The ABO blood group is a trans-species polymorphism in primates

Laure Ségurel[a,b,1,2], Emma E. Thompson[a,1], Timothée Flutre[a,c], Jessica Lovstad[a], Aarti Venkat[a], Susan W. Margulis[d,3], Jill Moyse[d], Steve Ross[d], Kathryn Gamble[d], Guy Sella[e], Carole Ober[a,2,4], and Molly Przeworski[a,b,f,2,4]

[a]Department of Human Genetics, [b]Howard Hughes Medical Institute, and [f]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637; [c]Department of Genetics and Plant Breeding, Institut National de la Recherche Agronomique, Unité de Recherche 1164, 78026 Versailles, France; [d]Lincoln Park Zoo, Chicago, IL 60614; and [e]Department of Ecology, Evolution and Behavior, The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem 91904, Israel

Edited by Marcus W. Feldman, Stanford University, Stanford, CA, and accepted by the Editorial Board September 12, 2012 (received for review June 22, 2012)

The ABO histo-blood group, the critical determinant of transfusion incompatibility, was the first genetic polymorphism discovered in humans. Remarkably, ABO antigens are also polymorphic in many other primates, with the same two amino acid changes responsible for A and B specificity in all species sequenced to date. Whether this recurrence of A and B antigens is the result of an ancient polymorphism maintained across species or due to numerous, more recent instances of convergent evolution has been debated for decades, with a current consensus in support of convergent evolution. We show instead that genetic variation data in humans and gibbons as well as in Old World monkeys are inconsistent with a model of convergent evolution and support the hypothesis of an ancient, multiallelic polymorphism of which some alleles are shared by descent among species. These results demonstrate that the A and B blood groups result from a trans-species polymorphism among distantly related species and has remained under balancing selection for tens of millions of years—to date, the only such example in hominoids and Old World monkeys outside of the major histocompatibility complex.

natural selection | balanced polymorphism | population genetics

Balancing selection pressures can maintain two or more alleles in the population for long periods of time—so long, that the polymorphism may be shared due to identity by descent among distinct species, leading to a trans-species polymorphism. In this scenario, the time to the most recent common ancestor (tMRCA) of the selected alleles will predate speciation times. Due to linkage, sites near the selected ones may also have old tMRCAs, resulting in unusually high diversity within species and shared alleles between species (1). Because of recombination events between the two allelic classes in the history of the sample, however, the tMRCA at linked sites can also be much more recent than at the selected sites, so diversity levels need not be unusually high. Moreover, the old tMRCA provides many opportunities for recombination, which will erode the segment that carries the high diversity signal (2, 3); as a result, ancient balancing selection will leave only a narrow footprint in genetic variation data and will often be hard to detect (1, 3).

Perhaps for this reason, only a handful of examples of trans-species polymorphisms are known at the molecular level. The most famous examples are the self-incompatibility loci in plants and the major histocompatibility complex (MHC) in vertebrates. The self-incompatibility loci prevent self-fertilization, and variation is likely maintained by negative frequency-dependent selection (4, 5). In turn, the nature of selective pressures at the MHC are unclear, but are believed to result from its central role in the recognition of pathogens (for a recent review, see ref. 6). Within humans, the MHC is the only region known to harbor variation shared identical by descent with other species. Variants at other loci have also been found to be shared with hominoid species, potentially indicating shared balancing selection pressures, but patterns of genetic variation do not support the hypothesis that the alleles are identical by descent as opposed to due to recurrent mutations [e.g., the phenylthiocarbamide (*PTC*) alleles in humans and chimpanzees] (7).

Arguably the best studied case of apparent convergent evolution is the ABO blood group (A, B, AB, O) (1, 8), the first molecular polymorphism to be characterized in humans. ABO blood groups are defined by the presence or absence of specific antigens that circulate in body fluids and are attached to lipids at the surface of various epithelial and endothelial cell types (notably in the gastrointestinal tract, but also, in hominoids only, on red blood cells) (9, 10). These antigens are associated with complementary immune antibodies produced in the gut after contact with bacteria and viruses carrying A-like and B-like antigens (11). Whereas the biological significance of ABO outside of its role in transfusion is unclear (12), histo-blood antigens in general are known to act as cellular receptors by which pathogens can initiate infections (13–15). Furthermore, variation in ABO has been associated with susceptibility to a number of infectious diseases (reviewed in ref. 16), pointing to a role of ABO in immune response.

The presence or absence of the A, B, and H (O) antigens result from allelic variation at the *ABO* gene, which encodes a glycosyltransferase. The A transferase, able to transfer *N*-acetyl-D-galactosamine to the H acceptor substrate, is encoded by the A allele and the B transferase, able to transfer D-galactose to the same acceptor substrate, by the B allele. A and B alleles at *ABO* are codominant, whereas the null O alleles are recessive (17). Two amino acids at positions 266 and 268 in exon 7 are responsible for the A and B enzymatic specificity in humans (18) and are surrounded by a peak of nucleotide diversity (3, 19, 20). As an illustration, the diversity level in exon 7 is 0.0058 in a Yoruba sample, a value equalled or exceeded in only 0.08% of comparable exonic windows in the genome (*SI Methods*, *SI Note S1*). This peak of diversity provides support for long-lived balancing selection acting on this locus (20, 21).

Author contributions: L.S., E.E.T., G.S., C.O., and M.P. designed research; L.S., E.E.T., and J.L. performed research; S.W.M., J.M., S.R., and K.G. contributed new reagents/analytic tools; L.S., E.E.T., T.F., A.V., G.S., and M.P. analyzed data; and L.S., G.S., C.O., and M.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. M.W.F. is a guest editor invited by the Editorial Board.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. JQ857042–JQ857076).

