Supporting Information

Ségurel et al. 10.1073/pnas.1210603109

SI Methods

SI Note S1. Genome-Wide Estimate of Exonic Diversity in 585-bp Sliding Windows. We used the primate orthologous exon database (http://giladlab.uchicago.edu/orthoExon/) to obtain the coordinates of unique, nonoverlapping orthologous exons between human and rhesus macaque genomes. This dataset represents about 29% of human exons and does not include the MHC region. We estimated human diversity in the YRI sample (individuals of Yoruban ancestry) from the 1000 Genome Pilot project (1). For each exon, we calculated the mean diversity in humans in a 585-bp sliding window. Windows with a gap greater than 10% of the sequence were excluded. We then considered the fraction of windows with diversity levels higher than in *ABO* exon 7 (where the diversity was 0.0058) and obtained a value of 0.08%. To control for differences in mutation rate, we verified that a similarly low *P* value was obtained for the ratio of diversity to divergence between human and rhesus macaque (this ratio is 0.16 in *ABO* exon 7; *P* value = 0.2%). The *P* values obtained in this way are likely to be overestimates, as some of the more extreme windows are probably the result of errors in SNP calls or in read mapping.

SI Note S2. Summary of Previous Studies About ABO Evolution. Phylogenetic approaches have been used to show that, in species of gibbons and siamangs (2) as well as in species of macaques (3–5), *ABO* haplotypes cluster by allelic type rather than by species. Additionally, in gibbons/siamangs, coalescence times between the A and B alleles were estimated to be older than split time between the *Hylobates* and *Symphalangus* genera. The case of baboon has led to more controversy, with Doxiadis et al. (3) finding that macaques and baboons shared the same trans-species polymorphism, whereas Kermarrec et al. (4) and Noda et al. (5) reported that olive and yellow baboons form a monophyletic cluster compared with macaques. Thus, among recently diverged subspecies or species (i.e., among gibbons/siamangs and among macaques), there is agreement that the data support a trans-species polymorphism, but not for more distant species.

Ref. 6 is the only study to have favored the trans-species polymorphism hypothesis among other hominoids. The authors' conclusion was based on the phylogenetic tree of two A chimpanzees, five B gorillas, and one A orangutan, which used 405-bp sequences of exon 7. When we compared the sequences for chimpanzee, gorilla, and orangutan obtained by different authors (4, 6–8), as well as by this study, however, we found that the dataset from ref. 6 appears to contain sequencing errors (Dataset S2).

All other studies concluded that ABO sharing results from convergent evolution. For example, ref. 9 shows a phylogenetic tree matching the species tree for human, chimpanzee, gorilla, orangutan, baboon, and macaque. To obtain this tree, however, the authors assumed that all nucleotides shared between species were due to parallel substitutions. Using a neighbor-joining method in ref. 4, the authors obtained a tree supporting the independent evolution of the B allele in humans, gorilla, and orangutan, but with insignificant bootstrap support.

In addition, in refs. 9 and 10, estimated divergence times are shown among human *ABO* alleles to be between 2.7 and 4.7 million years (Myr) (based on 405 bp in exon 7) and 4.8 Myr (based on ~2.5 kb covering exons 6 and 7), respectively, and these numbers are interpreted as evidence that the B allele originated more recently than the human–chimpanzee split. In ref. 11, using data from intron 5 and intron 6 in three humans, two chimpanzees, and two gorillas, the authors did not find any shared polymorphism between species and, on that basis, argued against a trans-species polymorphism. Similarly, in refs. 3 and 7, the existence of 10 substitutions differentiating Old World monkeys from hominoids is considered as evidence that the A and B alleles of Old World monkeys appeared independently from those of hominoids. However, these arguments implicitly ignore the effect of recombination, which restricts the presence of shared SNPs to only a small region around the functional sites, allowing for the accumulation of substitutions between Old World monkeys and hominoids outside of this region (as pointed out by refs. 12 and 13) and leading to a decreased estimate of the divergence time of *ABO* alleles if too large a window is considered (14).

SI Note S3. Identification of the A, B, and O Haplotypes. In humans, we used resequencing data from exon 6 and exon 7 to classify haplotypes in three main lineages: A, B, or O. Considering the negative strand on chromosome 9 in Build hg19, A haplotypes are defined as those carrying the A-specific alleles (codons) at the two functional sites: C at position 136.131.312 (leucine at amino acid 266) and G at position 136.131.315 (glycine at amino acid 268). B haplotypes are defined as those carrying an A (methionine) and a C (alanine) at these positions. O haplotypes are defined as those carrying the 1-bp deletion between position 136.132.908 and 136.132.907, and almost always carry the A-specific alleles at the two functional sites. We found two recombinant haplotypes between the two functional sites (at frequency 0.6% in the combined human population samples) and excluded them from subsequent analyses.

