Hum Genet (2008) 124:123–135
DOI 10.1007/s00439-008-0530-8

ORIGINAL INVESTIGATION

Evolutionary dynamics of the human *ABO* gene

Francesc Calafell · Francis Roubinet · Anna Ramírez-Soriano · Naruya Saitou · Jaume Bertranpetit · Antoine Blancher

Received: 27 May 2008 / Accepted: 3 July 2008 / Published online: 16 July 2008

Abstract The ABO polymorphism has long been suspected to be under balancing selection. To explore this possibility, we analyzed two datasets: (1) a set of 94 23-Kb sequences in European- and African-Americans produced by the Seattle SNPs project, and (2) a set of 814 2-Kb sequences in *O* alleles from seven worldwide populations. A phylogenetic analysis of the Seattle sequences showed a complex pattern in which the action of recombination and gene conversion are evident, and in which four main lineages could be individuated. The sequence patterns could be linked to the expected blood group phenotype; in particular, the main mutation giving rise to the null *O* allele is likely to have appeared at least three times in human evolution, giving rise to allele lineages *O02*, *O01*, and *O09*. However, the genealogy changes along the gene and variations of both numbers of branches and of their time depth were observed, which could result from a combined action of recombination and selection. Several neutrality tests clearly demonstrated deviations compatible with balancing selection, peaking at several locations along the gene. The time depth of the genealogy was also incompatible with neutral evolution, particularly in the region from exons 6 to 7, which codes for most of the catalytic domain.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-008-0530-8) contains supplementary material, which is available to authorized users.

F. Calafell (✉) · A. Ramírez-Soriano · J. Bertranpetit
Unitat de Biologia Evolutiva,
Departament de Ciències Experimentals i de la Salut,
Universitat Pompeu Fabra, Doctor Aiguader,
80, 08003 Barcelona, Catalonia, Spain
e-mail: francesc.calafell@upf.edu

F. Calafell · J. Bertranpetit
CIBER Epidemiología y Salud Pública (CIBERESP),
Barcelona, Spain

F. Roubinet
Etablissement Français du sang Centre Atlantique,
BP 52009, Tours, 37020 Cedex 1, France

F. Roubinet · A. Blancher
Laboratoire d'Immunogénétique Moléculaire,
Faculté de Médecine Purpan,
Université Paul Sabatier Toulouse III,
Bâtiment A2, 31062 Toulouse Cedex 4, France

N. Saitou
Division of Population Genetics,
National Institute of Genetics, Mishima 411-8540, Japan

Introduction

The *ABO* system was discovered by Karl Landsteiner (1901) and consists of three main alleles: two codominant *A* and *B* and one silent and recessive allele called *O*. The *A* and *B* alleles code for glycosyltransferases that add a *N*-acetyl galactosamine or a galactose, respectively, to various substrates generically referred to as H substance. These products result in A or B blood group specific antigens. The combinations of the three main alleles result in four major phenotypes, namely A, B, AB, and O, which are characterized by the presence (or absence) of A and B antigens on the surface of red cells and the presence in the serum of natural antibodies against the antigen absent at the surface of red blood cells. Indeed, because of the natural tolerance, the natural antibodies against the antigens possessed by the individual are not normally observed in physiological conditions in humans. Beside its utmost importance in medicine, the *ABO* system was also the first human genetic system to be applied to human population studies, which revealed the variation in *A*, *B* and *O* allele frequencies among populations (Mourant 1954).

The discovery of the *ABO* gene at the molecular level allowed a refinement of polymorphism knowledge (Yamamoto et al. 1990a, b, 1995). Particularly, it was evidenced that each of the three main antigenic classes (A, B, and O) comprises numerous alleles that can be defined by their coding and non-coding sequences. Currently, over 70 alleles have been defined at the molecular level (see Olsson and Chester 2001; Yamamoto 2004; Yip 2002), and http://www.bioc.aecom.yu.edu/bgmut/abo.htm for reviews), and *ABO* seems to be one of the most polymorphic genes in humans. The main A and B alleles (namely, *A101* and *B101*) differ at four amino acid residues: *A101* carries 176 Arg, 235 Gly, 266 Leu, 268 Gly, while *B101* carries 176 Gly, 235 Ser, 266 Met, 268 Ala. In vitro expression studies, cisAB alleles (i.e., alleles coding for an enzyme that can transfer both *N*-acetyl galactosamine and galactose), and B sequences in other primates have shown that the determining amino acid residues are 266 and 268. The two functional A and B allele classes were revealed to contain numerous sequence variants. For example, the A201 allele responsible for a serologically detectable A^2 phenotype with a 20- to 50-fold reduction in A activity, displays when compared to the A101 allele an insertion at genomic position 1,061, which results in a frameshift adding 21 additional amino acid residues to the protein.

The silent allele O is also greatly heterogeneous when studied at the gene sequence level. The most frequent human *O* alleles are *O01* and *O02*, which have been found at high frequencies in all populations studied so far. They differ in exons 6 and 7 by nine nucleotide substitutions (Olsson and Chester 1996b; Yamamoto et al. 1990a), and by an additional 14 positions in intron 6 (Roubinet et al. 2004, 2001) but share a point deletion of a G at position 261 in exon 6. This deletion, referred to as Δ261 as per the numbering in Yamamoto (2000), induces a frameshift and creates a premature stop codon (nucleotides 352–354), resulting in a truncated (117 amino acids) protein deprived of any glycosyltransferase activity (Yamamoto et al. 1990a). Numerous variants of *O01* and *O02* alleles have been described. They differ from *O01* or *O02* by a few point mutations (Chester and Olsson 2001; Ogasawara et al. 1996a, b, 2001; Olsson and Chester 2001; Olsson et al. 1997, 1998; Roubinet et al. 2004, 2001; Yamamoto 2000; Yip 2000), or result from inter-allelic exchanges between them or with *A* or *B* alleles [for a general review see (Yip 2002)]. Rarer alleles such as *O03* (Grunnet et al. 1994; Yamamoto et al. 1993), *O08* (Olsson and Chester 1996a); *O4* and *O5* (Olsson and Chester 2001); *O301* and *O302* (Ogasawara et al. 2001) carry different inactivating mutations. Previous studies of the exon 6 to exon 7 region had showed three main lineages: *A101*/*O01*, *B101*, and *O02* (Roubinet et al. 2004). Even in such a relative short region (1.8 Kb), the divergence between the main lineages allowed a precise delineation of the putatively recombinant sequences. Additional sequence of exons 1–5 does not contribute much to refining that knowledge, given that these exons add up to only 239 bp. Then, the review by Yip (2002), based mostly on coding sequence, showed mostly this three-lineage genealogy. Sequencing the whole gene (bar the huge intron 1) in selected alleles rather than in random population samples, Seltsam et al. (2003) defined a five-lineage genealogy, in which *A101* was separated from *O01*, and *O03* was revealed as a quite distinct lineage; they also demonstrated that the upstream sequence of *B101* was very similar to that of *A101*, and that *B101* was, in fact, a recombining allele, retaining only its singularity downstream of exon 5.

The adaptive value of the *ABO* polymorphism has long been studied. However, until the discovery of the *ABO* gene (Yamamoto et al. 1990a), the only feasible approach was to explore the association between *ABO* phenotypes and various diseases as reviewed in the classic book by Mourant et al. (1978). Special attention was devoted to infectious disease because of the widespread expression of A and B antigens by various infectious agents and also the use of human A and B substances as receptors by a number of infectious agents. As reviewed by Gagneux and Varki (1999), the *ABO* polymorphism can prevent that the species carrying it be endangered by a pathogen using a given carbohydrate as receptor. On the other hand, the *ABO* polymorphism leads to a polymorphic production of anti-A and anti-B natural antibodies, which potentially protect individuals from the various and numerous infectious agents expressing A and B motifs. From these observations, one can conclude that the silent allele *O*, although being a null allele could have had a selective value because, in homozygosity, it implies that natural anti-A and anti-B antibodies are produced. Moreover, the *O* homozygotes are potentially protected from infectious agents that use the A and B substances as receptors, but they are more sensitive to *Helicobacter pylori* (Borén et al. 1993) and are particularly exposed to severe forms of cholera (Swerdlow et al. 1994). The Δ261 deletion has been shown to be protective against severe malaria (Fry et al. 2007), probably because the O phenotype reduces red-cell rosetting, a virulence factor (Rowe et al. 2007); for a general review of ABO and malaria, see Cserti and Dzik (2007). Other examples of potentially selective agents are *Campylobacter jejuni* (Ruiz-Palacios et al. 2003) and the Norwalk virus (Lindesmith et al. 2003; Marionneau et al. 2002). In fact, the potential value of the *ABO* polymorphism should be understood in the context on its complex interaction with other polymorphic genes encoding fucosyltranferases *FUT1*, *FUT2*, and *FUT3*; that is, the genes previously known as H, Se, and Le (Marionneau et al. 2001). Given this complex pattern of putative selective agents, the direct demonstration of the action of selection is bound to be elusive. However, even if the actual mechanisms of selection cannot be pinpointed, the footprint of selection on the molecular diversity of *ABO* can be sought for, given the variety of tests available to assay for departures from neutral evolution. Saitou and Yamamoto (1997), using cDNA sequences, had already noted the extreme deep coalescence times among human *ABO* alleles, suggesting balancing selection. Stajich and Hahn (2005) found that, in a survey of public resequencing data for 151 genes in two North American populations, *ABO* sequence structure showed clear signs of balancing selection, a finding disputed by Bubb et al. (2006).