[1]L.S. and E.E.T. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: lsegurel@uchicago.edu, c-ober@genetics.uchicago.edu, or mfp@uchicago.edu.

[3]Present address: Animal Behavior, Ecology and Conservation, Canisius College, Buffalo, NY 14208.

[4]C.O. and M.P. contributed equally to this work.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1210603109/-/DCSupplemental.

Remarkably, the A, B, and H antigens exist not only in humans but in many other primates (reviewed in ref. 22), and the same two amino acids are responsible for A and B enzymatic specificity in all sequenced species (8, 18, 23–25). Thus, primates not only share their ABO blood group, but also the same genetic basis for the A/B polymorphism. O alleles, in contrast, result from loss-of-function alleles such as frame-shift mutations and appear to be species specific (26). That different species share the same two A/B alleles could be the result of convergent evolution in many lineages or of an ancestral polymorphism stably maintained for millions of years and inherited across (at least a subset of) species. The two possibilities have been debated for decades, with a consensus emerging that A is ancestral and the B allele has evolved independently at least six times in primates (in human, gorilla, orangutan, gibbon/siamang, macaque, and baboon) (8, 25, 26), in particular, that the human A/B polymorphism arose more recently than the split with chimpanzee (8, 20) (*SI Methods, SI Note S2*). We show instead that the remarkable distribution of *ABO* alleles across species reflects the persistence of an old ancestral polymorphism that originated at least 20 million years (My) ago and is shared identical by descent by humans and gibbons as well as among distantly related Old World monkeys.

Results

Previous to ours, 31 studies reported phenotypic and/or genetic data on ABO in non-human primates (Datasets S1 and S2). To supplement these data, we sequenced exon 7 of the *ABO* gene in four hominoid species (10 bonobos, 35 western chimpanzees, 31 lowland gorillas, and 12 orangutans from both Sumatra and Borneo), in two previously uncharacterized Old World monkeys (five colobus and four vervet monkeys), as well as, for the first time, in two New World monkeys (four marmosets and three black howler monkeys) (*Methods*, *SI Acknowledgments* and Datasets S3, S4, and S5). Among 40 non-human species, the ABO polymorphism is ubiquitous, with 19 species in 10 genera polymorphic for A and B, as well as 10 species in seven genera without a B in our sample and 11 species in six genera without an A (Dataset S1, *SI Methods*, *SI Note S3*, and Fig. 1); 15 species in 11 genera also have an unambiguous O allele. Contrasting sequences obtained for marmosets (which are all A in our sample) and black howler monkeys (all B in our sample) reveal that the genetic basis for A/B specificity is the same in New World monkeys as in hominoids (Fig. 1). For the A allele, we further observed that, compared with hominoids, colobus and vervet monkeys use a different codon to encode the same amino acid at one of the two functional positions, as was previously noted for macaques (24) and baboons (27) (Fig. 1). This phylogenetic pattern is consistent with a synonymous substitution in the A lineage leading to Old World monkeys.

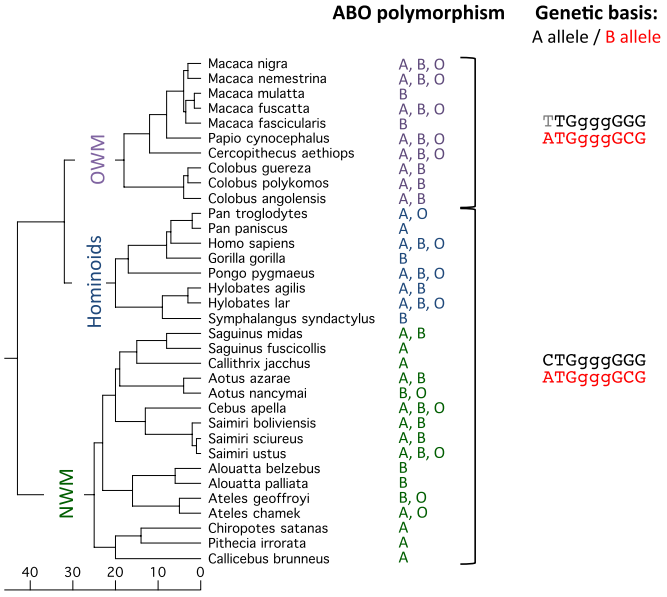

Fig. 1. The phylogenetic distribution of ABO phenotypes and genotypes. Shown is a phylogenetic tree of primate species, with a summary of phenotypic/genotypic information given in the first column, and the genetic basis for the A versus B phenotype provided in the second column (functionally important codons at positions 266 and 268 are in uppercase letters). See Dataset S1 for the source of information about phenotypes/genotypes. Only species with available divergence times are represented here (34 of 40). The phylogenetic tree is drawn to scale, with divergence times (on the *x* axis) in millions of years taken from ref. 29. OWM, Old World monkeys; NWM, New World monkeys. Under a model of convergent evolution, these data suggest that A is the ancestral allele, and a turnover (e.g., a neutral substitution) occurred on the branch leading to Old World monkeys. If instead, B were ancestral, all Old World monkeys would have had to serendipitously converge from ATG to TTG to encode a leucine, whereas all New World monkeys and hominoids would have had to converge to the CTG codon.

To distinguish between the maintenance of an old trans-species polymorphism (Fig. 2*A*) and more recent convergent evolution (Fig. 2*B*), we first sought to gain a sense of the length of the segment that should carry a signal of a trans-species polymorphism. To this end, we approximated the expected length of the (two-sided) segment contiguous to the focal sites, for plausible values of the salient parameters (see *Methods*, *SI Methods, SI Note S4,* and Fig. S1 for details). Consistent with previous modeling work (2), we found that it should be at most a few hundred base pairs (bp) in length, depending notably on the pair of species considered and the recombination rate (Fig. 2*C* and Fig. S2). One implication is that previous studies that considered larger windows or segments far from the selected sites may have missed the footprints of a trans-species polymorphism (*SI Methods, SI Note S2*).