In other species, A and B haplotypes were aligned to humans using the Phred-Phrap-Consed package (15) and analogously defined by their amino acids at positions 266 and 268. O haplotypes were considered only when defined on the basis of their presence in individuals with an O phenotype or if frame-shift mutations had been clearly identified. In chimpanzee, there are two such O haplotypes (16): Odel, which carries a 9-bp deletion in exon 7 and Ox, which carries a G at nucleotide position 791. An unambiguous O haplotype has been found in macaques and in baboons as well, even though the causal mutation has not been identified (3, 17). Finally, in gibbons, a 7-bp deletion found in exon 7 was considered to define a null allele (2). In baboons (17), an O haplotype (<5%) has been predicted by comparing the expected phenotype and genotypes of individuals; because this result could arise from phenotypic errors or alleles associated with low expression of antigens, however, we excluded this allele from consideration.

SI Note S4. Estimating the Length of the Segment Carrying Various Signals of a Trans-species Polymorphism. For brevity, we assumed that salient parameters (recombination rate, generation time, and frequencies of the selected classes) remained the same in the two lineages as well as in the ancestral population.

First, we considered a sample of two haplotypes from a single species, one from each selected class, and estimated the expected length of the contiguous segment that does not coalesce (backward in time) before time T (also estimated ref. 12). This is the relevant length scale over which a trans-species polymorphism will leave a signal in *intraspecies* comparisons (such as those in Fig. 4 *A* and *B*). The expected length can be bound from below by considering the segment on one side of the selected site that experienced no recombination in either sample during T generations. The length of this segment follows an exponential distribution $x \sim Exp(2cT)$, where c is the recombination rate per

base pair, per generation (ref. 12). Thus, the expectation of the length of the two-sided shared segment is $1/cT$.

Assuming that only recombination events with haplotypes from the other selected class reduce the segment size, then $x \sim Exp(cT)$. This follows because, if allele A has frequency p and allele B frequency q ($p + q = 1$), then the distribution of lengths of the segment contiguous to allele A that did not experience recombination with the other class is $x_A \sim Exp(cTq)$ and similarly $x_B \sim Exp(cTp)$; therefore, the length of the segment to experience recombination in either background is $x = \min\{x_A, x_B\} \sim Exp(cT(q+p))$. Thus, the expectation of the length of the two-sided segment is $2/cT$. This calculation neglects the possibility of a more complex sequence of recombination events. For example, consider a recombination event with a haplotype from the same selected class at distance K of the selected site; we ignored the possibility that the recombination event is with a haplotype that previously recombined with the other selected class between the selected site and K. It can be shown that, if $T >> 2N_e$ generations, which is satisfied for the cases considered here, the contribution of these events is negligible. We note further that, throughout, we considered only one chromosome from a given selected class in a given species. This is appropriate because coalescent times within each of the selected classes should be small compared with the species' split times (i.e., because $T >> 2N_e$).

Next, we considered a sample of two chromosomes, each from a different species. We asked about the expected length of the contiguous segment in which a pair in the same selected class coalesces before a pair from different selected classes. In such a segment, divergence levels for pairs from the same selected class are expected to be less than or equal to the levels between pairs from different selected class. This is therefore the relevant length scale for the interspecies comparisons in Fig. 4 *A* and *B*. We assumed, without loss of generality, that the pair from different selected classes is an A from species 1 and a B from species 2, and that the pair from the same selected class consists of the same A from species 1 and an A from species 2. By assumption, at the selected site, the A pair coalesces (in the ancestral population) before the A and B pair. This coalescent order would also hold for the segment contiguous to the selected site that experienced no recombination in any of the three lineages before the A pair coalesced. Because recombination events within the same selected class are unlikely to affect the expected order of coalescence events (assuming $T >> 2N_e$ and $2N_a$, where N_a is the effective size of the ancestral population), we derived the expected length of the segment by requiring that none of the three lineages recombine with haplotypes from the other selected class. The coalescence time of the A samples in the ancestral population follows an exponential distribution $t \sim Exp\{1/2N_a p\}$. Conditional on t, the length of the one-sided segment of interest is exponentially distributed with $x = min\{x_1^A, x_2^A, x_2^B\} \sim Exp\{c(T+t)(2q+p)\}$. The two-sided expected length is therefore

$$E(x) = 2\int_0^\infty E(x|t)f(t)dt = 2\int_0^\infty \frac{1}{c(2-p)(T+t)}\frac{e^{-t/2N_a p}}{2N_a p}dt$$
$$= \frac{2}{c(2-p)}\int_0^\infty \frac{e^{-z}}{(T+(2N_a p)z)}dz.$$