We have explored the phylogenetic structure of the main ABO lineages and their nucleotide diversity patterns for departures of neutrality. For that, we have used two data sets: the resequencing of the complete *ABO* gene by the Seattle SNPs project (Akey et al. 2004) in a random sample of 24 African-Americans and 23 European-Americans, and the cloning and resequencing of 1,875 bp of 814 O chromosomes from autochthonous individuals from Africa, Europe, Asia, and the Americas (Roubinet et al. 2004). Note that the total size of sequence produced in each set is similar (~2.2 and ~1.5 Mb, respectively), and that they complement each other in that Seattle SNPs has a limited geographical sampling and the data set by Roubinet et al. (2004) covers a subset of the phylogeny. The joint analysis of both data sets has allowed us to discover new lineages in ABO, to refine the hypotheses concerning the origin of the main alleles and the role of selection and recombination in those origins, and to provide additional evidence for the extensive action of balancing selection on *ABO*.

Methods

Datasets

Two datasets were analyzed: the *ABO* sequences produced by the Seattle SNPs project and the O allele sequences in Roubinet et al. (2004). The Seattle SNP Project aims to resequence a large number of genes involved in the inflammatory response in samples of European- and African-Americans (Akey et al. 2004); as of June 2008, sequence data for 320 genes were publicly available at http://pga.gs.washington.edu. The complete *ABO* gene (including 2.4 Kb upstream and 1.7 Kb downstream) had been resequenced in 24 African-Americans and 23 European-Americans; all individuals were unrelated to each other. The haplotypes posted at the Seattle SNP website had been statistically estimated from segregating sites with frequencies >0.05. Since all variation is needed for many neutrality tests (Kreitman and Di Rienzo 2004; Soldevila et al. 2005), we reconstructed haplotypes using all 214 segregating sites by means of the Bayesian algorithm implemented in the PHASE 2.1 software (Stephens and Scheet 2005; Stephens et al. 2001) available at http://www.stat.washington.edu/stephens/software.html. Each haplotype was attributed to a main allele (A, B, O) and lineage (such as *A101*, *A201*, *O01*, *O02*, and others) using the relevant positions in exon 6, intron 6, and exon 7 as defined by Roubinet et al. (2004) and Seltsam et al. (2003), and reviewed by Yip (2002). A pattern in which a haplotype departed from the consensus sequence in its lineage for at least three consecutive informative positions was taken as indicative of a recombination or gene conversion event.

Roubinet et al. (2004) determined the sequence of 1,875 bp of the *ABO* gene, comprising exon 6, intron 6, and exon 7, in 814 *O* chromosomes from seven human populations (Akan from the Ivory Coast, Berbers from Morocco, Basques from France and Spain, Han Chinese from Putien and Fujiou, and Cayapas and Aymaras from Bolivia). Haplotype phase was resolved molecularly for each individual by sequencing cloned amplification products.

Neutrality tests

Basic descriptive statistics was performed with DNAsp 4.10 (Rozas et al. 2003), available at http://www.ub.es/dnasp. Deviation from neutrality was tested by means of Tajima's *D* (Tajima 1989) and Fu and Li's *F* test (Fu and Li 1993) with a chimpanzee A sequence as an outgroup (The Chimpanzee Sequencing and Analysis Consortium 2005). Fu and Li's *F* test was chosen over the *D* test by the same authors because it showed slightly more statistical power under recombination (Ramírez-Soriano et al. 2008). The significance of both tests was estimated with a coalescent simulation with recombination, conditional on the number of segregating sites. The recombination rate at the *ABO* gene was estimated with data from Kong et al. (2002). According to their map, D9S754 is located ~200 Kb downstream from *ABO* and the recombination at that locus was estimated at 1.07 cM/Mb; thus, total recombination rates can be estimated as 0.025 cM for the Seattle SNP data set (23,759 bp) and as 2.01×10^{-3} cM for the Roubinet dataset (1,875 bp), and the $R = 4N_er$ parameter at 10.26 and 0.80, respectively, if the effective population size of humans is taken as 10,000 (Takahata et al. 1995). Ten thousand coalescent simulations were run for each test with DnaSP 4.10, and Tajima's *D* and Fu and Li's *F* values were obtained in each simulation under the neutral model. The one-tailed significance of the actual tests was given as the empirical fraction of simulations yielding more extreme test values than those observed. Tajima's *D* and Fu and Li's *F* were also computed in sliding windows along the sequence. The statistical significance of peaks in the sliding windows graph was tested also by a coalescent approach, using the scan-ms program (Ardell 2004) with 10,000 iterations.

Phylogenetic trees and dating

Maximum likelihood (ML) trees were produced with the PHYML v.2.4.4 software (Guindon and Gascuel 2003) using a general time reversible (GTR) model with a discrete-gamma distribution of substitution rates. Support for branches was estimated from 1,000 bootstrap iterations. Trees were also generated with a Bayesian method as implemented in MrBayes v. 3.1 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003), also with a GTR model with a discrete-gamma distribution of substitution rates. Other parameters were left in their default states as suggested in the software manual, except for generations, which were increased to one million.

Before a dating method was applied, we tested for the constancy of the molecular clock. We used the DAMBE 4.5 software (Xia and Xie 2001) to perform relative rate tests between pairs of human sequences, using the chimpanzee as an outgroup. We tested the molecular clock in all possible pairs of sequence lineages; since we had a choice of sequences in each lineage, we compared the sequence closest to the root in one lineage against the sequence farthest from the root in the other lineage. None of the tests were significant at the $\alpha = 0.05$ level.

The average number of mutations accumulated from an ancestral sequence (ρ) is related to time with $\rho = \mu t$, where μ is mutation rate (Bertranpetit and Calafell 1996; Morral et al. 1994; Saillard et al. 2000). We have used this method with two different aims: (1) to obtain an age estimate for the overall tree of ABO sequences; (2) to estimate the time needed to accumulate all mutations observed in a lineage from the root of the overall tree, since the distance from the tree root to its tips varies widely from lineage to lineage. A broad mutation rate was estimated by considering the genomewide divergence value between humans and chimpanzees (Ki = 0.0127) (The Chimpanzee Sequencing and Analysis Consortium 2005). Considering that this divergence has accumulated over ~6 million years, a substitution rate of 1.058×10^{-9} per nucleotide per year was estimated.

Results

Evolutionary analysis of the ABO sequences in Seattle SNPs

Sequences for 24 African-Americans and 23 Americans of European descent were determined for 23,759 bp covering the *ABO* gene as well as 2.4 Kb upstream and 1.7 Kb downstream. Three fragments were not sequenced: one of ~1 Kb in intron 1, and two of ~200 bp each, both in the 3′ end of the gene. Sixty-one different haplotypes were found, and summary statistics of variation can be found in Table 1. As expected given the genomewide trend, African-Americans show greater variability than European-Americans, both at the haplotype and the nucleotide levels. For the latter, African-Americans are more diverse than European-Americans in 258 out of the 319 genes (81%) sequenced by the Seattle SNPs project (May 2008). *ABO* is the most polymorphic gene sequenced so far by the Seattle SNPs project, almost five standard deviations above the mean nucleotide diversity (8.6×10^{-4}). Saitou and Yamamoto (1997), using cDNA sequences, had already noted the extreme differentiation among human *ABO* alleles.

The extent of polymorphism along the *ABO* gene is shown in Fig. 1. Three highly polymorphic regions appear to exist in intron 1, intron 3, and from intron 4 to exon 7. This pattern is found in both African-Americans and European-Americans (Spearman's $\rho = 0.941$), though the former show a much higher peak at the intron 4 to exon 7 region, probably due to the absence of *B* alleles in the European American sample.