Based on our calculations, we generated trees of *ABO* haplotypes for short regions around the functional sites, i.e., 300 bp in hominoids and 200 bp in Old World monkeys (*SI Methods, SI Note S5*). Strikingly, in the trees within hominoids (other than orangutan) and within Old World monkeys, A and B alleles cluster by type rather than by species (Fig. 3). The clustering by A and B types is consistent with a scenario in which the A and B allelic classes had a most recent common ancestor long ago and persisted across species (Fig. 2*A*). The lack of clustering for the O alleles could then reflect frequent turnovers (i.e., replacement) of these alleles within a balanced polymorphism, as expected from the high mutation rate to null alleles (28). Alternatively, the phylogenetic trees could reflect convergent evolution of multiple functional mutations in numerous lineages (Fig. 2*B*).

Because of recombination, diversity among *ABO* alleles should be highly heterogeneous along the sequence, with at most a few (and possibly no) short segments showing unusually high diversity levels (2). We therefore focused on small, sliding windows along exon 7 (see Figs. S3 and S4 for a slightly larger window choice), comparing the pairwise synonymous diversity among *ABO* allelic classes to the 95% confidence interval for synonymous divergence between the same allele (A or B) from different species (*Methods*). In comparisons of lineages from different species, a model of convergent evolution predicts that, near the selected sites, the divergence between allelic classes (e.g., A and B) will be the same as divergence within allelic classes (e.g., A and A), whereas a model of trans-species polymorphism predicts that the former will exceed the latter. The excess will be subtle, however, because the divergence between lineages from the same allelic class sampled in different species (e.g., A and A) is also inflated close to an ancient balanced polymorphism (due to recombination; see figure 1 in ref. 2). We note further that because O is a null allele, there is a high mutation rate to O, such

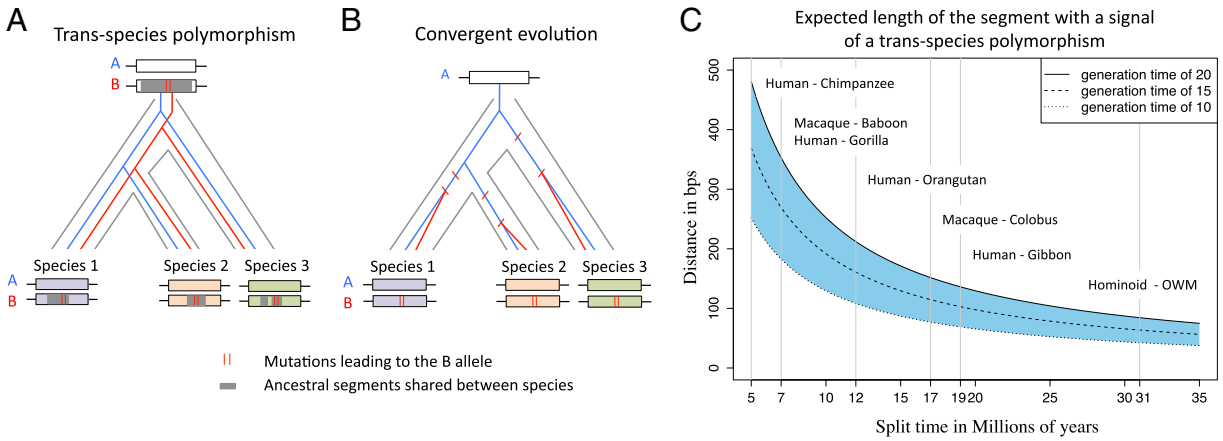

Fig. 2. Expectations under the two possible evolutionary hypotheses. Schematic of expectations for (*A*) the trans-species polymorphism and (*B*) the convergent evolution hypotheses. We illustrate the latter case assuming that A is ancestral. The population tree is outlined in gray, and the A and B lineages sampled from three species are shown in blue and red, respectively. Under the trans-species polymorphism hypothesis, the divergence between A and B is deeper than the species split and species share a short ancestral segment with shared polymorphisms; neither of these patterns is expected under a model of convergent evolution. (*C*) Plot of the expected length of the segment in which a signal of a trans-species polymorphism should be detectable. Specifically, we present the expected (two sided) segment contiguous to the selected site in which the divergence between A and B lineages from different species should exceed the divergence between A (or B) lineages from different species (*Methods* and *SI Methods, SI Note S4*). We considered a recombination rate of 1 cM/Mb, because the recombination rate in this exon is estimated to be ∼1 cM/Mb in humans (54) and higher in chimpanzees (55), and varied the generation time between 10 and 20. Higher recombination rates lead to shorter segment lengths (Fig. S2). OWM, Old World monkeys.

that the current O could be derived from any functional allele, including some that have not persisted to the present. As a consequence, divergence between O and A (or O and B) may be deep, possibly even deeper than A versus B, even if the O allele is itself recent.

Within Old World monkeys, there is tremendous synonymous diversity between A and B alleles in macaques, exceeding the divergence between macaque and baboon, and consistent with the divergence between macaque and colobus monkey ∼18 Mya (29) (Fig. 4*A*). Similarly high synonymous diversity is visible between A and B alleles within colobus monkeys, notably around the functional sites (Fig. 4*A*). A similar but weaker pattern is seen in baboons (Fig. 4*A*), possibly due to greater erosion of the ancestral segment by recombination. In accordance with the inheritance of an ancestral segment identical by descent across macaques and baboons, two synonymous polymorphisms are shared between the two species (Fig. 4*A*). Further evidence that the origin of the polymorphism predates their split comes from the comparison between baboon A and macaque O (or baboon O and macaque B), which reveal deeper coalescent times between allelic classes than seen within an allelic class (i.e., baboon and macaque A or B). Together, these findings strongly support the hypothesis of a trans-species polymorphism shared among macaque, baboon, and colobus monkey, with an origin as old as 18 Mya (or greater).

