SI Appendix

Evidence that the Rate of Strong Selective Sweeps Increases with Population Size in the Great Apes

Kiwoong Nam[a], Kasper Munch[a], Thomas Mailund[a], Alexander Nater[b,c], Maja Greminger[b,c], Michael Krützen[c], Tomàs Marquès-Bonet[d,e,f], Mikkel Heide Schierup[a,g]

[a]Bioinformatics Research Centre, Aarhus University, DK-8000 Aarhus C, Denmark.
[b]Department of Evolutionary Biology and Environmental Studies, University of Zurich, 8057 Zurich, Switzerland
[c]Anthropological Institute and Museum, University of Zurich, 8057 Zurich, Switzerland
[d]Institut de Biologia Evolutiva, (CSIC-Universitat Pompeu Fabra), PRBB, Doctor Aiguader 88, Barcelona, Catalonia 08003, Spain.
[e]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia 08010, Spain.
[f]Centro Nacional de Analisis Genomico (CRG-CNAG), Baldiri Reixach 4-8, 08023 Barcelona, Spain
[g]Department of Bioscience, Aarhus University, DK-8000 Aarhus C, Denmark.

Supplementary Tables

Table S1. **The information of taxa used in this study**

Common name	Scientific name	Number of individuals
Human	*Homo sapiens*	9
Bonobo	*Pan paniscus*	13
Central Chimpanzee	*Pan troglodytes troglodytes*	4
Eastern Chimpanzee	*Pan troglodytes schweinfurthii*	6
Western Chimpanzee	*Pan troglodytes verus*	5
Nigeria Cameroon Chimpanzee	*Pan troglodytes ellioti*	10
Eastern Lowland Gorilla	*Gorilla beringei graueri*	3
Western Lowland Gorilla	*Gorilla gorilla gorilla*	27
Sumatran Orangutan	*Pongo abelii*	5
Bornean Orangutan	*Pongo pygmaeus*	5

Table S2. **The information for each chromosome.** The length of chromosomes, the number of called positions and the number of SNPs in each taxon for each chromosome.

Chromo-some	Chromosome length	Called	Number of SNP									
			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
Sum	2,867,732,772	1,931,685,517	6,691,690	6,753,310	9,773,138	8,808,795	9,308,238	4,765,829	2,978,829	13,809,229	12,013,858	8,363,859

Table S3. **Rate of beneficial mutations and rate of positive selection in the simulated data** For different selection coefficient (0.01 and 0.02), we compared the number of positively selected sites in genes between simulations in which total numbers of arisen beneficial mutations are the same but *Ns* differs by a factor of two.

s	*N*	proportion beneficial mutation	no. of beneficial mutation expected	no. of fixed beneficial mutations expected (2*s)*	no. of positively selected sites
0.01	2000	0.0005	500	10	9.21
	1000	0.001	500	10	9.14
	2000	0.001	1000	20	18.19
	1000	0.002	1000	20	18.07
0.02	2000	0.0005	500	20	18.68
	1000	0.001	500	20	18.41
	2000	0.001	1000	40	36.27
	1000	0.002	1000	40	36.72

Supplementary Figures

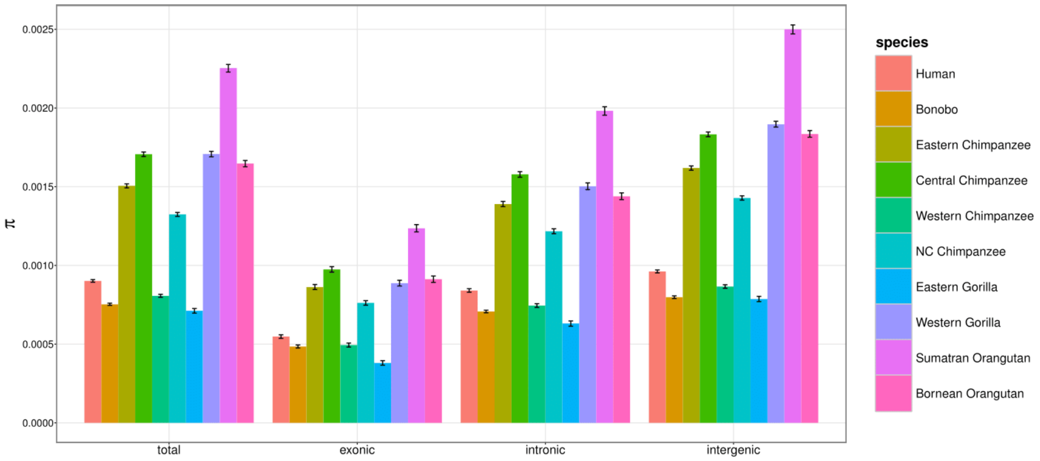

Fig. S1. **Diversity patterns** The diversity levels of total genomes, exons, introns, and intergenic regions calculated from each species.

