	352,990	187,643	119,672	526,443	474,900	326,162
chr14	106,368,585	64,504,906	219,620	226,685	313,562	284,567	308,687	177,637	94,913	461,982	398,533	276,116
chr15	100,338,915	56,493,900	196,969	190,789	271,299	247,355	261,973	145,614	80,995	379,143	327,720	236,253
chr16	88,827,254	49,858,432	198,075	190,344	269,525	247,237	260,878	159,158	86,384	363,871	310,499	223,654
chr17	78,774,742	50,696,851	165,695	170,816	248,630	222,955	231,674	96,969	67,633	327,025	306,148	210,212
chr18	76,117,153	57,568,302	205,893	208,815	301,583	267,285	282,747	121,829	97,290	433,232	376,497	259,327
chr19	63,811,651	29,695,141	107,273	110,796	145,318	133,290	141,827	89,916	44,932	213,185	191,953	137,617
chr20	62,435,964	44,682,113	156,021	160,191	227,739	205,784	214,291	113,235	53,552	322,647	271,899	183,520
chr21	46,944,323	24,441,662	95,257	96,903	140,852	125,625	134,921	76,251	40,093	210,164	188,300	125,382
chr22	49,691,432	21,290,071	78,197	82,028	110,440	98,838	105,023	49,288	33,617	155,687	145,205	95,163
chrX	154,913,754	105,036,574*	153,012	219,519	298,178	279,321	289,400	138,828	34,555	373,464	346,978	156,010
Sum	3,022,646,526	2,036,936,908	6,844,702	6,972,829	10,071,316	9,088,116	9,597,638	4,904,657	2,978,829	14,182,693	12,360,836	8,519,869

The length of chromosomes, the number of called positions, and the number of SNPs in each taxon for each chromosome are shown.
*This number excludes counts from pseudoautosomal regions and male heterozygous positions of a total of 105,251,391 called positions.