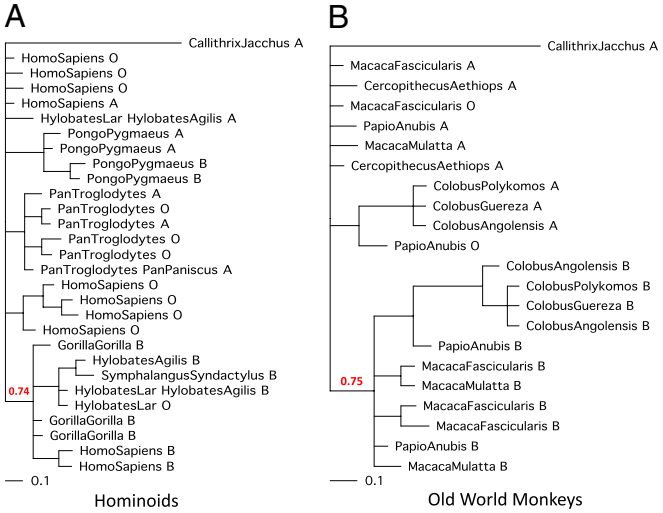

Fig. 3. Tree of *ABO* exon 7 alleles in (*A*) hominoids and (*B*) Old World monkeys. In *A*, the tree is based on 300 bp; in *B*, the tree is based on a smaller window of 200 bp, because the shorter generation time in Old World monkeys should lead to a smaller segment with a signal of a trans-species polymorphism (2) (Fig. 2*C* and *Methods*). The tree is centered on the two functional sites. We excluded rare recombinant haplotypes between functional classes (*SI Methods, SI Note S3*). In red is the median clade credibility (based on three runs that yielded identical consensus trees), i.e., the proportion of trees sampled from the posterior distribution that had this clade (56).

In hominoids, in turn, human *ABO* diversity is over an order of magnitude higher than typical polymorphism levels of ∼0.1% (30) (Fig. 4*B*), with synonymous diversity between *ABO* alleles similar to divergence levels among African apes and compatible with a divergence as high as between human and gibbon lineages, who last had a common ancestor ∼20 Mya (29). This situation is mirrored in gibbons (Fig. 4*B*), as expected if humans and gibbons share A and B alleles identical by descent. Furthermore, divergence between chimpanzee A and human B alleles significantly exceeds the divergence among African ape lineages (Fig. 4*B*), which is predicted only if the A/B polymorphism predates the species' split. In orangutan, in contrast, A and B lineages are highly similar (Fig. S4), suggesting they are recently derived; because this case could represent a turnover event in this lineage, however, it does not preclude the possibility that the A/B balanced polymorphism is older (28). Thus, synonymous pairwise differences in hominoids point to an origin of the A/B polymorphism before the divergence of extant African apes ∼8 Mya, and are consistent with the maintenance of a balanced polymorphism since the root of hominoids ∼20 Mya, which persisted in humans and gibbons.

Because of the stochasticity of recombination and mutation events in a small window, the divergence levels are noisy and do not establish whether the polymorphism is significantly older than 8 My, i.e., whether A and B alleles arose twice in hominoids, once in the ancestor of the African apes and once in the ancestor of gibbons and siamangs, or only once, before the human–gibbon