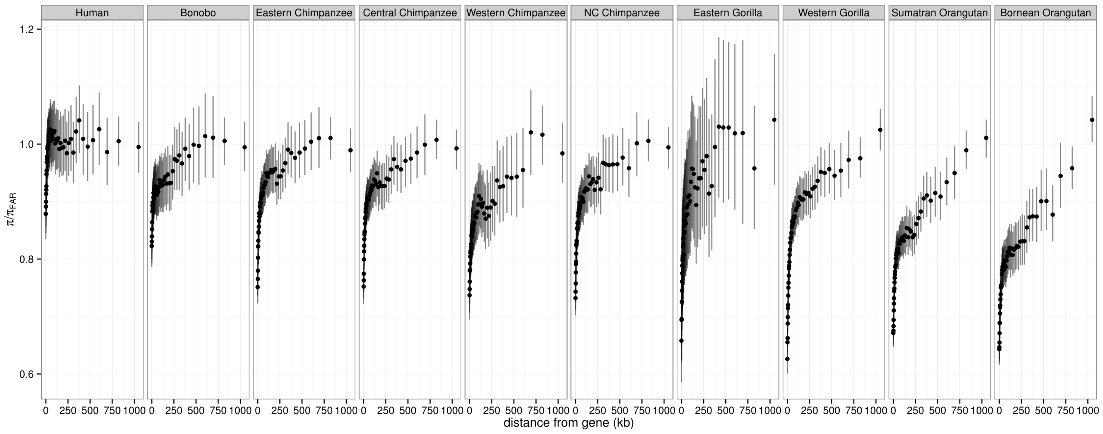

Fig. S2. **The reduced diversity near gene** The plots show a relationship between nucleotide diversity, π, and physical distance from the nearest genes. The diversity is normalized to the diversity far away from genes (> 823 kb). The error bars indicate 95% of the confidence intervals calculated from bootstrapping with 1,000 replicates resampled from 1mb windows.

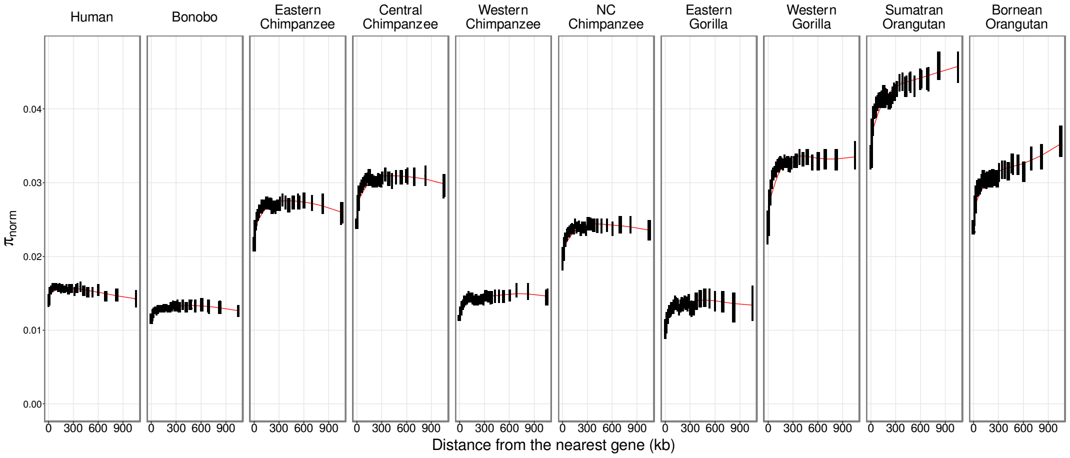

Fig. S3. **The normalized diversity according to the distance from gene** The plot shows a relationship between normalized π with the divergence between humans and macaques and physical distance from the nearest genes. The error bars indicate 95% of the confidence intervals calculated from bootstrapping with 1,000 replicates resampled from 1mb windows. The red bars indicate the non-linear regression curves generated by smooth spline (df=4).