Figure 2 shows the haplotypes that can be reconstructed from the Seattle SNPs ABO sequences. The two closest haplotypes to the chimpanzee sequence carried an intron 6 sequence that was very close to that of allele *O47* ("Blood group antigen gene mutation database"; http://www.bioc.aecom.yu.edu/bgmut/index.php), also called *Ovartlse20* in Roubinet et al. (2004) (see the Appendix for a complete list of correspondences between the two nomenclature systems). This ancestral-like sequence extended only up to the middle of intron 4, where it was replaced by a sequence very similar to that of the consensus *O02* (see below): in ~15 Kb, the two *O47* haplotypes showed only four fixed substitutions from *O02*; however, in the 3′-most 3.3 Kb, these two haplotypes accumulate six differences from *O02*. In summary, these two haplotypes could be ancient lineages in which a central section has been replaced by recombination from *O02* (or, in case of haplotype H17, from *A201*). In Fig. 2, the *O47*- specific sequence is marked in magenta, as opposed to the brown *O02* (see the online version for colour).

Table 1 Summary statistics of Seattle SNPs ABO sequences

Population	*N*	*S*	*k*	*H*	π
African-Americans	48	207	39	0.9867	29
European-Americans	46	161	28	0.9623	26
Total	94	214	61	0.9838	28

N number of chromosomes, *S* number of segregating sites, *k* number of different haplotypes, *H* haplotype diversity, π nucleotide diversity ($\times 10^{-4}$)

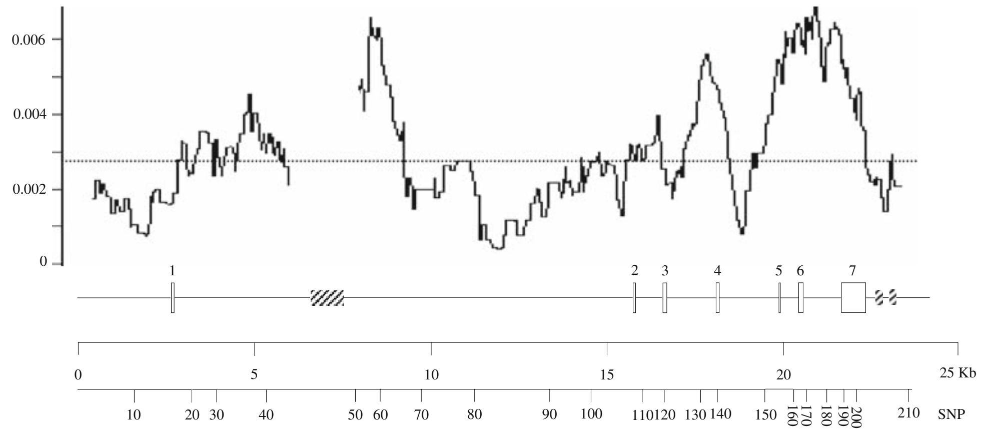
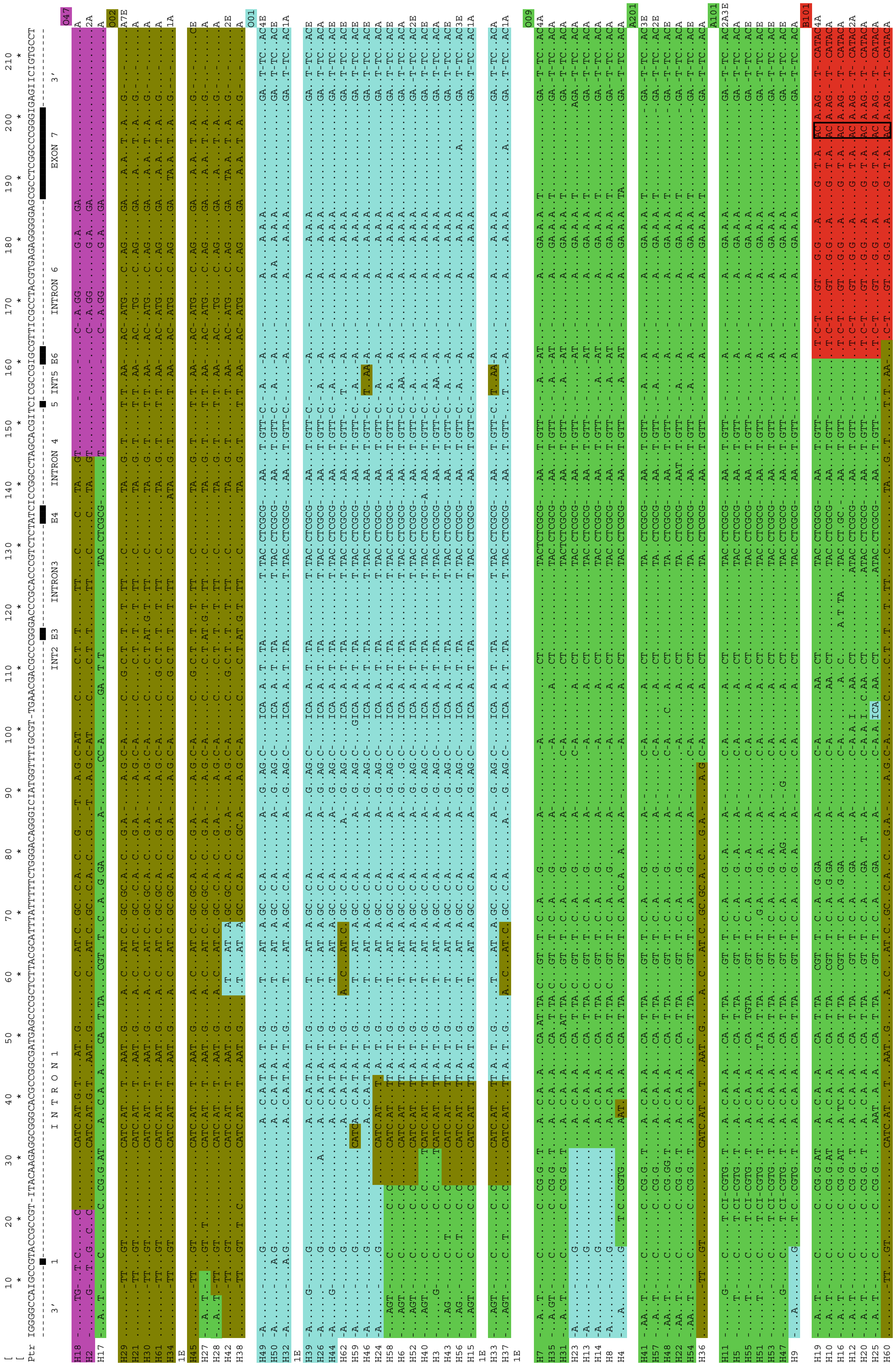


Fig. 1 Nucleotide diversity along the ABO gene. Diversity is plotted for windows of 1,000 bp moved 25 bp. On the *X*-axis, from top to bottom: i) a cartoon of the ABO gene, with exons numbered and regions not sequenced by Seattle SNPs as striped boxes; ii) a physical scale; iii) position of the 214 segregating sites (multiples of 10 only)

O02 lineages (in brown) are easily recognizable from their exon 6 to exon 7 sequences (Roubinet et al. 2004; Seltsam et al. 2003). Although they share with *O01*, the Δ261 deletion that inactivates the *ABO* gene, both lineages are very divergent, with an average 100.4 substitutions (0.00423 per nucleotide) between *O02* and *O01* sequences. This makes it unlikely that *O01* and *O02* share Δ261 because of a recent common origin of both lineages; two alternate possibilities are (1) that *O01* and *O02* have an ancient common origin, diverged a long time ago and retained Δ261 probably because of selection, or (2) that *O01* and *O02* are two independent lineages that converged to Δ261 by parallel mutation or gene conversion. Under the first hypothesis, it is expected that divergence were minimal around Δ261 and would increase outwards. However, this is not the case, since divergence between *O01* and *O02* around Δ261 is over twice the mean (Fig. 3), and the expected pattern of divergence increasing away from Δ261 was not found.

O01 lineages are indicated in blue in Fig. 2. Their *O01* sequence was described as quite similar to that of *A101*, but, overall, *O01* is rather divergent from *A101* (Dxy = 0.0244). Three groups of haplotypes with different phenotypes (namely, *A101*, *A201*, and *O09*) carried similar sequences and formed a definite cluster (in green in Figure 2). Finally, haplotypes corresponding to B alleles (characterized by the L266 M and G268A amino acid changes) carried a distinctive sequence only downstream from exon 6, whereas upstream they carried a sequence related to *A201*. From the start of the Seattle SNP sequence to position 20,150 (in the middle of exon 6), *B101* and *A201* haplotypes had Dxy = 0.00055 and four fixed differences, whereas from position 20,151 to the end of the sequence (at position 23,759), divergence was an order of magnitude larger (Dxy = 0.00737) and the number of fixed differences grew to 26. A *B101* haplotype (H60) carried an *O02* upstream sequence instead. These results are in accordance with the findings of Seltsam et al. (2003) for introns 2–4.