EVOLUTION

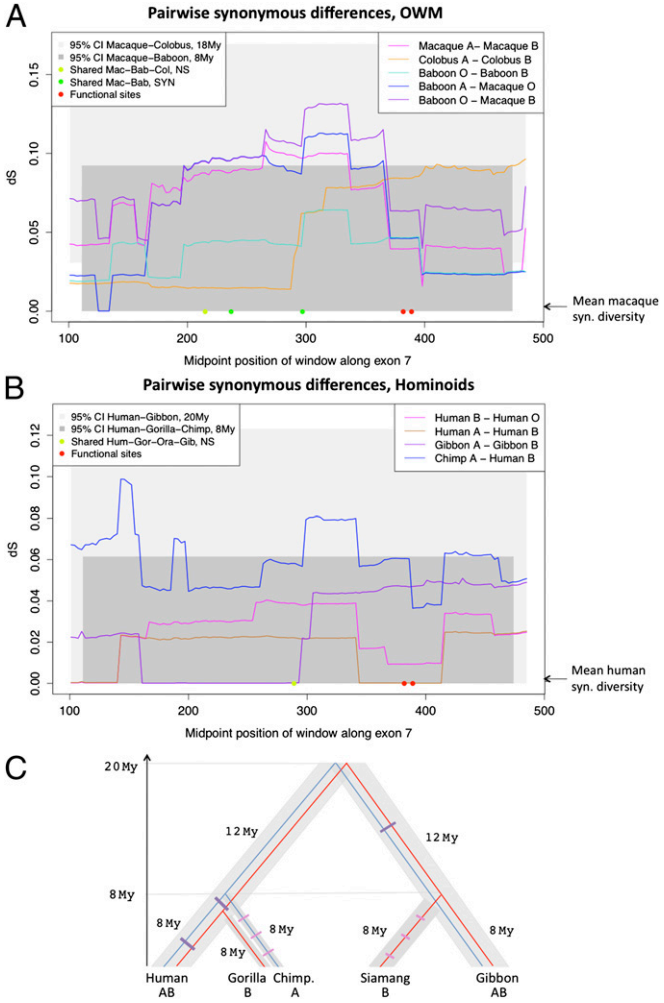

Fig. 4. Evidence for a trans-species polymorphism in Old World monkeys and hominoids. Shown in *A* for Old World monkeys and *B* for hominoids are the synonymous pairwise differences (dS) among *ABO* haplotypic classes in 201-bp (i.e., 67 codons) sliding windows, as well as the shared SNPs between the species compared. For details about how dS and the 95% confidence interval were estimated, see *Methods*. Genome-wide mean synonymous diversity estimates within species were taken from ref. 45. When multiple species were available per genera, we chose one representative with the largest sample size, namely *Macaca mulatta*, *Colobus angolensis*, and *Hylobates lar*. Only the most informative comparisons are presented here, with the rest shown in Figs. S3 and S4 (which also includes similar figures using a larger sliding window choice of 300 bp). Because of recombination, the segment carrying the footprint of a trans-species polymorphism will not necessarily be contiguous and hence, even if there were no stochasticity in the mutation process, diversity levels may be jagged and may only be unusually deep in (at most) small windows (2). In *C*, a depiction is shown of the synonymous substitutions inferred to have occurred in the hominoid phylogeny (represented as ticks, in pink for monomorphic lineages with only one allelic class and in purple for polymorphic lineages with multiple allelic classes). Numbers along the lineages represent millions of years. *A* and *B* lineages are shown in blue and red, respectively. Orangutan is not shown because it appears that a recent turnover occurred in this species, so the lineage is not informative for our test; similarly, the branch from the common ancestor with chimpanzee to bonobo is too short to be informative (*Methods*). Parsimony was used to assign synonymous changes to hominoid lineages. Gorilla is shown closest to human because a substitution (at a non-CpG site) is inferred to have occurred in the ancestor of humans and gorillas but is not found in chimpanzees, suggesting incomplete lineage sorting in this region of the genome (49) (Dataset S3); treating it instead as a multiple hit in humans and gorillas only strengthens our conclusions.

split. We therefore considered an additional prediction that allows us to distinguish between these two cases. When multiple allelic classes are maintained in the population, new mutations can drift up to high frequency within one class, but cannot fix in the species until they recombine onto the other class (31). If A and B date back to the origin of hominoids ~20 Mya, then the maintenance of the two allelic classes will have reduced the number of synonymous fixations at linked sites relative to what would be expected in the absence of a balanced polymorphism; in other words, the deeper coalescent times at linked sites will have left less time for fixations to occur. In contrast, under a model in which B (or A) evolved twice, only one allelic class would have been present in the first 12 My of hominoid evolution (whether A or B), and no such slowdown in synonymous fixations would be expected. We tested the null hypothesis that only one *ABO* allelic class was present during the first 12 My of hominoid evolution by comparing the observed rate of synonymous substitutions in extant "monomorphic" branches that do not have both A and B (i.e., chimpanzee, gorilla, and siamang) to the rate in the two internal branches of the hominoid phylogeny (using 585 bp of exon 7; Fig. 4*C* and *Methods*). We found that the internal branches evolved significantly more slowly than did monomorphic lineages (P value = 0.01). In contrast, we could not reject a null model in which the two internal branches evolved at the same rate as did branches polymorphic for A and B (human and gibbon; P value = 0.18). Thus, the data are inconsistent with a model where only one allelic class existed during the first 12 My of hominoid evolution (i.e., convergent evolution in humans and gibbons) and are best explained by the persistence of the A and B allelic class for the ~20 My since divergence of humans and gibbons.

Discussion

Our results indicate that ABO is a trans-species polymorphism inherited identical by descent in humans and gibbons as well as among all Old World monkeys studied (macaques, baboons, and colobus monkeys), and which was therefore maintained over tens of millions of years. *ABO* is the second example of a locus at which human variation traces back to the origin of hominoids. Moreover, given that the signal of a trans-species polymorphism is expected to decrease over time, eventually becoming undetectable (2, 3) (Fig. 2*C*), we cannot exclude an older origin of the balanced polymorphism, which led to allele sharing among hominoids and Old World monkeys (albeit with a turnover of the A allele), or possibly even with New World monkeys. This finding points to selection pressures that have remained strong relative to genetic drift throughout the evolution of these species (28).

As remains the case for the MHC (6), the selection mechanism maintaining this polymorphism across so many primate species is largely unknown. One possibility, heterozygote advantage (seen e.g., in the *HBB* gene in response to malaria and sickle-cell anemia) (32), may be unlikely to underlie long-lived balancing selection, instead representing a transient solution to balancing selection pressures until a single allele that confers the heterozygote phenotype arises or a duplication occurs (33). In support of this argument, the AB phenotype can be created by single "*cis*-AB" alleles (34) and yet such alleles have not reached fixation in any of the surveyed species and are rare in humans (Dataset S3). Moreover, the presence of the ABO polymorphism throughout primates implies that the selected phenotype is probably not tied to the expression of antigens on red blood cells, a trait restricted to hominoids (35). Humans are known to have many histo-blood subgroups (notably among A types), which are interchangeable for transfusion purposes, but differ in quantity and quality of antigens (36, 37); similarly, chimpanzees, gorillas, orangutans, and gibbons have been reported to have variable A and B subgroups, respectively, that differ in antigenic properties (26, 38). Thus, although A, B, and O are clearly of functional importance and may denote the strongest fitness differences

among variants of the *ABO* gene, these histo-blood labels are unlikely to provide a complete description of the allelic classes acted on by natural selection. These considerations suggest that variation at *ABO* reflects a multiallelic balanced polymorphism, with cryptic differences in function among A and B alleles. As expected from an ancient multiallelic balanced polymorphism (21, 28), there was occasional turnover within allelic classes, including of a codon in the lineage leading to Old World monkeys and of A and B alleles within orangutan, as well as frequent turnovers of O alleles in all species (28). Numerous losses of A and B have also occurred (Fig. 1), possibly as a result of bottlenecks in the history of the species (39) or due to differences in selection pressures among lineages.