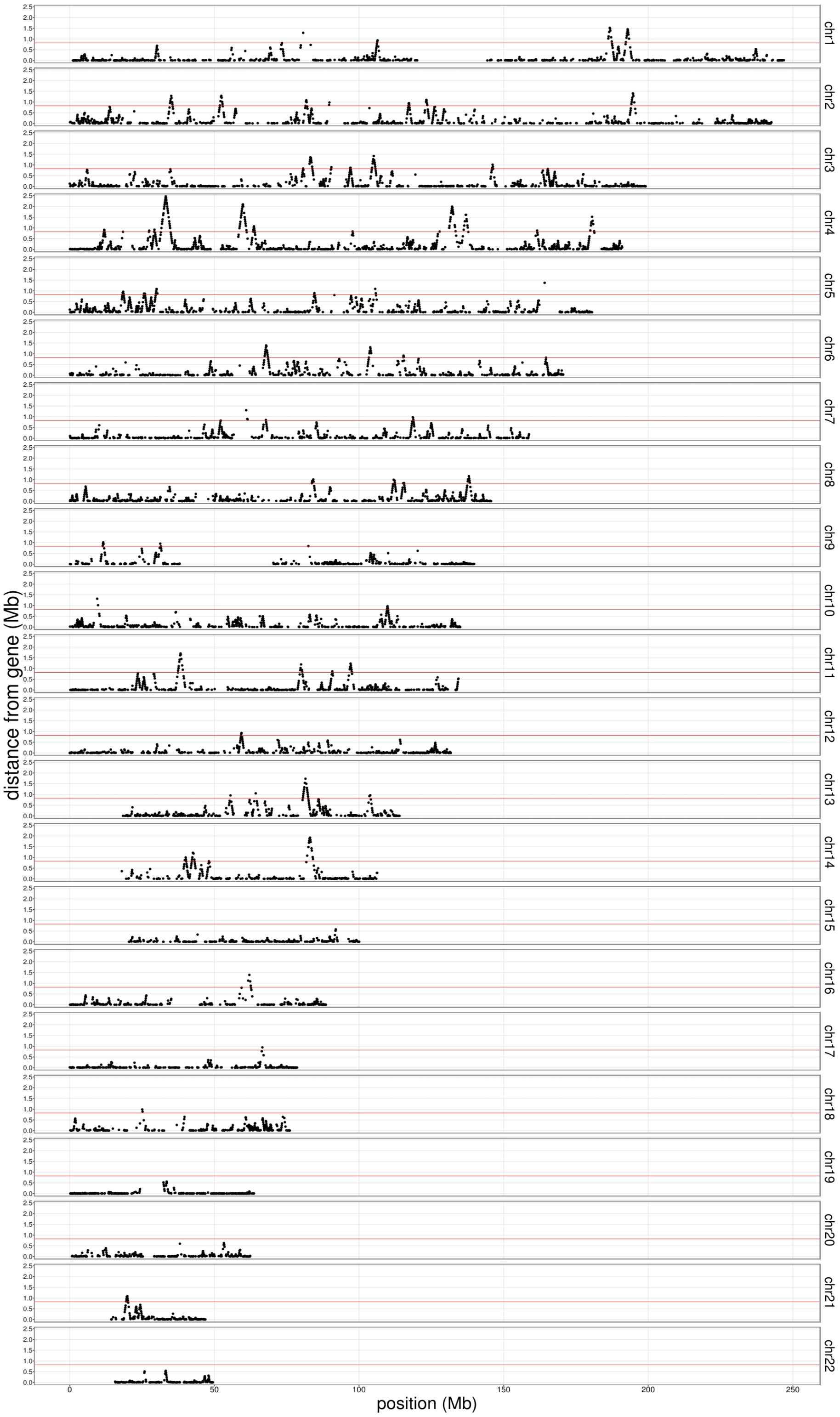

Fig. S4. **The spatial distribution of distance from genes**. The x axis represents the coordinates of each chromosome and y axis is the distance from the nearest genes (Mb). The red bars indicate the criterion on distance fro m genes used to identify gene-deserts.