When $T > 2N_a$, this length increases with p and is therefore bound by the values at $P = 0$ and $P = 1$. In Fig. 2*C* and Fig. S2, we plot this equation for $P = 0.5$.

Last, we considered a sample of four haplotypes, one from each selected class in each of the two species. We asked about the expected length of the contiguous segment in which the coalescent order of the four haplotypes clusters by selected class (as opposed to by species or neither). This is the relevant scale at which to expect a tree that clusters by selected class and not by species and the scale on which we would expect to see shared polymorphisms between the two species (besides the selected site). Thus, it is the relevant scale for Fig. 3. In this case, we were interested in the time at which both A lineages and both B lineages have coalesced at the selected site, but A and B lineages have not. This time is given by $t = max\{t_A, t_B\}$, where $t_A \sim Exp\{1/2N_a p\}$ and $t_B \sim Exp\{1/2N_a q\}$. The segment (contiguous to the selected site) in which none of the four lineages recombined with the other selected class will have the same tree topology as the selected site, i.e., in it, the haplotypes will group by selected class. Conditional on t, the length of the one-sided segment is exponentially distributed with $x = min\{x_1^A, x_2^A, x_1^B, x_2^B\} \sim Exp\{2c(T+t)\}$. In turn, the two-sided expected segment length is

$$E(x) = 2\int_0^\infty E(x|t)f(t)dt$$
$$= \frac{1}{c}\int_0^\infty \frac{1}{T+t}\left[\left(1-e^{-t/2N_a p}\right)\frac{e^{-t/2N_a q}}{2N_a q} + \left(1-e^{-t/2N_a q}\right)\frac{e^{-t/2N_a p}}{2N_a p}\right]dt.$$

When $T > 2N_a$, this length is maximized for $P = q = 0.5$ and minimized for p or $q = 0$. This equation yields an expectation similar to the one plotted in Fig. 2*C* (see also Fig. S1).

Within this segment, the number of shared SNPs will depend on the age of the balanced polymorphism. On average, there should be more differences between selected backgrounds than expected from the heterozygosity in the ancestral population (because coalescent times between haplotypes from different selected classes are older than average). However, because of recombination in the ancestral population, the contiguous segment becomes shorter and shorter with increasing age of the balanced polymorphism. For a very old polymorphism, the density of shared SNPs will therefore increase with decreasing distance to the selected site.

SI Note S5. Tree Representations of the Data. Based on our expectations for the expected length of the segment in which we expect haplotypes to cluster by selected class (see above), we used a sequence of 200 and 300 bp to generate the trees of Old World monkeys and hominoids, respectively. Trees were based on nucleotides 649–948 and 699–898 in hominoids and Old World monkeys, respectively, starting at the first ATG codon in the mRNA sequence (GenBank NM_020469.2). In humans, we excluded rare recombinants that carried the deletion in exon 6 together with B-specific alleles at the two functional sites (1.3% in the combined human population samples) as well as haplotypes with recombination events between the A and the B background within the window considered (0.3%).

Phylogenetic trees were obtained with the program MrBayes version 3.1.2 (18). We used a general time reversible model of molecular evolution with a proportion of invariable sites, and a γ-shaped distribution of rates across sites; other parameters were left at default settings. Using the same rates of substitutions for all nucleotides, instead, did not affect our results. The program was run three times, each time with two simultaneous Markov chains running for 3,000,000 generations, discarding the initial 25% of the trees as burn-in. The three replicates gave identical consensus trees, with a tight range of credibility values for the clade of B alleles. Trees were plotted with TreeView version 0.5.0 (19).