In any case, the maintenance of ABO across so many primates reveals previously unknown and important functions of a heavily studied gene. More generally, this study illustrates a general approach that can be used to scan for ancient balancing selection in the genome and raises the possibility that, with the availability of genome-wide polymorphism data from closely related species, this mode of selection will turn out to be more common than currently believed.

Methods

Resequencing Data for *ABO*. We used previously published data: in humans, 60 individuals of European ancestry (CEU), 60 individuals of South East Asian ancestry (CHB + JPT) and 59 individuals of Sub-Saharan ancestry (YRI) (30), 31 olive baboons (*Papio anubis*) (27), 13 macaques (seven cynomolgus or crab-eating macaques, *Macaca fascicularis* and six rhesus macaques, *Macaca mulatta*) (24), 17 gibbons (five agile gibbons, *Hylobates agilis* and 12 white-handed gibbons, *Hylobates lar*), and six siamangs (*Symphalangus syndactylus*) (25). In addition, we sequenced hominoid samples from Lincoln Park Zoo (40) (*SI Acknowledgments*): 10 bonobos (*Pan paniscus*), 35 western chimpanzees (*Pan troglodytes*), 31 lowland western gorillas (*Gorilla gorilla*), and nine orangutans (*Pongo pygmaeus*: three orangutans from Sumatra, two from Borneo, and four hybrids). Three black howler monkeys (*Alouatta caraya*) samples were also obtained from Lincoln Park Zoo. Additionally, samples from three Sumatran orangutans were purchased from the San Diego Zoo, five colobus monkeys (three *Colobus angolensis*, one *Colobus polykomos*, and one *Colobus guereza*) from the Integrated Primate Biomaterials and Information Resource (IPBIR) through the Coriell Institute, four vervet monkeys (*Chlorocebus aethiops*) from Alpha Genesis, and four marmosets (*Callithrix jacchus*) from the Southwest Foundation for Biomedical Research.

When needed, genomic DNA was extracted from blood using the Puregene DNA Isolation kit (Gentra Systems). Amplification of exon 7 was performed using primers and conditions described in Dataset S5. Sequencing reactions were performed using the Big Dye Terminator v3.1 Cycle Sequencing kit (Applied Biosystems). Chromatograms were aligned and analyzed using the Phred-Phrap-Consed package (41).

Haplotypes were estimated in each species separately using PHASE2.1 (42). The program was run twice, with different seeds (option –S); the two outputs were identical other than for six SNPs in colobus monkeys (of which five SNPs were in perfect linkage disequilibrium) and one SNP in humans (in the CHB + JPT population). Trees and diversity plots rely on the first output, but identical conclusions were obtained with the second one. The substitutions and polymorphisms requiring more than one mutation given the accepted species tree (29) are listed in Dataset S3 and the numbers of polymorphic sites found per species are listed in Dataset S4.

Estimating the Length of the Segment Carrying a Signal of a Trans-species Polymorphism. The presence of a trans-species polymorphism at one site distorts patterns of genetic variation data at linked sites, but only over a short distance (2). To estimate this distance, we assumed that balancing selection has maintained two alleles, A and B, at a selected site (without turnover) in two species since before the time of their split, T generations ago. Specifically, we assumed that, at the selected site, the time to the most recent common ancestor for two A alleles or two B alleles sampled from different species is less than the coalescent time for an A and a B allele sampled from different species. We then derived expressions for the expected length of the segment contiguous to the selected site on which we expect to see various signals of a trans-species polymorphism (*SI Methods*, *SI Note S4*).

To estimate the expected segment length for specific species pairs, we used a generation time of 15 and an ancestral effective population size N_a of ~30,000, a typical value for ancestral hominoids (43, 44). This value is also roughly equivalent to the current effective population size of rhesus macaques (~36,000, based on diversity estimates of 0.29% per base pair (45), and a mutation rate of 2×10^{-8} per base pair per generation, which is higher than in hominoids and consistent with the higher synonymous divergence in Old World monkeys). This value of N_a leads to an expected pairwise coalescence time of $2gN_a = 0.9$ My, and, using divergence times from refs. 25 and 29, to the following split time estimates: ~7 My for macaque–baboon, ~17 My for macaque–colobus, ~31 My for macaque–human, and ~19 My for human–gibbon. In turn, the split between human and chimpanzee was taken to be ~5 My, between human and gorilla ~7 My, and between human and orangutan ~12 My (44, 46–49).

Calculating Synonymous Pairwise Differences, d_s. The average synonymous pairwise differences between alleles (d_s) per 201 bp (Fig. 4 *A* and *B*) and 300 (Figs. S3 and S4) sliding window in *ABO* exon 7 were calculated with the maximum likelihood method (50) implemented in the program *codeml* from PAML (51). The codon substitution model was one in which the equilibrium codon frequencies are estimated from the average nucleotide frequencies in the sequence (option CodonFreq = 1). The pairwise comparison was used (runmode = −2), to avoid relying on an underlying species tree for all of the sequences. To obtain a more precise estimate, the transition-to-transversion ratio was estimated for all species all together from 585 bp (the sequence for which we have data for all species). The obtained value ($\kappa = 5.8$) was then fixed when estimating the d_s per window. Using the approximation from ref. 52, implemented in the program yn00 from PAML (51) instead, i.e., allowing the transition–transversion bias to vary along the sequence and estimating each equilibrium codon frequency from the data (option CodonFreq = 3), had a considerable effect on the denominator of d_s, but did not change any qualitative difference (i.e., the ordering of the comparisons). In humans, we used the low coverage pilot data from the 1000 Genomes YRI (30); however, instead considering data generated by the Seattle SNP project (http://pga.gs.washington.edu/) using Sanger sequencing yielded highly similar results.