Two features of the *ABO* sequence genealogy emerge from Fig. 2: the presence of haplotypes of mixed origin (given the high polymorphism of *ABO*, such events can be detected with more precision), and the presence of a varying number of lineages along the sequence. Actually, the changes in tree topology along a sequence have been proposed as the means to detect recombination (Kosakovsky Pond et al. 2006). We constructed an ML tree with all presumably non-recombined sequences and rooted it with the chimpanzee sequence (Fig. 4); a Bayesianly-inferred tree showed the same topology. Three main lineages can be individuated: *O02*, *O01*, and *A101-A201-O09*. Note, though, that the divergent *B101* and *O47* lineages were not included in the tree since they are apparently recombinant sequences. Moreover, the number of main lineages varies along the sequence, as revealed by a summary visual inspection of Fig. 2. We chose to analyze this phenomenon by splitting the original sequence length in three segments defined as to minimize recombinant sequences within each segment, and building trees in each segment. We placed the segment limits at positions 6,000 (that is, between SNPs 45 and 46), and 20,124 that is, between SNPs 159 and 160. Taking the first 6 Kb (from the beginning of the sequence to 3.5 Kb into intron 1), three lineages can be discerned: *O02*, *O01*, and the rest of the haplotypes, now with only *A101* haplotypes clustering (Fig. 5a). In a central section, from positions 6,000 to 20,124 (that is, from intron 1 to the end of intron 5), *O02* and *O01* are distinct, and subdivisions in the *B101-A101-A201-O09* lineage are very shallow (Fig. 5b). Finally, from 20,124 to the 3′ end of the sequence (comprising the functionally relevant sites in exons 6 and 7), the tree obtained (Fig. 5c) is deeper (nucleotide diversity is $\pi = 0.00421$ vs. $\pi = 0.00226$ and $\pi = 0.00251$ in the 5′ and central sections) and lineages are more markedly differentiated: *O47* is clearly different from *O02*, and *B101* becomes detached from a cluster now containing *O01* as well as *A101*, *A201*, and *O09*. Thus, the gene region containing the variation that is involved in the functional difference


Ptr
3'
INTRON 1
INT2 E3
INTRON3
E4
INTRON 4
5 INT5 E6
INTRON 6
EXON 7
3'
O47
O02
O01
O09
A201
A101
B101
H18
H2
H17
H29
H21
H30
H61
H34
H45
H27
H28
H42
H38
H49
H50
H32
H39
H26
H44
H62
H59
H46
H24
H58
H6
H52
H40
H3
H43
H56
H15
H33
H37
H7
H35
H31
H23
H13
H14
H8
H4
H41
H57
H48
H22
H54
H36
H11
H5
H55
H51
H53
H47
H9
H19
H10
H16
H12
H20
H25
H60

◀ **Fig. 2** Haplotypes reconstructed from the Seattle SNPs ABO sequences. *Top row* alleles at a chimpanzee allele sequence, *below* exon/intron structure. *Dots* indicate identity with the chimpanzee sequence. *I/-* stand for the long/short alleles in insertion/deletion polymorphisms and do not necessarily imply the direction of the actual mutation event. The "*I*" allele corresponds to the addition of CCCTTCCT in SNP1; SNP9, GAGGAATTGC CACAATTTTT TCCTGGCCTG CACC; SNP10, TGCACC; SNP23, A; SNP24, TG; SNP87, TAAA; SNP89, GGCAGTTT; SNP97, TAGTGGTGGGCG; SNP102, GG; SNP137, GTGTGGACAGAAG; SNP151, C; SNP154, CCC; SNP160, G; SNP165, T; SNP166, TGGGGCTCG; SNP201, C; SNP205, CACA; SNP206, CACA; SNP208, ACACACAGACACATAGA. Codons 266 and 268, which determine A/B activity are *boxed* in the B101 haplotypes. Δ261 is site 160. Lineages are *color-coded* (only in the online version of the figure) and labelled on the *right-hand side*: *magenta* for *O47*, *brown* for *O02*, *blue* for *O01*, *green* for *A101/A201/O09*, *red* for *B101*. Haplotypes carrying sequences of different colors may be the result of recombination or gene conversion. Absolute population frequencies are also indicated on the *right-hand side*, with *A* for African-Americans and *E* for European-Americans: for example, haplotype H29 is labelled "*A7E*" because it was found in one African-American and seven European-Americans

between A and B alleles appears more deeply subdivided into well-defined lineages.

Table 2 shows the lineage frequencies in European- and African-Americans. Again, African-Americans are more diverse. F_{ST} based on lineage frequencies is 0.1069 ($P < 0.001$). However, at the sequence level, F_{ST} between the two populations is 0.0573 ($P = 0.01$); this lower value can be explained by parallel substitutions and recombination among lineages. Both values are well within the known distribution of F_{ST}, which points to the absence of population-specific selective pressures, at least between Africans and Europeans.

Next, we tested for departures from the neutral model. Tajima's D was 1.982 ($P < 0.0001$) for the whole sample, 1.662 ($P = 0.0022$) for African-Americans and 2.383 ($P < 0.0001$) for European-Americans. It should be taken into account that only one, two, and six out of 320 genes had larger Tajima's D values in Seattle SNP genes respectively in the whole sample, African-Americans, and European-Americans. This shows that ABO has an extreme allele frequency spectrum as mesaured by Tajima's D, and that it is unlikely that demographic history, which acts on the whole genome, would be the only factor modelling this spectrum. Therefore, we can infer that balancing selection has had a role in preserving a diversity of ancient lineages. Along the gene (Fig. 6), Tajima's D shows significant peaks in intron 1, exon 2, intron 4, and the 3' region; about 30% of the total sequence length has significant (two-tailed $P < 0.05$) Tajima's D values, as ascertained with a sliding-window-specific heuristic (Ardell 2004). When considering the two populations separately, most of the significant windows in the overall analysis remained significant, and new significant peaks were detected in intron 1 for both populations, and in intron 2 to intron 3, intron 4, and intron 5 to intron 6 (including exon 6) only in European-Americans. It is possible, as discussed below, that linkage disequilibrium may be strong enough in this 23 Kb region to prevent the precise pinpointing of particular segments (or, even less, particular polymorphisms) as being the biological targets of balancing selection.

Fu and Li's F test with the chimpanzee sequence as outgroup was $F = 2.253$ overall, 2.037 in African-Americans and 2.693 in European-Americans, all significant with $P < 0.001$. This implies that the average difference between human sequences is larger than expected considering the divergence between humans and chimpanzees. Along the gene (data not shown), Fu and Li's F is high in the first half of intron 1 (as Tajima's D), but it also peaks at the very end of intron 1, in intron 3 and from exons 5 to 6.

The total divergence between human and chimpanzee is 0.0142. For comparison purposes, this total figure can be

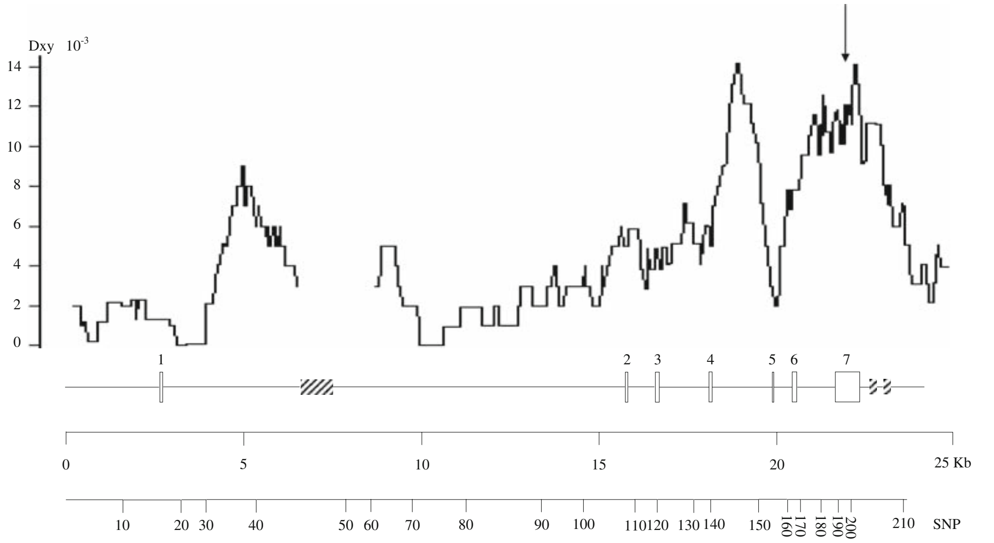


Fig. 3 Divergence between O01 and O02 alleles, expressed as Dxy (×10^{-3}). *X*-axis as in Fig. 1. The *arrow* indicates the position of Δ261