SI ACKNOWLEDGMENTS. The authors appreciate general support from Species Survival Plan coordinators and veterinary advisors (Gay Reinartz and Victoria Clyde, Bonobo SSP; Kay Backues, Common Chimpanzee SSP; Dan Wharton and Tom Meehan, Gorilla SSP; and Lori Perkins, Rita McManamon, and Chris Bonar, Orangutan SSP) and Kristin Lukas, Elizabeth Lonsdorf, and

Maureen Leahy. Contributing institutions are acknowledged for their provision of samples and animal identification, as described in ref. 20: *Species Survival Plan* Busch Gardens Tampa Bay, Tampa, FL; Columbus Zoo and Aquarium, Powell, OH; Denver Zoological Gardens, Denver, CO; Fort Wayne Children's Zoo, Fort Wayne, IN; Gladys Porter Zoo, Brownsville, TX; Honolulu Zoo, Honolulu, HI; Houston Zoo, Inc., Houston, TX; Jacksonville Zoo and Gardens, Jacksonville, FL; Lincoln Park Zoo, Chicago, IL; The Maryland Zoo in Baltimore, Baltimore, MD; Memphis Zoo, Memphis, TN; Miami Metrozoo, Miami, FL; Riverbanks Zoo and Garden, Columbia, SC; San Diego Zoo's Wild Animal Park, Escondido, CA; Sedgwick County Zoo, Wichita, KS; Smithsonian National Zoological Park, Washington, DC; Utah's Hogle Zoo, Salt Lake City, UT; and Zoo New England, Stoneham, MA.

1. 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061–1073.
2. Kitano T, Noda R, Takenaka O, Saitou N (2009) Relic of ancient recombinations in gibbon ABO blood group genes deciphered through phylogenetic network analysis. *Mol Phylogenet Evol* 51(3):465–471.
3. Doxiadis GG, et al. (1998) Characterization of the ABO blood group genes in macaques: Evidence for convergent evolution. *Tissue Antigens* 51(4 Pt 1):321–326.
4. Kermarrec N, Roubinet F, Apoil PA, Blancher A (1999) Comparison of allele O sequences of the human and non-human primate ABO system. *Immunogenetics* 49(6): 517–526.
5. Noda R, Kitano T, Takenaka O, Saitou N (2000) Evolution of the ABO blood group gene in Japanese macaque. *Genes Genet Syst* 75(3):141–147.
6. Martinko JM, Vincek V, Klein D, Klein J (1993) Primate ABO glycosyltransferases: Evidence for trans-species evolution. *Immunogenetics* 37(4):274–278.
7. Kominato Y, et al. (1992) Animal histo-blood group ABO genes. *Biochem Biophys Res Commun* 189(1):154–164.
8. Sumiyama K, Kitano T, Noda R, Ferrell RE, Saitou N (2000) Gene diversity of chimpanzee ABO blood group genes elucidated from exon 7 sequences. *Gene* 259(1-2):75–79.
9. Saitou N, Yamamoto F (1997) Evolution of primate ABO blood group genes and their homologous genes. *Mol Biol Evol* 14(4):399–411.
10. Calafell F, et al. (2008) Evolutionary dynamics of the human ABO gene. *Hum Genet* 124(2):123–135.
11. O'hUigin C, Sato A, Klein J (1997) Evidence for convergent evolution of A and B blood group antigens in primates. *Hum Genet* 101(2):141–148.
12. Wiuf C, Zhao K, Innan H, Nordborg M (2004) The probability and chromosomal extent of trans-specific polymorphism. *Genetics* 168(4):2363–2372.
13. Bubb KL, et al. (2006) Scan of human genome reveals no new Loci under ancient balancing selection. *Genetics* 173(4):2165–2177.
14. Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet* 2(4):e64.
15. Nickerson DA, Tobe VO, Taylor SL (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res* 25(14):2745–2751.
16. Moor-Jankowski J, Wiener AS, Rogers CM (1964) Human blood group factors in non-human primates. *Nature* 202:663–665.
17. Diamond DC, Fagoaga OR, Nehlsen-Cannarella SL, Bailey LL, Szalay AA (1997) Sequence comparison of baboon ABO histo-blood group alleles: Lesions found in O alleles differ between human and baboon. *Blood Cells Mol Dis* 23(2):242–251.
18. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12):1572–1574.
19. Page RD (1996) TreeView: An application to display phylogenetic trees on personal computers. *Comput Appl Biosci* 12(4):357–358.
20. Gamble KC, Moyse JA, Lovstad JN, Ober CB, Thompson EE (2010) Blood groups in the species survival plan((R)), European endangered species program, and managed in situ populations of bonobo (Pan paniscus), common chimpanzee (Pan troglodytes), gorilla (Gorilla ssp.), and orangutan (Pongo pygmaeus ssp.). *Zoo Biol*.