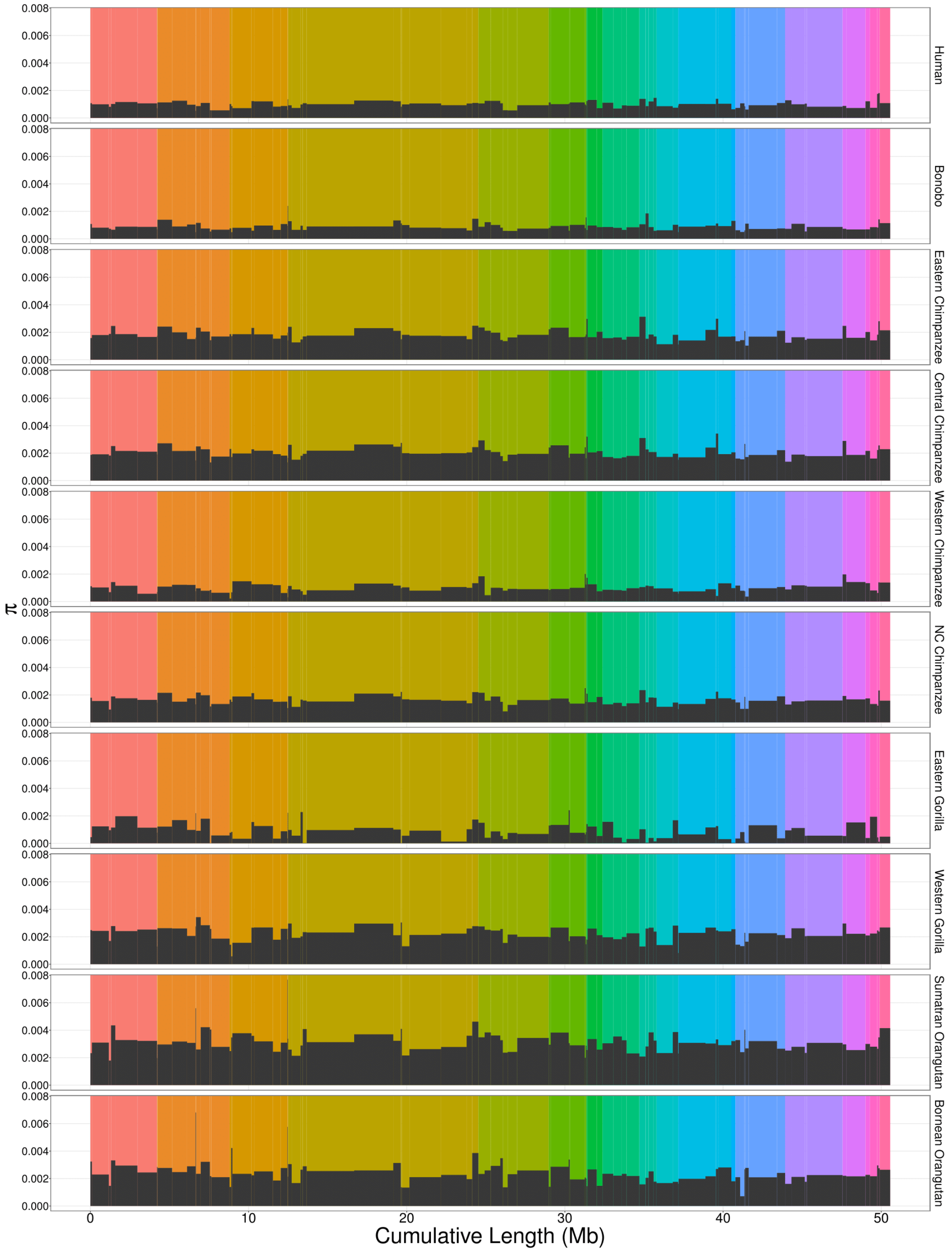

Fig. S5. **The nucleotide diversity of genomic loci far from genes**.

The width of each block represents the size of genomic loci far from genes (823 kb) and each background color shows chromosomes where these regions are found.

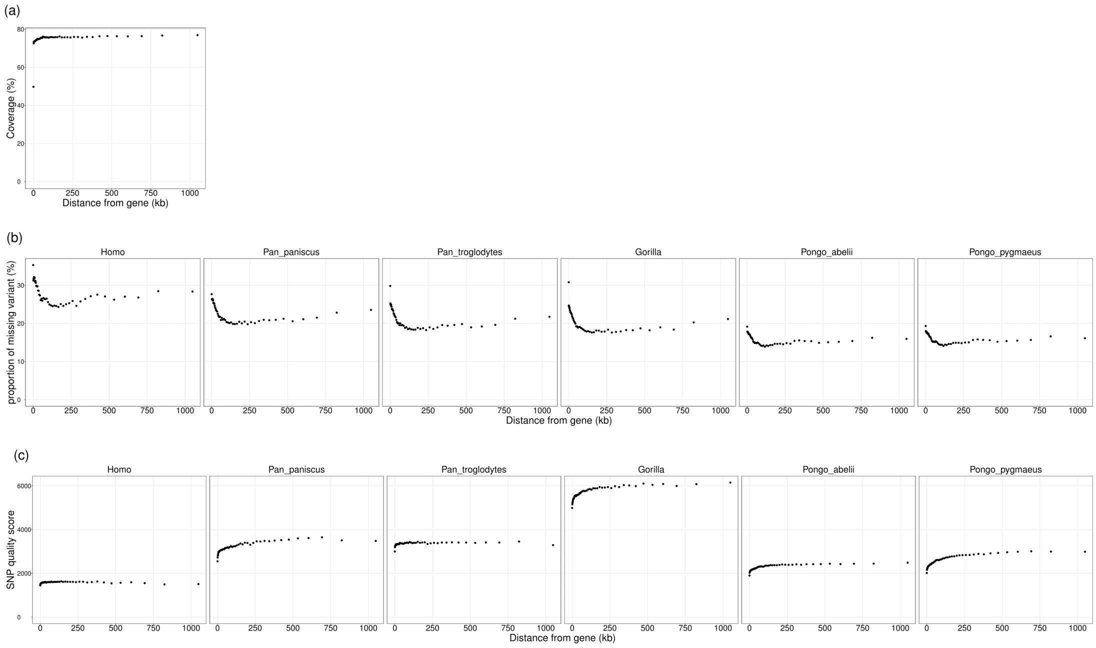

Fig. S6. **SNP quality according to the distance from genes**

(a) proportion of covered sequences after removing masked positions from at least one species, (b) the proportion of filtered out variant (c) and the SNP quality score.