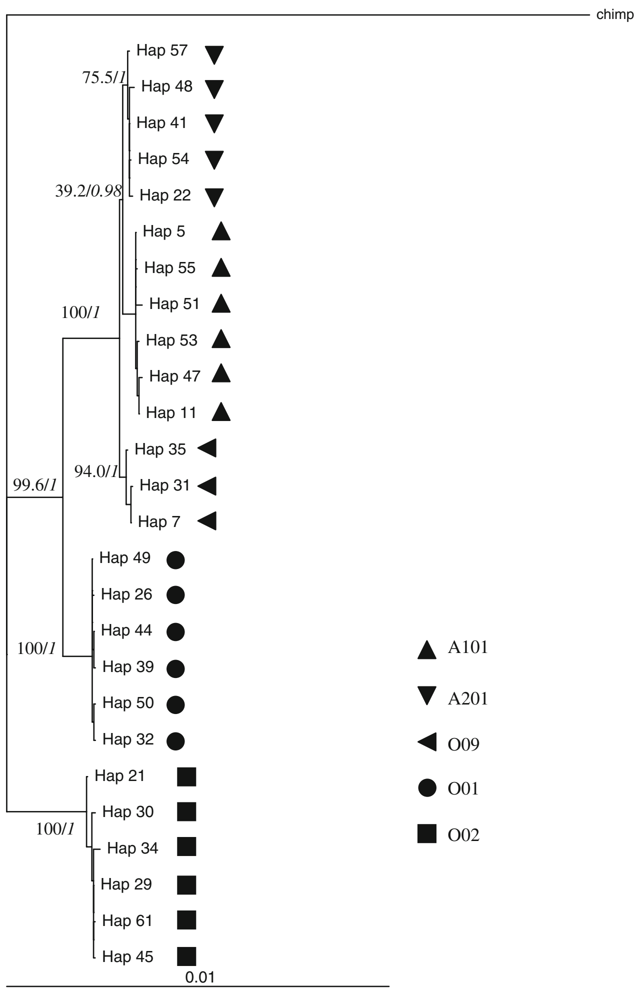


Fig. 4 Maximum-likelihood (ML) tree of the presumably non-recombinant ABO haplotypes in Seattle SNPs. Lineages are indicated by *symbols*. *Figures* indicate percent bootstrap support of the ML tree for main nodes, and, after the *slash*, the posterior probabilities obtained with a Bayesian model. Haplotypes are labelled as in Fig. 2

influenced by the proportion of coding and non-coding sequence. The non-coding divergence in *ABO* (Ki = 0.0141) is above the mean (0.0127) for 12,997 autosomal genes (The Chimpanzee Sequencing and Analysis Consortium 2005), but 20.4% of those have higher divergences than *ABO*. In the coding regions of *ABO*, the synonymous divergence (Ks = 0.0423) was larger than the genomewide mean (0.0143), though not extreme (3.6% of the genes have higher values). Finally, the nonsynonymous divergence was Ka = 0.0092, again larger than the mean but not out of the genomewide distribution (8.7% of genes have larger divergences). The Ka/Ks ratio was 0.211, and the McDonald and Kreitman's test was not significant ($P = 0.489$).

Recombination has conspicuously acted in the genealogy of *ABO* sequences, probably contributing to the differences in terminal branch length (see Fig. 5c); balancing selection probably has also had a role in shaping the genealogy of *ABO*. Under these circumstances, it is extremely complex to obtain reliable time estimates; therefore, what follows should be taken as a general indication of the time depth of *ABO* rather than an attempt to a precise chronology. From the average number of nucleotide changes to the tree root, and assuming that all of them have been generated by mutation, we obtain a total depth for the overall tree of 2.65 ± 0.50 Mya (million years ago), in the lower end of the interval estimated by Saitou and Yamamoto (1997) based on a few cDNA sequences. This is a very long tree: scaled in units of Ne generations, it would be over 13, or far beyond the expected depth of a neutral tree. However, this average age masks differences both across lineages and along the sequence. From nucleotide positions 1 to 6,000 (from the beginning of the sequence to 3.5 Kb into intron 1), the average time would be just 2.05 ± 0.67 Mya. In the central segment we defined above, the average time is 2.83 ± 0.70 Mya. Finally, in the 3′ segment, the genealogy becomes much longer, reaching an average 4.84 ± 0.85 Mya, close to the range of ages estimated for the human-chimpanzee split (5–6 Mya).

Population genetics of O allele diversity

Roubinet et al. (2004) sequenced 1,875 bp of the ABO gene, comprising exon 6, intron 6, and exon 7, in 814 O chromosomes from seven populations (Akan from the Ivory Coast, Berbers from Morocco, Basques from France and Spain, Han Chinese from Putien and Fujiou, and Cayapas and Aymaras from Bolivia). The basic descriptive parameters of sequence variation for *O* allele sequences are shown in Table 3. Haplotype (H) and nucleotide (π) variation are highest in the African Akans, a trend that is shown by a number of African populations in many (though not all) genes (Calafell et al. 1998; Mateu et al. 2001; Tishkoff et al. 1996, 1998), and that is compatible with a recent and African origin of modern humans. It should be noted that nucleotide diversity is extremely high in any population, but slightly lower than for the same region in Seattle SNPs (60.5×10^{-4}).

An analysis of the molecular variance (AMOVA) (Excoffier et al. 1992) among the seven populations, showed that, within the *O* alleles, 7.02% (significantly different from zero, $P < 10^{-5}$) of the genetic variation was accounted for by differences among populations. These values are below the average (~15%) but within the range observed for other genes (Barbujani et al. 1997; Romualdi et al. 2002), although this reference can depend on the number and distribution of populations analyzed. The fact that genetic variation within *O* allele sequences is relatively homogeneous among populations may be interpreted as the result of geographically homogeneous selective pressures,

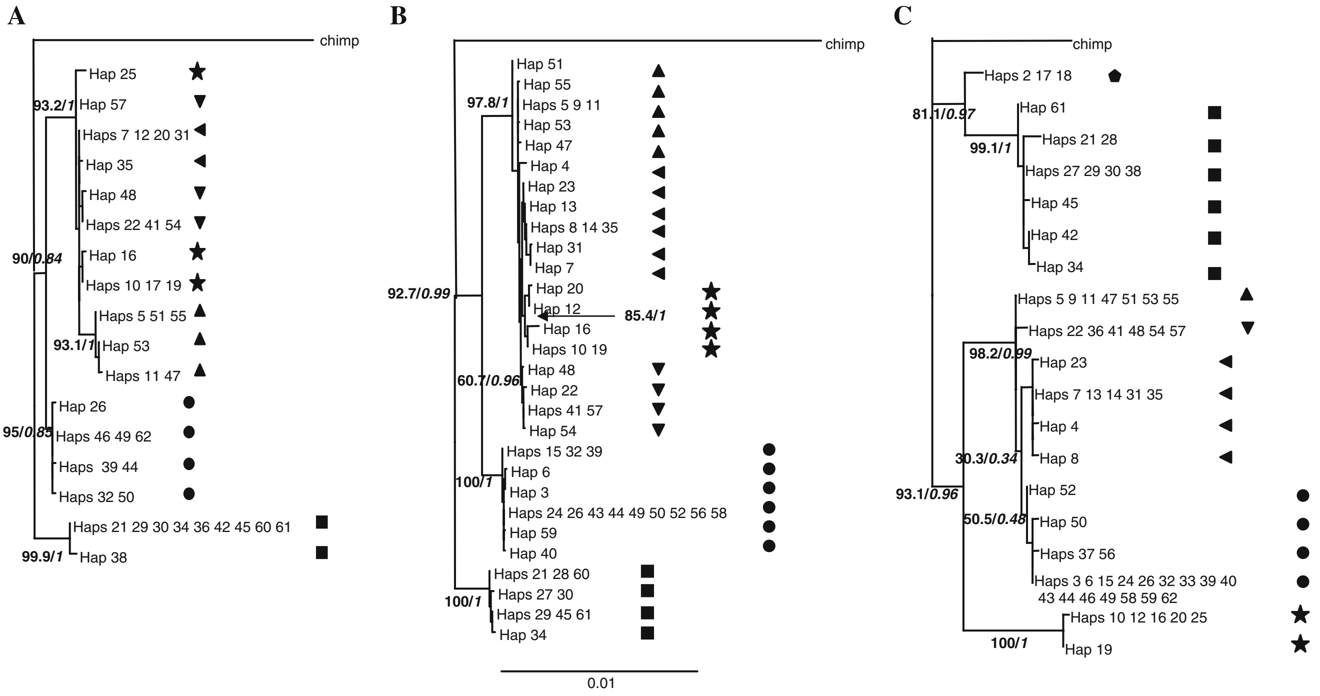


Fig. 5 Maximum-likelihood trees of the presumably non-recombinant ABO haplotypes in Seattle SNPs. **a** from positions 1 to 6,000, **b** from positions 6,001 to 20,124, **c** from positions 20,125 to 23,759. Recombinant haplotypes have not entered the trees corresponding to the region where the putative recombination event took place (for instance, haplotype 25 appears to have experienced gene conversion in the **b** region and was excluded from that tree, whereas it appears in segments **a** and **c**). This explains why the number of haplotypes is not the same across the three trees. *Symbols* are as in Fig. 4, with the addition of *stars* (B101) and *pentagons* (O47). Figures indicate percent bootstrap support of the ML tree for main nodes, and, after the slash, the posterior probabilities obtained with a Bayesian model