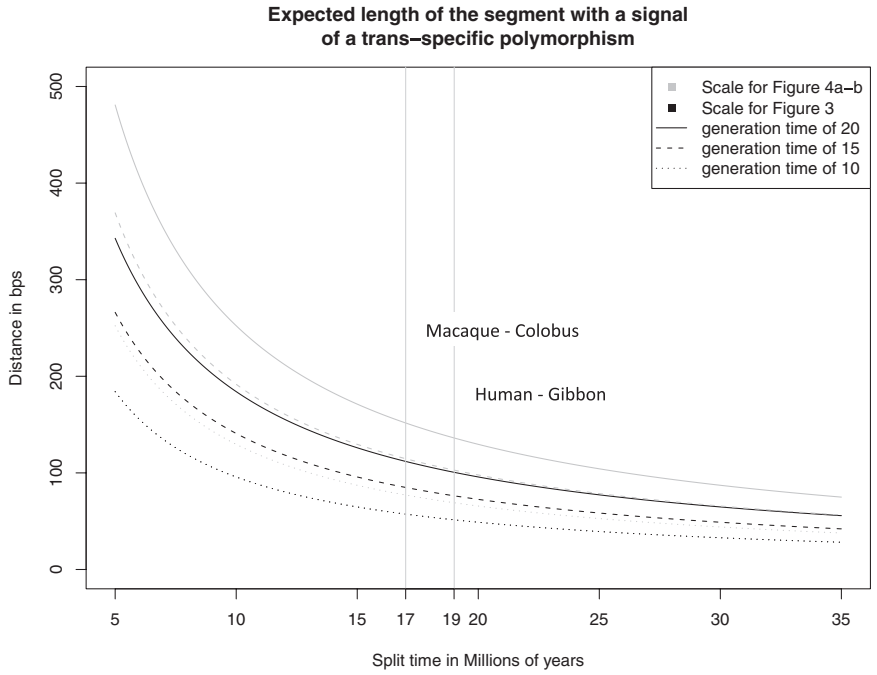

Fig. S1. Plot of the expected length of the segment in which a signal of a trans-species polymorphism should be detectable. Shown are the relevant scale for interspecies comparisons (Fig. 4 *A* and *B*) and the one for the tree (Fig. 3) (*Methods*). For other details, see Fig. 2*C* legend.

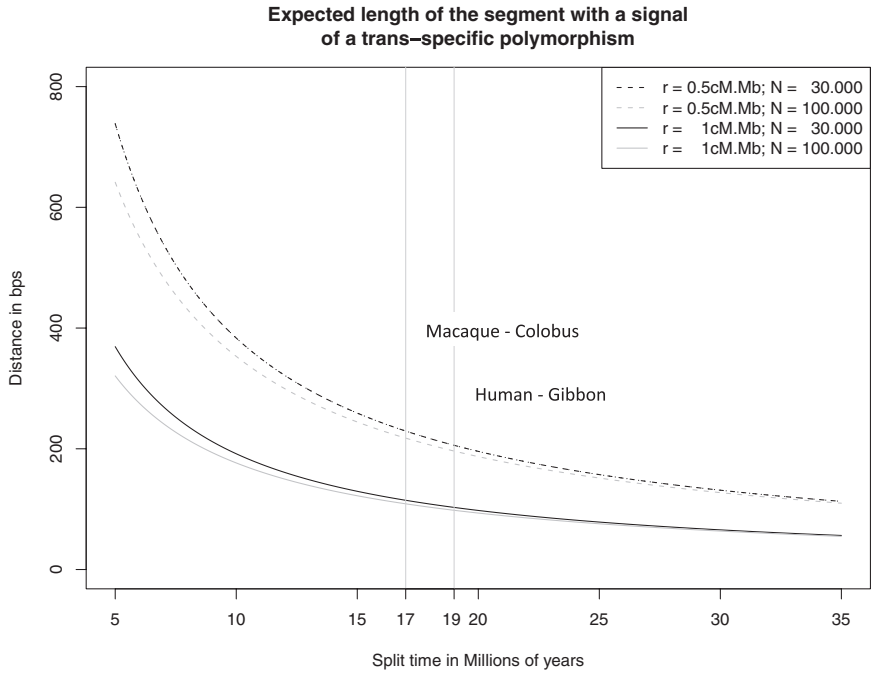

Fig. S2. Sensitivity of the expected segment length to the recombination rate and effective population size. The derivation is described in *Methods* and *SI Methods*, *SI Note S4*; *r* is the recombination rate per base pair per generation and *N* the effective population size. For the sensitivity to the generation time, see Fig. 2*C*. For other details, see Fig. 2*C* legend.