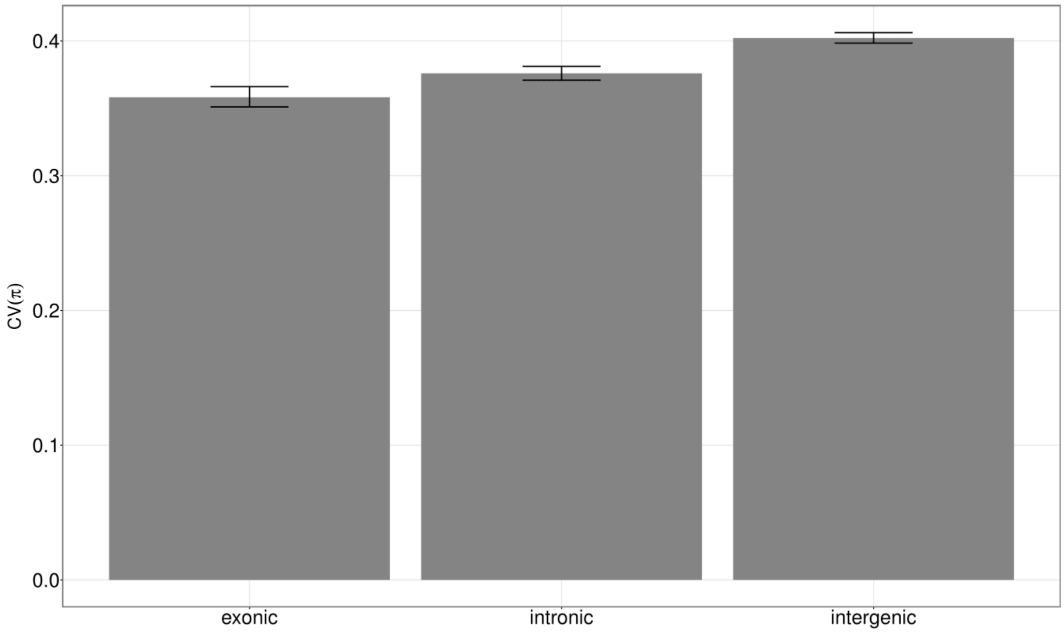

Fig. S7. **Heterogeneity of diversity according to annotation** The coefficient of variance of π in exonic, intronic, and intergenic sequences are shown. The error bars indicate 95% confidence intervals calculated from bootstrapping with 1,000 replicates resampled from 1Mb windows.

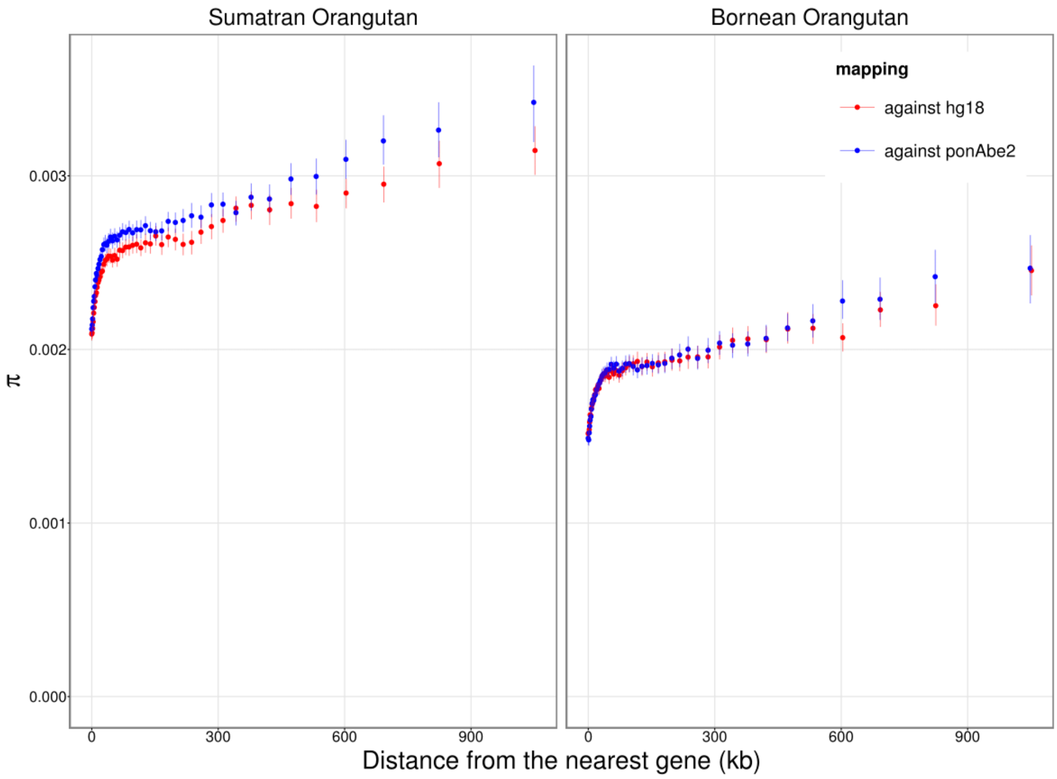

Fig. S8. **The diversity pattern away from genes in orangutans with different reference mappings.** Red points show estimated pi for mapping against hg18 (human), identical to results from Figure 1, blue points show estimates for mapping against ponabe2 (orangutan).