Table 2 Frequencies for the different ABO lineages in the Seattle SNP dataset

Lineage	Afr. Am.	Eur. Am.
O47	4 (0.083)	–
O02	8 (0.167)	11 (0.239)
O01	8 (0.167)	21 (0.457)
O09	11 (0.229)	–
A101	4 (0.083)	7 (0.152)
A201	2 (0.042)	7 (0.152)
B101	11 (0.229)	–

Each haplotype was allocated to the lineage it bore in the 5′ end of the sequence. See also Fig. 2

or of selective pressures that predate the expansion of anatomically modern humans.

Discussion

Analysis of two complementary data sets of *ABO* sequence variation has shown the depth and complexity of the genealogical relations among the allele lineages, highlighting a role for recombination; evidence of non-neutral evolution is manifest and points to balancing selection, and ancient lineages have been preserved even within the O null-allele class.

The *ABO* complete sequences produced by Seattle SNPs show a complex pattern in which the action of gene conversion and recombination is evident, and in which the genealogical patterns and the depth of the phylogeny change along the sequence. This is the standard behavior of the autosomal genome, in which recombination creates a mosaic of different gene genealogies with different evolutionary history. The present challenge in understanding and using the genome variation in autosomes is highly dependent of our ability to recognize past recombination events in the present genetic structure. Nonetheless, a number of distinct lineages were revealed: (1) *O02*, and, as a quite distinct subclade, *O47,* although divergence from *O02* extended only to the 3′ and 5′ ends of the gene; (2) *O01*, which, beyond the exon 6 to exon 7 region was clearly separated from *A101*; (3) the *A101*/*A201*/*O09* group; and (4) *B101*, which, as described by Seltsam et al. (2003), carries an *A201*-like sequence upstream of exon 6. This means that, in any case, and with more extensive sequence, we could not recover the five-lineage phylogeny proposed by Seltsam et al. (2003). Beyond the typological description, a phylogenetic analysis allows to trace the evolution of ABO

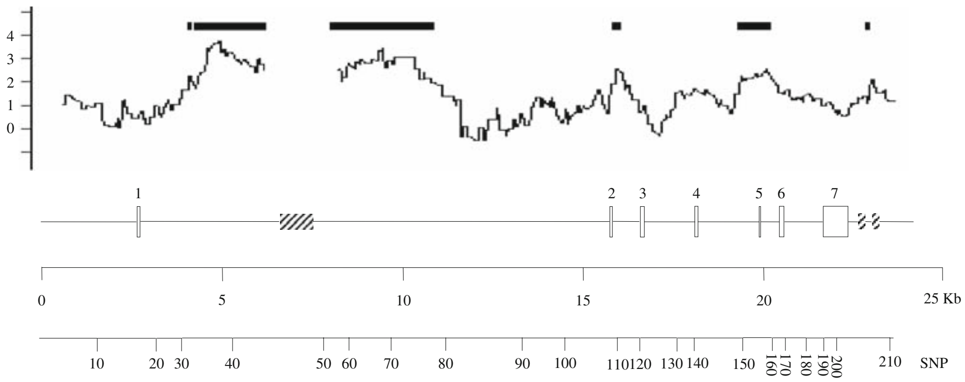


Fig. 6 Tajima's *D* along the ABO sequence. Window length and *X*-axis as in Fig. 1. Statistically significant peaks are marked with *black bars* above them

Table 3 Sequence variability parameters in *O* alleles from Roubinet et al. (2004)

	N	*k*	*S*	*H*	*π*
Akans	136	15	42	0.868 ± 0.011	62
Berbers	78	9	33	0.656 ± 0.042	52
Basques	220	11	40	0.583 ± 0.026	51
Putien	94	2	19	0.491 ± 0.019	50
Fujiou	86	2	19	0.506 ± 0.009	52
Cayapas	74	5	21	0.580 ± 0.034	52
Aymaras	126	5	22	0.563 ± 0.035	42
Total	814	23	51	0.666 ± 0.011	55

N sample size, *k* number of different haplotypes, *S* number of polymorphic sites, *H* haplotype diversity, *π* nucleotide diversity

sequences. Based on the shared functionality of the *A* alleles in humans, chimpanzees, and bonobos (Blancher and Socha 1997), and the different inactivation mechanism of *O* alleles in chimpanzees (Kermarrec et al. 1999), it can be proposed that the ancestral human sequence was *A*, as noted also by Saitou and Yamamoto (1997). The first non-A allele to appear was *B101*, which, using the divergence between *A101* and *B101* in the region downstream of exon 6, can be dated at ~3.5 Mya (95% CI: 2.64–4.36 Mya) (henceforth, all dates given are based on the raw mean divergence between lineages and on the substitution rate described in the "Methods" section; all caveats to dating lineages in ABO discussed above apply). This age would place the origin of B well after the split of the human and chimpanzee lineages, and confirms the absence of extense trans-specific polymorphism (beyond the functional substitutions at codons 266 and 268 shared by humans, gorillas, and orangutans) as modelled by Wiuf et al. (2004). The next oldest allele appears to be *O02*, which may have diverged from *A101* ~2.5 Mya (95% CI: 2.20–2.80 Mya). One of the mutations that appeared in the *O02* branch and that inactivated the gene is Δ261. As discussed above, the divergence pattern between *O02* and *O01* makes it unlikely that these two lineages share Δ261 from a common ancestor; on the contrary, *O01* seems to have sprung from *A101* ~1.15 Mya (95% CI: 0.95–1.35 Mya), and, after that, acquired Δ261 by mutation or gene conversion. This position falls in an exonic, non-repeat region, and the sequence context does not seem to facilitate repeated deletion events. Note that a third lineage acquired Δ261; namely, *O09*, which is closely related to *A101*, from which it may have separated ~316 thousand years ago (Kya) (95% CI: 231–401 Kya). *A101* also generated *A201* ~288 Kya (95% CI: 205–371 Kya); shortly after that (at ~260 Kya, 95% CI: 167–353 Kya), *A201* recombined with *B101*, obliterating the original *B101* sequence upstream of exon 6; sometime later, the previous non-recombinant forms of *B101* were lost (or reduced to such low frequencies that they have not yet been sampled).

We have shown that different regions of the *ABO* gene show different phylogenetic depths, with variation at the region coding for the catalytic domain carrying the amount of variation that would require the longest time to accumulate. This, with other evidence discussed below, can be interpreted as the footprint of balancing selection acting on the *A*, *B*, and *O* alleles. Stajich and Hahn (2005) found extremely positive Tajima's *D* values in complete samples of ABO variation from European- and African-Americans, which they interpreted as balancing selection involving the three (*A*, *B*, and *O*) main alleles. We have replicated that result, put it into the context of the functional lineages of *ABO*, and found that Tajima's *D* and Fu and Li's *F* significantly positive values extend beyond the functionally crucial exon 7 and peak elsewhere, particularly in intron 1. Moreover, extreme sequence divergence can also be found between lineages that are functionally equivalent such as *O01* and *O02*. Neutrality statistics based on amino acid divergence from the chimpanzee (such as the Ka/Ks ratio and McDonald and Kreitman's test) were not significant: balancing selection may have acted on only two amino acid replacements, or on any change that inactivated the gene.

The highest peaks for Tajima's *D* are found in intron 1; note, though, that significant peaks were also found elsewhere, and that significance depends on the population being analyzed. SNPs in intron 1 were not found to fall in any of the putatively functional categories (namely, triplex sequences, intron boundaries, and mouse-conserved regions) considered by Pupasuite (http://pupasuite.bioinfo.cipf.es/) (Conde et al. 2006). Actually, most of the human ABO intron 1 sequence is comprised of repeated elements, and it is unlikely that intron 1 has a major biological role in *ABO* expression or functionality. Thus, we interpret the Tajima's *D* peaks at intron 1 as the result of the random accumulation of old, frequent, and probably neutral mutations in a context of balancing selection and linkage disequilibrium. Even though most lineages at *ABO* are quite ancient, this is a relatively short (23 Kb) region, with apparently an average recombination rate. LD would, thus, reduce the precision with which a particular target of balancing selection can be spotted, and generate random peaks and troughs in neutrality statistics, fluctuating around a significantly positive baseline value.