The 95% confidence intervals for d_s were calculated as follows: we estimated the mean d_s in 585 bp for the A alleles using *codeml* with the species tree specified, then divided it by the length of the tree to obtain the expected number of mutations per base pair per million years. We did the same for the B alleles, and took the mean of the estimates from the two trees, m, which was 0.0013/My/bp in hominoids and 0.0025/My/bp in Old World monkeys. We then assumed that synonymous mutations are neutral, so that the number of synonymous mutations in a lineage is Poisson distributed with mean λ, where $\lambda = mT^*L$, T^* is the divergence time between lineages from different species and L is the number of synonymous sites (estimated by PAML). T^* values were taken from previously published estimates, notably from ref 29: 20 My for human–gibbon, 18 My for colobus–macaque, 17 My for human–orangutan, 12 My for vervet–macaque, and 8 My for baboon–macaque. For African apes, a synonymous allele (G at nucleotide 813) is fixed in human and gorilla but absent from chimpanzee, suggesting that gorilla is closer to human than chimpanzee/bonobo in this region, i.e., indicating a case of incomplete lineage sorting (49). If so, then lineages from humans, chimpanzees, and gorillas coalesced before the split of the three species. To account for this possibility, we used 8 My as the divergence time between human, chimpanzee, and gorilla lineages. This is conservative for our purposes, as a more recent divergence would imply a lower 95% CI and an even greater signal of trans-species polymorphism.

Test of Convergent Evolution Using the Internal Branches in the Hominoid Tree. Using the 585-bp sequence for which we have data for all species, we inferred where synonymous substitutions occurred along the tree of hominoids by parsimony, verifying that PAML output yielded similar ancestral sequences (using the maximum likelihood method based on an a priori species tree, with gorilla closest to human). We then calculated the rate of synonymous substitutions on extant lineages that are monomorphic (chimpanzee, gorilla, and siamang) or polymorphic (human and gibbon). For this purpose, we ignored the bonobo lineage, because the addition of the bonobo lineage (which split from chimpanzee <1 Mya (53) does not provide much time for neutral mutations to arise and fix, so is relatively uninformative. In addition, the orangutan lineage was excluded from this analysis because it appears to have experienced a recent turnover.

We assumed that B arose ~8 Mya (the minimum age indicated by Fig. 2*B*) and tested a null model in which there was only one *ABO* class during the first 12 My of hominoid evolution. Specifically, we compared rates of substitutions in exon 7 on internal branches (one substitution in 24 My) to that seen in monomorphic lineages (six substitutions in 24 My). Assuming that

the number of synonymous fixations is Poisson distributed, this yields P value = 0.01. In contrast, when we asked whether the internal branches differ from extant branches that are polymorphic (two mutations in 16 My), we could not reject the null model: P value = 0.18. Comparing the two tests, the posterior odds for a model of trans-species polymorphism rather than convergent evolution are 10:1. Considering instead that this region is not a case of incomplete lineage sorting among African apes (i.e., that the site at nucleotide 813 experienced a recurrent mutation in human and gorilla) does not change the conclusions (in this case, the posterior odds are 34:1). A similar test was not performed in Old World monkeys because we do not have sequence data for monomorphic species from which to estimate the rate of synonymous divergence.

We note that, for monomorphic species, we did not substract the tMRCA of the sample when considering over what time period substitutions accumulated; this is conservative for our purposes, as it leads to a slower rate of substitutions on these branches. This test further assumes (again conservatively) that lineages polymorphic at present were always polymorphic and lineages monomorphic at present lost the polymorphism immediately after speciation. Beyond requiring that the balanced polymorphism be stably maintained in polymorphic lineages, it does not make assumptions about the strength or mechanism of selection.

ACKNOWLEDGMENTS. We thank P. Andolfatto, G. Coop, R. Hudson, G. Perry, J. Pritchard, and M. Stephens for helpful discussions, and G. Coop and J. Pritchard for comments on an earlier version of the manuscript. This study used biological materials obtained from the Southwest National Primate Research Center, which is supported by National Institutes of Health-National Center for Research Resources Grant P51 RR013986. This work was supported by a Rosalind Franklin award and R01 GM72861 (to M.P.). C.O. was partially funded by Grant R01 HD21244. E.E.T. was supported by Grant K12 HL090003. G.S. was funded by a Flegg fellowship and Israel Science Foundation Grant 1492/10. M.P. is a Howard Hughes early career scientist.

1. Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet* 2(4):e64.
2. Wiuf C, Zhao K, Innan H, Nordborg M (2004) The probability and chromosomal extent of trans-specific polymorphism. *Genetics* 168(4):2363–2372.
3. Bubb KL, et al. (2006) Scan of human genome reveals no new Loci under ancient balancing selection. *Genetics* 173(4):2165–2177.
4. Wright S (1939) The distribution of self-sterility alleles in populations. *Genetics* 24(4): 538–552.
5. Castric V, Vekemans X (2004) Plant self-incompatibility in natural populations: A critical assessment of recent theoretical and empirical advances. *Mol Ecol* 13(10): 2873–2889.
6. Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proc Biol Sci* 277(1684):979–988.
7. Wooding S, et al. (2006) Independent evolution of bitter-taste sensitivity in humans and chimpanzees. *Nature* 440(7086):930–934.
8. Saitou N, Yamamoto F (1997) Evolution of primate ABO blood group genes and their homologous genes. *Mol Biol Evol* 14(4):399–411.
9. Szulman AE (1960) The histological distribution of blood group substances A and B in man. *J Exp Med* 111:785–800.
10. Oriol R, Le Pendu J, Mollicone R (1986) Genetics of ABO, H, Lewis, X and related antigens. *Vox Sang* 51(3):161–171.
11. Springer GF, Horton RE, Forbes M (1959) Origin of anti-human blood group B agglutinins in white Leghorn chicks. *J Exp Med* 110(2):221–244.
12. Reid ME, Mohandas N (2004) Red blood cell blood group antigens: Structure and function. *Semin Hematol* 41(2):93–117.
13. Borén T, Falk P, Roth KA, Larson G, Normark S (1993) Attachment of Helicobacter pylori to human gastric epithelium mediated by blood group antigens. *Science* 262 (5141):1892–1895.
14. Moulds JM, Nowicki S, Moulds JJ, Nowicki BJ (1996) Human blood groups: Incidental receptors for viruses and bacteria. *Transfusion* 36(4):362–374.
15. Frattali Eder A, Spitalnik SL (1997) in *Molecular Biology and Evolution of Blood Group and MHC Antigens in Primates*, eds Blancher A, Klein J, Socha WW (Springer, New York), pp 268–304.
16. Garratty G (2005) Relationship of blood groups to disease: Do blood group antigens have a biological role? *Rev Med Inst Mex Seguro Soc* 43(Suppl 1):113–121.
17. Yamamoto F, Clausen H, White T, Marken J, Hakomori S (1990) Molecular genetic basis of the histo-blood group ABO system. *Nature* 345(6272):229–233.
18. 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.
19. Stajich JE, Hahn MW (2005) Disentangling the effects of demography and selection in human history. *Mol Biol Evol* 22(1):63–73.
20. Calafell F, et al. (2008) Evolutionary dynamics of the human ABO gene. *Hum Genet* 124(2):123–135.
21. Charlesworth B, Charlesworth D (2010) *Elements of Evolutionary Genetics* (Roberts and Co., Greenwood Village, CO), p xxvii.
22. Blancher A, Socha WW (1997) in *Molecular Evolution of Blood Group and MHC Antigens in Primates*, eds Blancher A, Klein J, Socha WW (Springer, New York), pp 30–92.
23. Kominato Y, et al. (1992) Animal histo-blood group ABO genes. *Biochem Biophys Res Commun* 189(1):154–164.
24. 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.
25. 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.
26. 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.
27. 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.
28. Takahata N (1990) A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. *Proc Natl Acad Sci USA* 87(7): 2419–2423.
29. Perelman P, et al. (2011) A molecular phylogeny of living primates. *PLoS Genet* 7(3): e1001342.
30. 1000 Genomes Project Consortium (2010) (A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061–1073.
31. Hudson RR, Kaplan NL (1988) The coalescent process in models with selection and recombination. *Genetics* 120(3):831–840.
32. Allison AC (1954) Protection afforded by sickle-cell trait against subtertian malareal infection. *BMJ* 1(4857):290–294.
33. Spofford BS (1969) Heterosis and the evolution of duplications. *Am Nat* 103(932): 407–432.
34. Yamamoto F, et al. (1993) Molecular genetic analysis of the ABO blood group system: 2. cis-AB alleles. *Vox Sang* 64(2):120–123.
35. Blancher A, Socha WW (1997) The ABO, Hh and Lewis blood group in humans and nonhuman primates. *Molecular Biology and Evolution of Blood Group and MHC Antigens in Primates*, eds Blancher A, Klein J, Socha WW (Springer, New York).
36. E von Dungern LH (1911) Concerning the group-specific structures of the blood (III). *Immunitatsforschung* 8:526–530.
37. Clausen H, Hakomori S (1989) ABH and related histo-blood group antigens; immunochemical differences in carrier isotypes and their distribution. *Vox Sang* 56(1):1–20.
38. Socha WW, Moor-Jankowski J (1979) Blood groups of anthropoid apes and their relationship to human blood groups. *J Hum Evol* 8(4):453–465.
39. Takahata N (1991) Trans-species polymorphism of HLA molecules, founder principle, and human evolution. *Molecular Evolution of the Major Histocompatibility Complex*, eds Klein J, Klein D (Springer, Heidelberg), pp 29–49.
40. 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* 30:427–444.
41. 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.
42. Stephens M, Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet* 76(3):449–462.
43. Becquet C, Przeworski M (2007) A new approach to estimate parameters of speciation models with application to apes. *Genome Res* 17(10):1505–1519.
44. Wall JD (2003) Estimating ancestral population sizes and divergence times. *Genetics* 163(1):395–404.
45. Perry GH, et al. (2012) Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Res* 22(4):602–610.
46. Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D (2006) Genetic evidence for complex speciation of humans and chimpanzees. *Nature* 441(7097):1103–1108.
47. Hobolth A, Christensen OF, Mailund T, Schierup MH (2007) Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet* 3(2):e7.
48. Hobolth A, Dutheil JY, Hawks J, Schierup MH, Mailund T (2011) Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Res* 21(3):349–356.
49. Scally A, et al. (2012) Insights into hominid evolution from the gorilla genome sequence. *Nature* 483(7388):169–175.
50. Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* 11(5):725–736.
51. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24 (8):1586–1591.
52. Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* 17(1):32–43.
53. Becquet C, Patterson N, Stone AC, Przeworski M, Reich D (2007) Genetic structure of chimpanzee populations. *PLoS Genet* 3(4):e66.
54. McVean GA, et al. (2004) The fine-scale structure of recombination rate variation in the human genome. *Science* 304(5670):581–584.
55. Auton A, et al. (2012) A fine-scale chimpanzee genetic map from population sequencing. *Science* 336(6078):193–198.
56. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12):1572–1574.