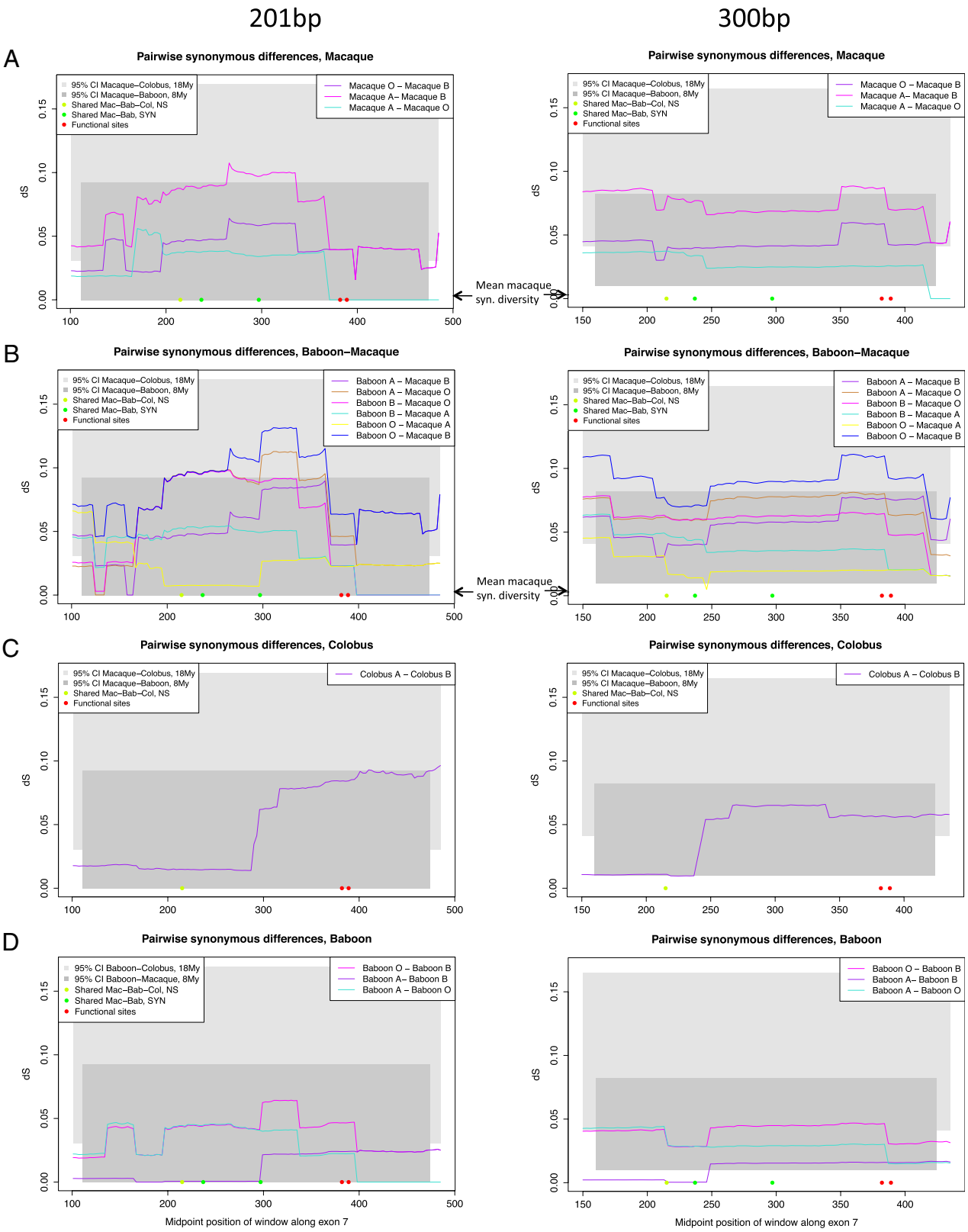

Fig. S3. Synonymous pairwise differences (dS) among *ABO* alleles in Old World monkeys: (*A*) in macaques, (*B*) in colobus monkeys, and (*C*) in baboons, for 201-bp (*Left*) and 300-bp (*Right*) sliding windows. For further details, see Fig. 4 *A* and *B* legend and *Methods*. We note that erosion of the ancestral segment by recombination can lead to divergence times that are much more recent than the origin of the ABO polymorphism (see main text and ref. 1). Thus, deeper divergence levels than expected are evidence for a trans-species polymorphism, but shallow divergence levels are not evidence against it.