Long extended haplotypes have been detected upstream of the *ABO* gene in the HapMap samples (Sabeti et al. 2006), as well as a segment with an extremely low F_{ST} value. Fry et al. (2007) used these two pieces of evidence to conclude that balancing selection at *ABO* may have acted upstream of the antigen-defining region of exon 6 to exon 7. Note, though, that selection at ABO can be extremely complex and local, depending on endemic pathogens (as discussed below), and that is unlikely to have operate globally with exactly the same selective coefficients, as required to maintain low interpopulation differentiation (F_{ST}). Using extensive resequencing data and a more accurate phylogeographic analysis, we can confirm that, indeed, the strongest signal for selection at *ABO* appears to be at intron 1. Actually, our approach may be more powerful to detect an apparently old mutation event, since resequencing allows the full description of the spectrum of allele frequency variation (Soldevila et al. 2005), which is more likely to have retained and ancient signal of selection (see, for instance, Fig. 1 in ref. (Sabeti et al. 2006)). On the contrary, Bubb et al. (2006) used whole genome simulations to conclude that polymorphism at *ABO* was not significantly different from that expected under neutral expectations; however, they only used the last four exons of the *ABO* gene and therefore may have lacked sufficient statistical power as given by additional sequence.

It is highly unlikely that a small, stable population such as that of the human lineage before the modern expansions could have retained three neutral alleles at a locus for millions of years. The probability that three chromosomes sampled at random do not coalesce in 125,000 generations (2.5 Mya at 20 years/generation) in a stable, panmictic population of $Ne = 10{,}000$ is $\sim 10^{-17}$ (Eq. 3 in Nordborg (2001)), or that of four alleles not coalescing in 1.15 Mya is $\sim 10^{-15}$. HLA could be a model system for *ABO*, in which a selective advantage for heterozygotes could have led to the conservation of an extremely diverse allele repertoire. However, there is an obvious difference between *ABO* and HLA: in the former, a null (but phenotypically relevant) allele is among those maintained at a high frequency, and, further to that, variation *within* that null allele seems to be rather ancient. In a context of balancing selection in which the null alleles may have been selected for, as discussed below, *O01*, created by a parallel mutation or gene conversion event may have been driven to high frequencies at the expense of A and B alleles but not of the pre-existing *O02*. This may partially explain why two very divergent but functionally equivalent alleles may coexist.

It may be the case that the selective pressures are not exclusively exerted on the A and B antigens, but on the anti-A and anti-B antibodies as well. AB individuals lack both specificities, which may be a selective disadvantage, as hinted by several clues: no mammal species presents exactly *A* and *B*; only the mouse glycosyltranferase (which is monomorphic) produces both A and B antigens, and, in humans, although such alleles exist [*cis-AB* and *B(A)*], they are exceedingly rare. Thus, not only a simple model of heterozygote advantage between *A* and *B* seems to be ruled out, but on the contrary, AB phenotypes may have been selected against. That would lead to an unstable equilibrium and to the loss of either *A* or *B*. However, the situation can be stabilized with a recessive null allele, which increases the relative frequencies of *A* and *B* over *AB* individuals by adding the *AO* and *BO* heterozygote individuals. Alternatively, frequency-depending selection can have operated on the *ABO* locus. Since different pathogens anchor to different antigens, their transmissibility and the mortality they inflict on populations might increase with the frequency of their *ABO* allele product (sugar chain) receptor in the population, thus favoring the other alleles. The following epidemic may have had the opposite effect, and this pattern of selection would keep the high polymorphism at *ABO,* including a large number of different lineages. Such selective events would have contributed to the raise in frequency of any lineage carrying Δ261. In a third scenario, polymorphic species of pathogens can also stabilize *ABO* polymorphism in its host depending on the selection coefficients of each strain of pathogen against each host genotype (Fischer et al. 1998).

Conclusions

We have unraveled the phylogeny of the several functional and non-functional lineages in the human *ABO* gene, and have highlighted how that phylogeny changes along the gene due to past recombination events. We have undertaken the most detailed evolutionary analysis of the human *ABO* gene so far, and have found clear evidence of balancing selection in it. We propose several hypotheses for the cause of that selection, which most likely involved interactions with multiple pathogens at different geographic regions and time scales.

The complexity of a clear Mendelian trait as the ABO blood group comes through the intricacy of cell surface interactions, and, at the molecular level has the clear footprint of a dynamics driven by selection through a deep past in human history, which is difficult to isolate and measure due to the complex dynamics of the genome, mainly due to recombination.

Acknowledgments The authors wish to especially thank Deborah Nickerson at the University of Washington, for making the excellently curated Seattle SNP sequence database openly available to researchers. This work was supported by grants BFU2004-02002 and BFU2007-63657 from the Spanish Ministry of Education and Science, and DURSI, Generalitat de Catalunya (Grup de Recerca Consolidat 2005SGR00608), as well as with funds from Etablissement Français du Sang (EFS) Pyrénées-Méditerranée and Centre Atlantique, and from Ministère Français de la Recherche (EA3034).

References

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol 2:e286

Ardell DH (2004) SCANMS: adjusting for multiple comparisons in sliding window neutrality tests. Bioinformatics 20:1986–1988

Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL (1997) An apportionment of human DNA diversity. Proc Natl Acad Sci USA 94:4516–4519

Bertranpetit J, Calafell F (1996) Genetic and geographical variability in cystic fibrosis: evolutionary considerations. Ciba Found Symp 197:97–114

Blancher A, Socha WW (1997) The ABO, Hh and Lewis blood groups in man and nonhuman primates. In: Blancher A, Jan Klein J, Socha WW (eds) Molecular biology and evolution of blood group and mhc antigens in primates. Springer, Heidelberg, pp 30–92

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 262:1892–1895

Bubb KL, Bovee D, Buckley D, Haugen E, Kibukawa M, Paddock M, Palmieri A, Subramanian S, Zhou Y, Kaul R, Green P, Olson MV (2006) Scan of human genome reveals no new Loci under ancient balancing selection. Genetics 173:2165–2177

Calafell F, Shuster A, Speed WC, Kidd JR, Kidd KK (1998) Short tandem repeat polymorphism evolution in humans. Eur J Hum Genet 6:38–49

Chester MA, Olsson ML (2001) The ABO blood group gene: a locus of considerable genetic diversity. Transfus Med Rev 15:177–200

Conde L, Vaquerizas J, Dopazo H, Arbiza L, Reumers J, Rousseau F, Schymkowitz J, Dopazo J (2006) PupaSuite: finding functional SNPs for large-scale genotyping purposes. Nucleic Acids Res 34:621–625

Cserti CM, Dzik WH (2007) The ABO blood group system and Plasmodium falciparum malaria. Blood 110:2250–2258

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491

Fischer C, Jock B, Vogel F (1998) Interplay between humans and infective agents: a population genetic study. Hum Genet 102:415–422

Fry AE, Griffiths MJ, Auburn S, Diakite M, Forton JT, Green A, Richardson A, Wilson J, Jallow M, Sisay-Joof F, Pinder M, Peshu N, Williams TN, Marsh K, Molyneux ME, Taylor TE, Rockett KA, Kwiatkowski DP (2007) Common variation in the ABO glycosyltransferase is associated with susceptibility to severe Plasmodium falciparum malaria. Hum Mol Genet 17:567–576

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133:693–709

Gagneux P, Varki A (1999) Evolutionary considerations in relating oligosaccharide diversity to biological function. Glycobiology 9:747–755

Grunnet N, Steffensen R, Bennett EP, Clausen H (1994) Evaluation of histo-blood group ABO genotyping in a Danish population: frequency of a novel O allele defined as O2. Vox Sang 67:210–215

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755

Kermarrec N, Roubinet F, Apoil PA, Blancher A (1999) Comparison of allele O sequences of the human and non-human primate ABO system. Immunogenetics 49:517–526

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002) A high-resolution recombination map of the human genome. Nat Genet 31:241–247

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD (2006) Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol 23:1891–1901

Kreitman M, Di Rienzo A (2004) Balancing claims for balancing selection. Trends Genet 20:300–304

Landsteiner K (1901) Über Agglutinationserscheinungen normalen menschlichen Blutes. Wien Klin Wochenschr 14:1132–1134

Lindesmith L, Moe C, Marionneau S, Ruvoen N, Jiang X, Lindblad L, Stewart P, LePendu J, Baric R (2003) Human susceptibility and resistance to Norwalk virus infection. Nat Med 9:548–553

