

The Functional A Allele Was Resurrected via Recombination in the Human ABO Blood Group Gene

Takashi Kitano,^{*1} Antoine Blancher,² and Naruya Saitou³

¹Department of Biomolecular Functional Engineering, College of Engineering, Ibaraki University, Hitachi, Japan

²Laboratoire d'Immunogénétique Moléculaire (LIMT, EA3034), Faculté de Médecine Purpan, Université Paul Sabatier, Toulouse III, France

³Division of Population Genetics, National Institute of Genetics, Mishima, Japan

*Corresponding author: E-mail: tkitano@mx.ibaraki.ac.jp.

Associate editor: Naoko Takezaki

Abstract

Functional A and B alleles are distinguished at two critical sites in exon 7 of the human ABO blood group gene. The most frequent nonfunctional O alleles have one-base deletion in exon 6 producing a frameshift, and it has the A type signature in two critical sites in exon 7. Previous studies indicated that B and O alleles were derived from A allele in human lineage. In this study, we conducted a phylogenetic network analysis using six representative haplotypes: A101, A201, B101, O01, O02, and O09. The result indicated that the A allele, possibly once extinct in the human lineage a long time ago, was resurrected by a recombination between B and O alleles less than 300,000 years ago.

Key words: ABO blood group, phylogenetic network, recombination.

Introduction

The human ABO blood group system consists of three allele groups: A, B, and O. The A and B alleles, differing at two critical sites, code for glycosyltransferases, which transfer N-acetylgalactosamine and galactose, respectively, to a common precursor called H (Yamamoto et al. 1990; Yamamoto and Hakomori 1990) (for an updated review of the ABO, see also Yamamoto et al. 2012). The most frequent null O alleles have a point deletion ($\Delta 261$) in exon 6 (Yamamoto 2000), which induces a frameshift, resulting in a truncated protein deprived of any glycosyltransferase activity (Yamamoto et al. 1990). The main conclusions of the previous studies about the evolution of the ABO gene in the human lineage (e.g., Saitou and Yamamoto 1997; Seltsam et al. 2003; Roubinet et al. 2004; Calafell et al. 2008) can be summarized as follows. Major alleles A101, B101, O01, and O02 are observed in most human populations. The ancestral allele was A, and B diverged from A with substitutions on the two critical sites in exon 7, and O diverged from A with one-base deletion in exon 6, followed by the divergence of O01 and O02. These major alleles are maintained by some kind of positive selection. Although this study agrees with this rough outline, there is a further finding: the current A type, that is A101, had been formed by recombination between B101 and O01. We reveal this by conducting a phylogenetic network analysis, which is a powerful tool to find recombination events.

Materials and Methods

We used the ABO data produced by the Seattle SNPs project (Akey et al. 2004), which is a set of 90 23-kb sequences in European American and African American. These data are the same as those used by Calafell et al. (2008), except

that sequences of O47, a minor haplotype, were excluded. A distance-based phylogenetic network (supplementary fig. S1, Supplementary Material online) of 58 haplotypes (from 90 sequences) of the human ABO blood group genes (supplementary table S1, Supplementary Material online) was constructed by using the Neighbor-Net method (Bryant and Moulton 2004) implemented in the SplitsTree4 program (Huson and Bryant 2006). When a gorilla sequence (Kitano T, Kim CG, Kohara Y, Saitou N, unpublished data) was included in the phylogenetic network (supplementary fig. S1, Supplementary Material online), it clustered with the chimpanzee sequence in an outgroup to all the human sequences. Thus, we decided to use only the chimpanzee sequence (supplementary table S1, Supplementary Material online) as the outgroup in further analyses. The six representative human haplotypes (O02_H29, O01_H49, O09_H07, A201_H41, A101_H11, and B101_H19), which represent the most frequent alleles of each haplogroup (Calafell et al. 2008), were used for further analyses (fig. 1). Based on the phylogenetic network of 58 haplotype sequences of the human ABO blood group genes (supplementary fig. S1, Supplementary Material online), it is clearly shown that these six representative haplotypes are not outliers but are representative of each haplogroup.

Phylogenetic networks were also constructed manually following the procedure of Bandelt (1994) and Saitou and Yamamoto (1997). Phylogenetic trees were constructed by using the neighbor joining method (Saitou and Nei 1987) with *p* distance. MEGA version 4 (Tamura et al. 2007) was used to edit variable positions, which were used in the construction. Gene conversion among three haplotypes (O02_H29, O01_H49, and B101_H19) was detected by means of GENECONV (Sawyer 1989). Because it is clear