1. Wiuf C, Zhao K, Innan H, Nordborg M (2004) The probability and chromosomal extent of trans-specific polymorphism. *Genetics* 168(4):2363–2372.

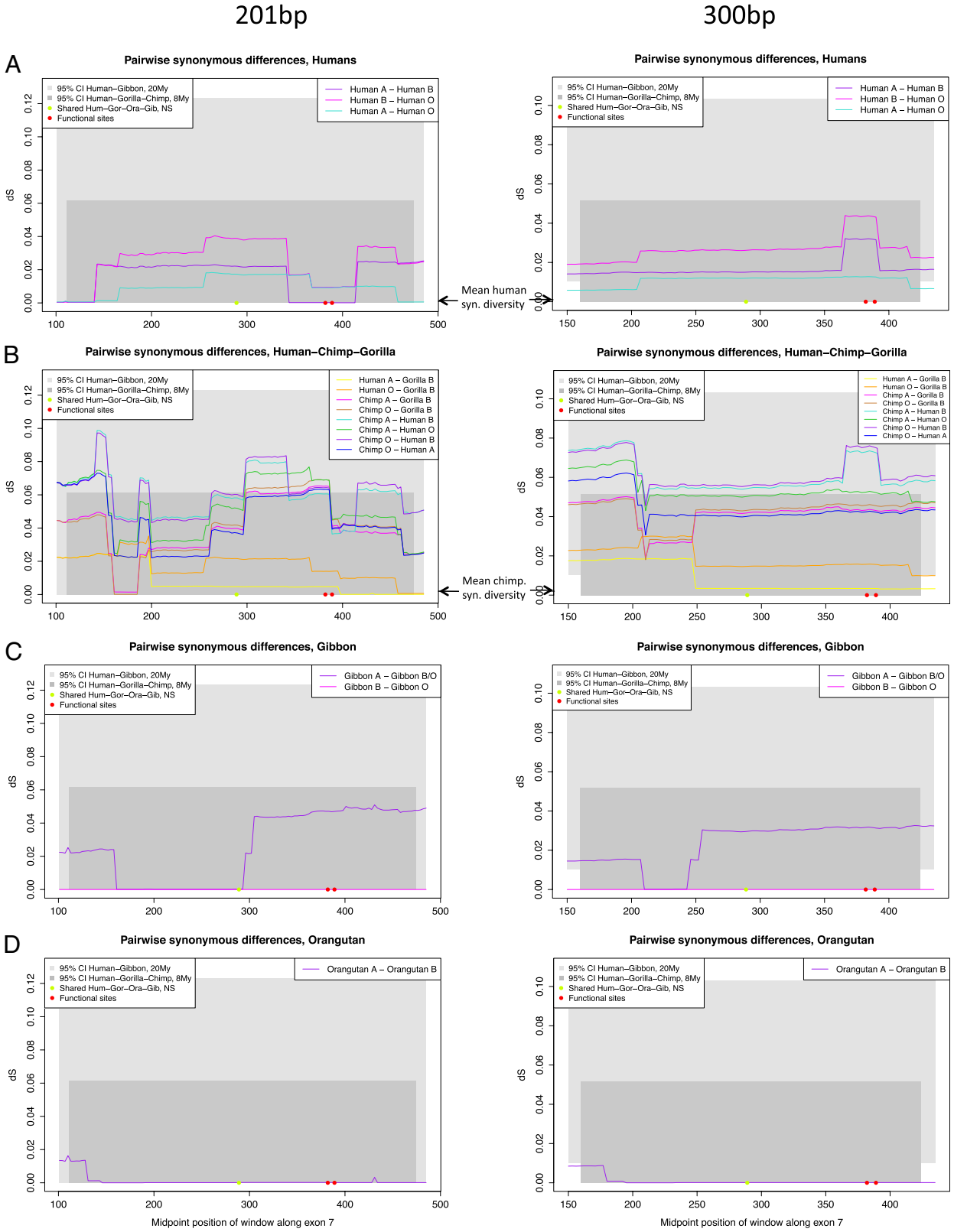

Fig. S4. Synonymous pairwise differences (dS) among *ABO* alleles in Hominoids: (*A*) in humans, (*B*) among human, chimpanzee, and gorilla, (*C*) in gibbons, and (*D*) in orangutans, for 201-bp (*Left*) and 300-bp (*Right*) sliding windows. Other details are as in Fig. S3.