Marionneau S, Cailleau-Thomas A, Rocher J, Le Moullac-Vaidye B, Ruvoen N, Clement M, Le Pendu J (2001) ABH and Lewis histo-blood group antigens, a model for the meaning of oligosaccharide diversity in the face of a changing world. Biochimie 83:565–573

Marionneau S, Ruvoen N, Le Moullac-Vaidye B, Clement M, Cailleau-Thomas A, Ruiz-Palacois G, Huang P, Jiang X, Le Pendu J (2002) Norwalk virus binds to histo-blood group antigens present on gastroduodenal epithelial cells of secretor individuals. Gastroenterology 122:1967–1977

Mateu E, Calafell F, Lao O, Bonne-Tamir B, Kidd JR, Pakstis A, Kidd KK, Bertranpetit J (2001) Worldwide genetic analysis of the CFTR region. Am J Hum Genet 68:103–117

Morral N, Bertranpetit J, Estivill X, Nunes V, Casals T, Giménez J, Reis A, Varon-Mateeva R, Macek M, Kalaydjieva L, Angelicheva D, Dancheva R, Romeo G, Russo MP, Garnerone S, Restagno G, Ferrari M, Magnani C, Claustres M, Desgeorges M, Schwartz M, Schwarz M, Dallapiccola B, Novelli G, Ferec C, de Arce M, Nemeti M, Kere J, Anvret M, Dahl N, Ferak V (1994) Tracing the origin of the major cystic fibrosis mutation (ΔF508) in European populations. Nature Genet 7:169–175

Mourant AE (1954) The ABO blood groups. Blackwell, Oxford

Mourant AE, Kopec AC, Domaniewska-Sobczak K (1978) Blood groups and diseases. Oxford University Press, Oxford

Nordborg M (2001) Coalescent theory. In: Balding J, Bishop M, Cannings C (eds) Handbook of statistical genetics. Wiley, Chichester, pp 179–208

Ogasawara K, Bannai M, Saitou N, Yabe R, Nakata K, Takenaka M, Fujisawa K, Uchikawa M, Ishikawa Y, Juji T, Tokunaga K (1996a) Extensive polymorphism of ABO blood group gene: three major lineages of the alleles for the common ABO phenotypes. Hum Genet 97:777–783

Ogasawara K, Yabe R, Uchikawa M, Saitou N, Bannai M, Nakata K, Takenaka M, Fujisawa K, Ishikawa Y, Juji T, Tokunaga K (1996b) Molecular genetic analysis of variant phenotypes of the ABO blood group system. Blood 88:2732–2737

Ogasawara K, Yabe R, Uchikawa M, Nakata K, Watanabe J, Takahashi Y, Tokunaga K (2001) Recombination and gene conversion-like events may contribute to ABO gene diversity causing various phenotypes. Immunogenetics 53:190–199

Olsson ML, Chester MA (1996a) Evidence for a new type of O allele at the ABO locus, due to a combination of the A2 nucleotide deletion and the Ael nucleotide insertion. Vox Sang 71:113–117

Olsson ML, Chester MA (1996b) Frequent occurrence of a variant O1 gene at the blood group ABO locus. Vox Sang 70:26–30

Olsson ML, Chester MA (2001) Polymorphism and recombination events at the ABO locus: a major challenge for genomic ABO blood grouping strategies. Transfus Med 11:295–313

Olsson ML, Guerreiro JF, Zago MA, Chester MA (1997) Molecular analysis of the O alleles at the blood group ABO locus in populations of different ethnic origin reveals novel crossing-over events and point mutations. Biochem Biophys Res Commun 234:779–782

Olsson ML, Santos SE, Guerreiro JF, Zago MA, Chester MA (1998) Heterogeneity of the O alleles at the blood group ABO locus in Amerindians. Vox Sang 74:46–50

Ramírez-Soriano A, Ramos-Onsins SE, Rozas J, Calafell F, Navarro A (2008) Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. Genetics 179:555–567

Romualdi C, Balding D, Nasidze IS, Risch G, Robichaux M, Sherry ST, Stoneking M, Batzer MA, Barbujani G (2002) Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. Genome Res 12:602–612

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574

Roubinet F, Kermarrec N, Despiau S, Apoil PA, Dugoujon JM, Blancher A (2001) Molecular polymorphism of O alleles in five populations of different ethnic origins. Immunogenetics 53:95–104

Roubinet F, Despiau S, Calafell F, Jin F, Bertanpetit J, Saitou N, Blancher A (2004) Evolution of the O alleles of the human ABO blood group gene. Transfusion 44:707–715

Rowe JA, Handel IG, Thera MA, Deans AM, Lyke KE, Kone A, Diallo DA, Raza A, Kai O, Marsh K, Plowe CV, Doumbo OK, Moulds JM (2007) Blood group O protects against severe Plasmodium falciparum malaria through the mechanism of reduced rosetting. Proc Natl Acad Sci USA 104:17471–17476

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496–2497

Ruiz-Palacios GM, Cervantes LE, Ramos P, Chavez-Munguia B, Newburg DS (2003) Campylobacter jejuni binds intestinal H(O) antigen (Fuc alpha 1, 2Gal beta 1, 4GlcNAc), and fucosyloligosaccharides of human milk inhibit its binding and infection. J Biol Chem 278:14112–14120

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES (2006) Positive natural selection in the human lineage. Science 312:1614–1620

Saillard J, Forster P, Lynnerup N, Bandelt HJ, Nørby S (2000) mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. Am J Hum Genet 67:718–726

Saitou N, Yamamoto F (1997) Evolution of primate ABO blood group genes and their homologous genes. Mol Biol Evol 14:399–411

Seltsam A, Hallensleben M, Kollmann A, Blasczyk R (2003) The nature of diversity and diversification at the ABO locus. Blood 102:3035–3042

Soldevila M, Calafell F, Heigason A, Stefansson K, Bertranpetit J (2005) Assessing the signatures of selection in PRNP from polymorphism data: results support Kreitman and Di Rienzo's opinion. Trends Genet 21:389–391

Stajich JE, Hahn MW (2005) Disentangling the effects of demography and selection in human history. Mol Biol Evol 22:63–73

Stephens M, Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76:449–462

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989

Swerdlow DL, Mintz ED, Rodriguez M, Tejada E, Ocampo C, Espejo L, Barrett TJ, Petzelt J, Bean NH, Seminario L, Tauxe RV (1994) Severe life-threatening cholera associated with blood group O in Peru: implications for the Latin American epidemic. J Infect Dis 170:468–472

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Takahata N, Satta Y, Klein J (1995) Divergence and population size in the lineage leading to modern humans. Theor Popul Biol 48:198–221

The Chimpanzee Sequencing Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87

Tishkoff SA, Dietzch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonné-Tamir B, Santachiara-Benerecetti S, Moral P, Krings M, Pääbo S, Watson E, Risch N, Jenkins T, Kidd KK (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271:1380–1387

Tishkoff SA, Goldman A, Calafell F, Speed WC, Deinard AS, Bonne-Tamir B, Kidd JR, Pakstis AJ, Jenkins T, Kidd KK (1998) A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. Am J Hum Genet 62:1389–1402

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

Xia X, Xie Z (2001) DAMBE: software package for data analysis in molecular biology and evolution. J Hered 92:371–373

Yamamoto F (2000) Molecular genetics of ABO. Vox Sang 78(Suppl 2):91–103

Yamamoto F (2004) Review: ABO blood group system–ABH oligosaccharide antigens, anti-A and anti-B, A and B glycosyltransferases, and ABO genes. Immunohematol 20:3–22

Yamamoto F, Clausen H, White T, Marken J, Hakomori S (1990a) Molecular genetic basis of the histo-blood group ABO system. Nature 345:229–233

Yamamoto F, Marken J, Tsuji T, White T, Clausen H, Hakomori S (1990b) Cloning and characterization of DNA complementary to human UDP-GalNAc: Fuc alpha 1—2Gal alpha 1—3GalNAc transferase (histo-blood group A transferase) mRNA. J Biol Chem 265:1146–1151

Yamamoto F, McNeill PD, Yamamoto M, Hakomori S, Bromilow IM, Duguid JK (1993) Molecular genetic analysis of the ABO blood group system: 4. Another type of O allele. Vox Sang 64:175–178

Yamamoto F, McNeill PD, Hakomori S (1995) Genomic organization of human histo-blood group ABO genes. Glycobiology 5:51–58

Yip SP (2000) Single-tube multiplex PCR-SSCP analysis distinguishes 7 common ABO alleles and readily identifies new alleles. Blood 95:1487–1492

Yip SP (2002) Sequence variation at the human ABO locus. Ann Hum Genet 66:1–27