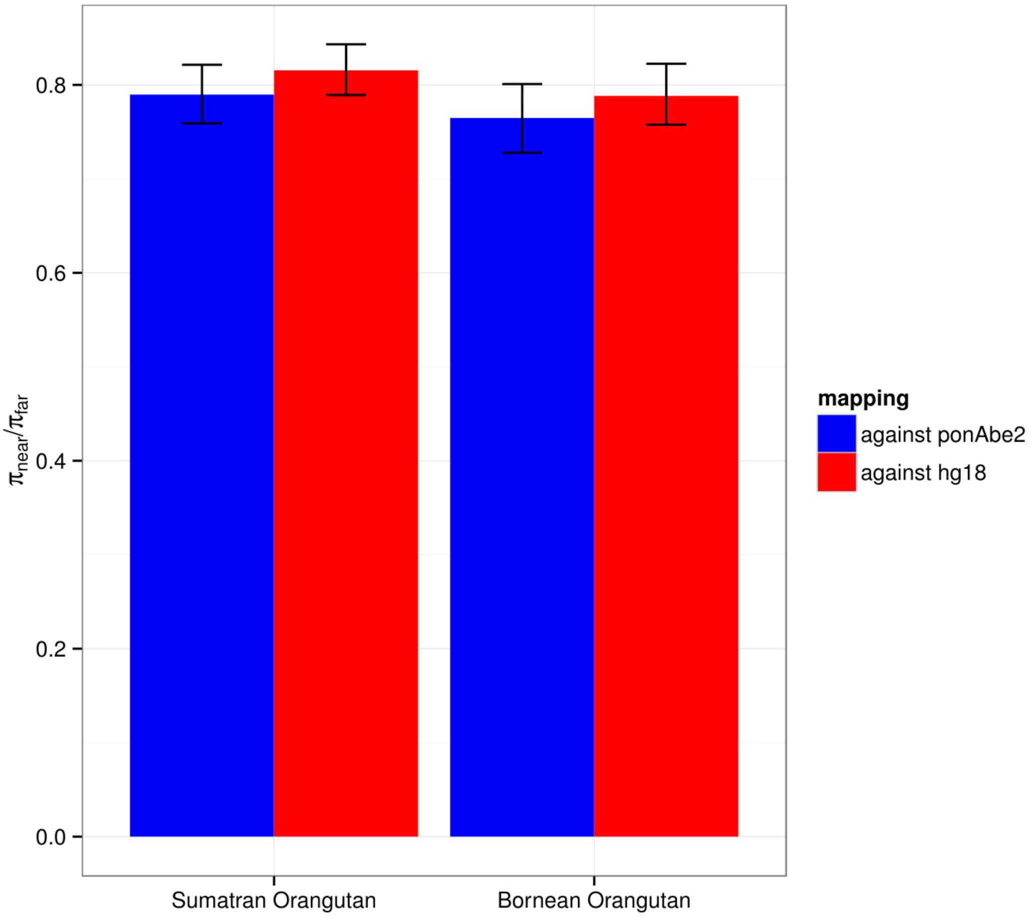

Fig. S9. **Reduction in diversity levels near genes in orangutans** The diversity ratio of gene deserts (in which distance from genes is larger than 823 kb) to the rest of intergenic regions in orangutans, calculated from the mapping against human (hg18) and orangutan (PonAbe2) reference genome. The error bars indicate 95% confidence intervals calculated from 1,000 times of bootstrapping.

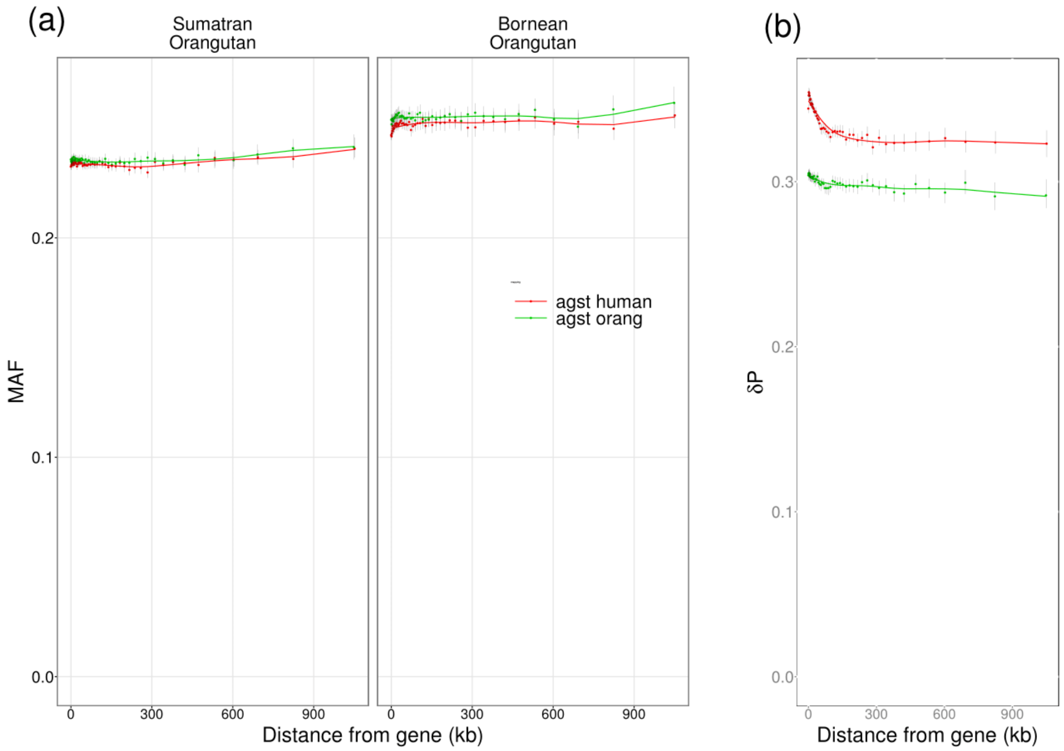

Fig. S10. **The diversity pattern from different mappings**. The relationship between distance from genes and the (a) diversity levels and (b) population differentiation, calculated from variants of orangutans identified by the mapping against human (red) and orangutan (green) reference genomes.