Dataset S1. Phenotype and genotype studies of ABO in primates

Dataset S1

Included are all 40 non-human species with phenotype or genotype information for ABO in the literature (1). We did not consider phenotypic studies with fewer than five individuals (except for siamang, to bolster the evidence that they are monomorphic for B). When available, the observed phenotypes are given; the exception is when only the allele frequencies were reported, in which case the alleles are highlighted in blue. Lines in gray represent species without information about divergence times in ref. 1 (therefore not shown on the tree in Fig. 1) or cases in which ABO phenotypes are only given for a combination of species.

*In gorillas, phenotypic tests using red blood cells are unreliable, as gorillas have few (if any) antigens at the surface of these cells. Gorillas were therefore classified on the basis of their saliva or serum phenotype (2, 3) as well as based on their genotype, both of which consistently categorize them as monomorphic for the B allele.

1. Perelman P, et al. (2011) A molecular phylogeny of living primates. *PLoS Genet* 7(3):e1001342.
2. Socha WW, Wiener AS, Moor-Jankowski J, Mortelmans J (1973) Blood groups of mountain gorillas (Gorilla gorilla beringei). *J Med Primatol* 2(6):364–368.
3. Wiener AS, Moor-Jankowski J, Gordon EB (1971) Blood groups of gorillas. *Kriminalistik und forensische Wissenschaften* 6:31–43.

Dataset S2. Polymorphic sites in *ABO* exon 7 found in chimpanzee, gorilla, and orangutan

Dataset S2

Data from previous studies and this one are summarized. Gray boxes represent polymorphic sites, and orange boxes highlight the inconsistencies between ref. 1 and other studies. The nomenclature of each allele is as it was in the original study. Nucleotide positions were taken from the mRNA sequence, starting at the first ATG (nucleotide sequence NM_020469.2 in GenBank). As can be seen, the data from ref. 1 appear to be unreliable.

1. Martinko JM, Vincek V, Klein D, Klein J (1993) Primate ABO glycosyltransferases: Evidence for trans-species evolution. *Immunogenetics* 37(4):274–278.

Dataset S3. Table of substitutions and polymorphisms that require more than one mutation given the species tree: ((((((human, gorilla), chimpanzee), orangutan), (gibbon, siamang)), (((baboon, macaque), vervet), colobus)), (marmoset, howler))

Dataset S3

Parsimony was used to assign substitutions to lineages. Each row represents a unique haplotype; *N* denotes the number of chromosomes surveyed in the subspecies/species (*left* column). The two functional sites are shown in red. Identical polymorphisms shared between two or more species are shown in green; these could arise from mutations on an ancestral segment shared by descent or from recurrent mutations. For comparison, nonidentical shared polymorphisms, which reflect recurrent mutations, are shown in blue. Nucleotide positions start at the first ATG codon in the mRNA sequence (GenBank NM_020469.2). We did not include SNPs shared among species of macaques, species of colobus monkeys, or species of gibbons, because they are more recently diverged and therefore more likely to share SNPs because of neutral incomplete lineage sorting (1). The sequence data for yellow baboon are from ref. 2; they are not analyzed elsewhere because they are not a population sample.

**cis*-AB allele (having both A and B activity). We note that, consistent with old tMRCAs between *ABO* alleles, a SNP is shared between humans, gorillas, orangutans, and gibbons on the B background. This nonsynonymous SNP (amino acid 235) is known to affect ABO activity in humans (3), so may denote a subclass of B alleles.

1. Hudson RR, Coyne JA (2002) Mathematical consequences of the genealogical species concept. *Evolution* 56(8):1557–1565.
2. Kominato Y, et al. (1992) Animal histo-blood group ABO genes. *Biochem Biophys Res Commun* 189(1):154–164.
3. Yamamoto F, Hakomori S (1990) Sugar-nucleotide donor specificity of histo-blood group A and B transferases is based on amino acid substitutions. *J Biol Chem* 265(31):19257–19262.

Dataset S4. Number of polymorphic sites found per species, both for synonymous and nonsynonymous sites, along with the total number of polymorphic sites across the dataset (i.e., polymorphic in at least one species)

Dataset S4

N, number of individuals; *S*, number of polymorphic sites; NSyn, number of nonsynonymous polymorphic sites; Syn, number of synonymous polymorphic sites.

Dataset S5. Primer sets used for amplification and sequencing of *ABO* exon 7

Dataset S5

Degenerate positions are given according to the International Union of Pure and Applied Chemistry code (S = G/C; R = A/G; K = G/T).