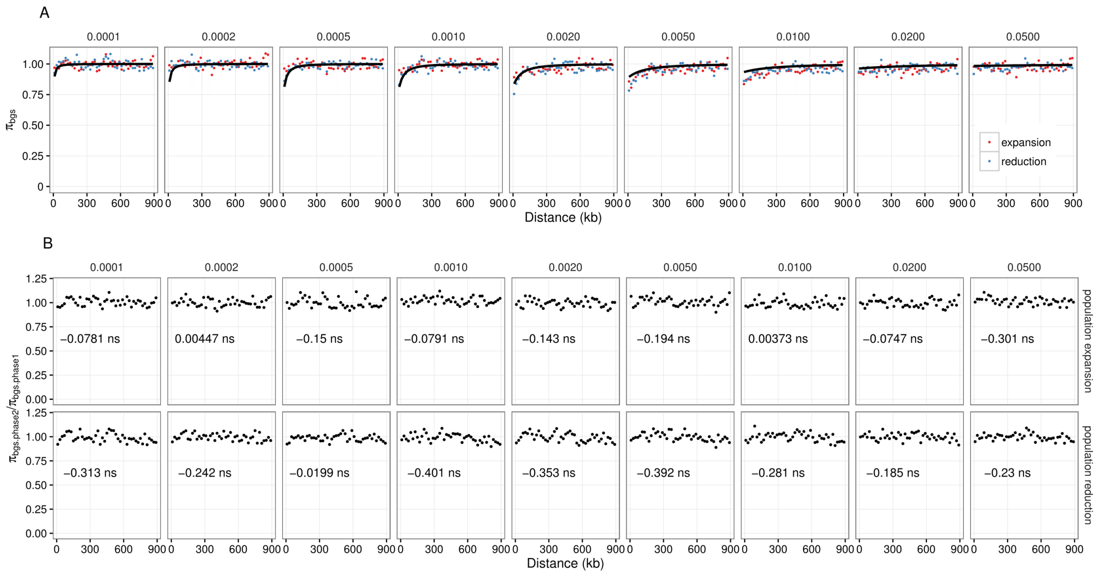

Fig. S11. **The effect of changes in *N* in background selection** The diversity pattern in the sequences flanking the genic region under evolutionary constraint (selection coefficients ranging from 0.0001 to 0.05) , when a population experiences a recent expansion (from N = 1,000 to N = 2,000) or reduction (from N = 2,000 to N = 1,000) 100 generations ago. For each set of parameters, we performed 1,000 independent simulations and report the average π. (a) The relationship between the distance from genes and π_{BGS}, diversity reduction due to background selection. The black lines indicate theoretical predictions of reduction in diversity by background selection (Durrett, 2008). Red and blue points represent population expansion and reduction, respectively. (b) The relationship between the distance from genes and the ratio of $\pi_{BGS.phase2}$ (π_{BGS} after changes in N) to $\pi_{BGS.phase1}$ (π_{BGS} before changes in N). The Spearman's correlation coefficient and the significance are shown in each panel (***, **, *, and ns indicate Bonferroni-corrected p-values with < 0.001, < 0.01, < 0.05, and ≥ 0.05, respectively).

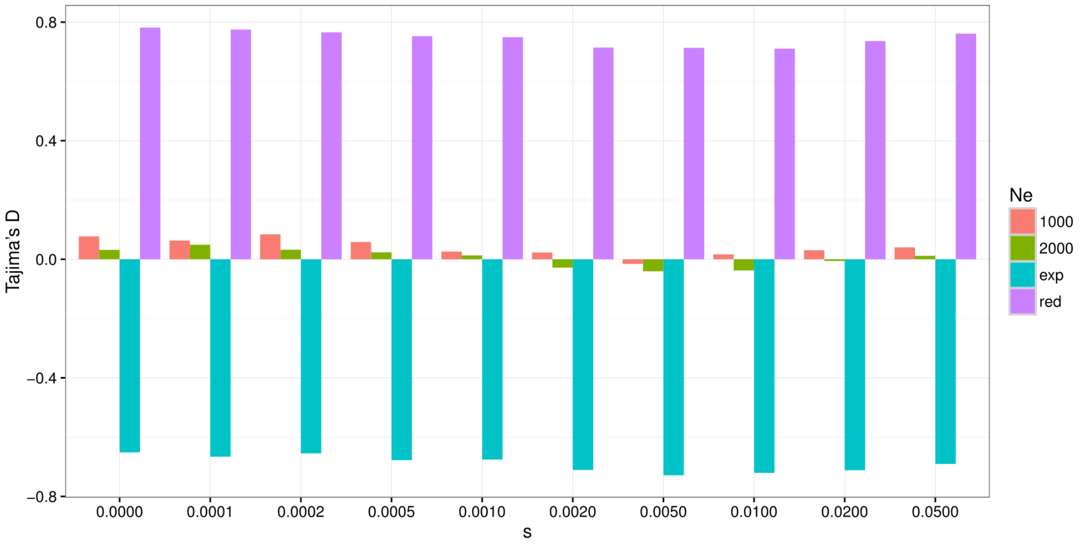

Fig. S12. **The skewness of site frequency spectrum** Tajima's D is calculated from different selection coefficients of deleterious mutations (ranging from 0 to 0.05), when a population size is constant (*N* = 1,000 or *N* = 2,000) or a population experiences a sudden expansion (*N* = 1,000 to *N* = 2,000) or reduction (*N* = 2,000 to *N* = 1,000) 100 generations ago.

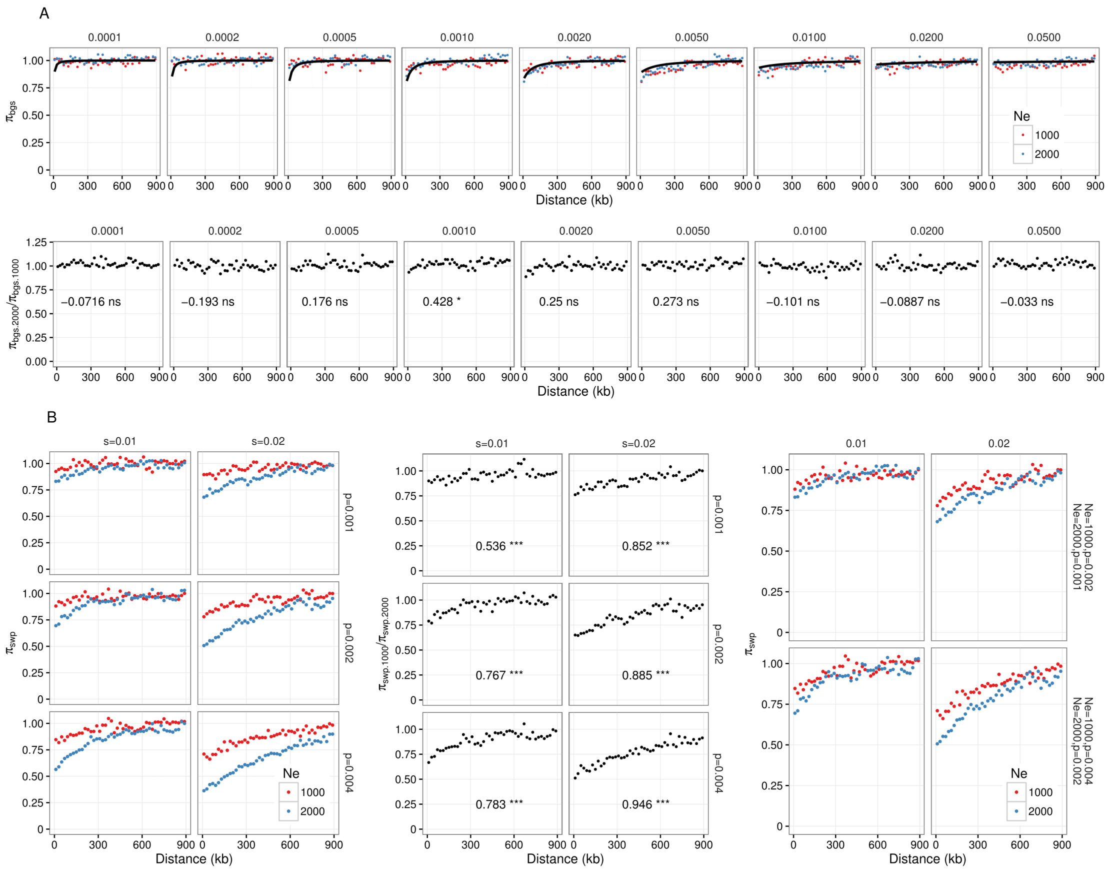

Fig. S13. **Simulations of diversity reduction by selective sweeps and background selection** The reduction in diversity due to background selection, π_{BGS}, and selective sweeps, π_{SWP}, is calculated from 900 kb intergenic sequences that are flanked by 100 kb genic sequences. The mutation rate is 1.2×10^{-8} per site per generation. (a) The upper panel shows the relationship between the distance from gene and π_{BGS} with selection coefficient ranging from 0.0001 to 0.05. The lower panel shows the relationship between the distance from genes and the ratio of π_{BGS} when $N = 2{,}000$ to π_{BGS} to $N = 1{,}000$. (b) The left panel shows the relationship between the distance from genes and π_{SWP}, when the proportion of beneficial mutation ranges 0.001 to 0.004 and selection coefficient is 0.01 or 0.02. The middle panel shows the relationship between the distance from genes and the ratio of π_{SWP} when $N = 2{,}000$ to π_{SWP} to $N = 1{,}000$. The right panel shows π_{SWP} as a function of distance from genes when the number of beneficial mutations per generation is the same but N differs by a factor of two.