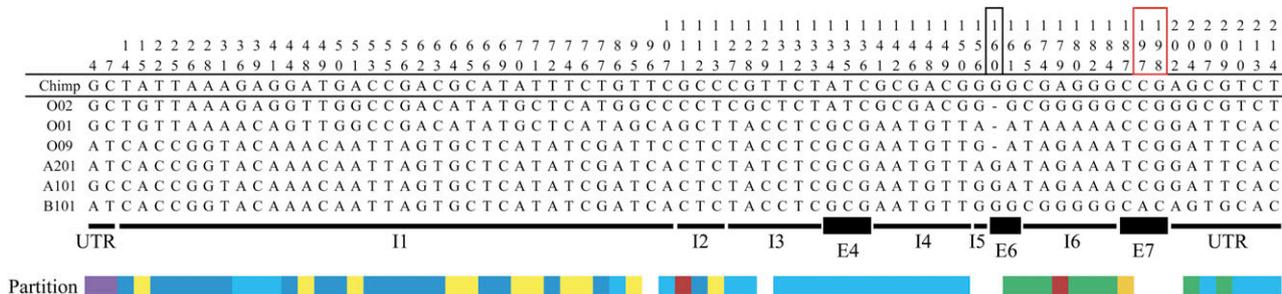


Fig. 1. Parsimony informative sites for six major haplotypes (O02, O01, O09, A201, A101, and B101). Corresponding chimpanzee A sites are also shown. Site numbers followed figure 2 of Calafell et al. (2008). Two sites (197 and 198) enclosed by a red square are two critical sites (codons 266 and 268), which determine A/B activity in exon 7. The site 160 enclosed by a black square is $\Delta 261$ in exon 6. Some exons are shown by boxes below. Untranslated regions (UTRs), some exons (E), and introns (I) are indicated. Partitions indicate dividing patterns: dark blue ([A101, A201, O09, B101], [ChimpA, O02, O01]), yellow ([O02, O01], [ChimpA, A101, A201, O09, B101]), purple ([Chimp, O02, O01, A101], [O09, A201, B101]), light blue ([O01, A101, A201, O09, B101], [ChimpA, O02]), green ([O01, A101, A201, O09], [ChimpA, O02, B101]), brown ([Chimp, O01], [O02, O09, A201, A101, B101]), orange ([Chimp, O02, O01, A101, B101], [O09, A201]).

that A101 as well as with A201 and O09 evolved from a recombinant allele, three haplotypes were excluded from the search of gene conversion analysis.

Results and Discussion

Six representative human ABO haplotype sequences were selected (see Materials and Methods; fig. 1) for the phylogenetic network analysis. As shown in figure 1, 11 sites are shared by only O01 and O02 (yellow partitions). Although O09 shares $\Delta 261$ with O01 and O02 alleles, its sequence is, by and large, quite similar to A101. Thus, the O09 allele most probably evolved from an A101-like common ancestral allele by a gene conversion on exon 6, introducing $\Delta 261$ from some other O allele. The enzyme encoded by the allele A201 is a weaker NACGal-transferase than the enzyme encoded by the A101 allele and has 21 supplementary amino acids of the C-terminal compared with the A101 protein because of a frameshift caused by a point deletion in the antepenultimate codon resulting (Yamamoto et al. 1992).

Large reticulations are observed in this phylogenetic network of the six haplotypes (fig. 2). Kitano et al. (2009) showed the relationship among recombinant and parental alleles in a phylogenetic network. To explain how to infer a recombination event from a phylogenetic network, we use model data (fig. 3). First, an ancestral sequence (anc) produces two sequences (p1 and p2) by accumulating five (indicated by blue letters in fig. 3A and blue lines in fig. 3B) and four substitutions (indicated by red letters in fig. 3A and red lines in fig. 3B), respectively. Then, a recombination between sites 6 and 7 produces two recombinants, r1 and r2 (fig. 3A). Assuming that r1 and r2 were produced by a single recombination event, transmission of both recombinant alleles to the next generation is highly improbable. We assumed that r2 had disappeared. After the recombination, three nucleotide substitutions (sites 7, 12, and 13 indicated by green letters in fig. 3A and green lines in fig. 3B) accumulate to p1, p2, and r1, respectively, and also three nucleotide substitutions (sites 3, 10, and 14 indicated by orange letters in fig. 3A and orange lines in

fig. 3B) accumulate to produce an outgroup (out). The phylogenetic network (fig. 3B) represents the relationship of the three present-day alleles (p1, p2, and r1) and an outgroup allele (out). Two parental alleles (p1 and p2) are located on opposing vertices of the rectangle with long external branches, whereas a recombinant allele (r1) has a short external branch and is located on the vertex opposing to the outgroup allele (out). A nucleotide substitution (site 7) on the short external branch of a recombinant allele (r1) is regarded as a nucleotide substitution, which occurs after the recombination event. In contrast, long external branches of two parental alleles (p1 and p2) contain nucleotide substitutions both before (9 and 15 for p1, and 1 and 6 for p2) and after (13 for p1 and 12 for p2) the recombination event.

Following this model, we can infer that A101 is a recombinant lineage, whereas B101 and O01 are parental allele lineages because O01 and B101 are located on opposing vertices of the rectangle with long external branches, whereas A101 has a short external branch and is located on the vertex opposing to the outgroup (fig. 2). In fact,

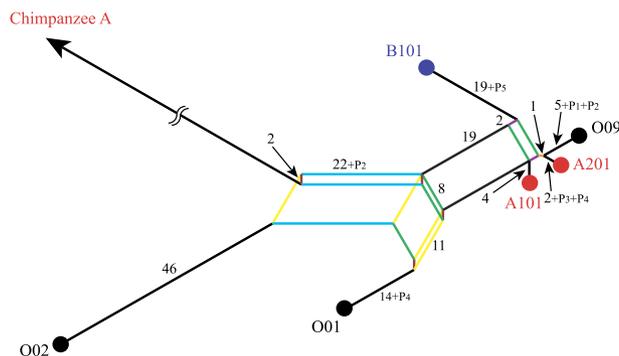


Fig. 2. The phylogenetic network for six major haplotypes (O02, O01, O09, A101, A201, and B101) of the human ABO genes with chimpanzee A. The color of each branch indicates each partition in figure 1. Numbers on each branch denote numbers of substitutions. Five parallel substitutions are also shown (site 96: P1, site 107: P2, site 129: P3, site 156: P4, site 202: P5, in which P1, P3, and P5 were parallel with chimpanzee A).

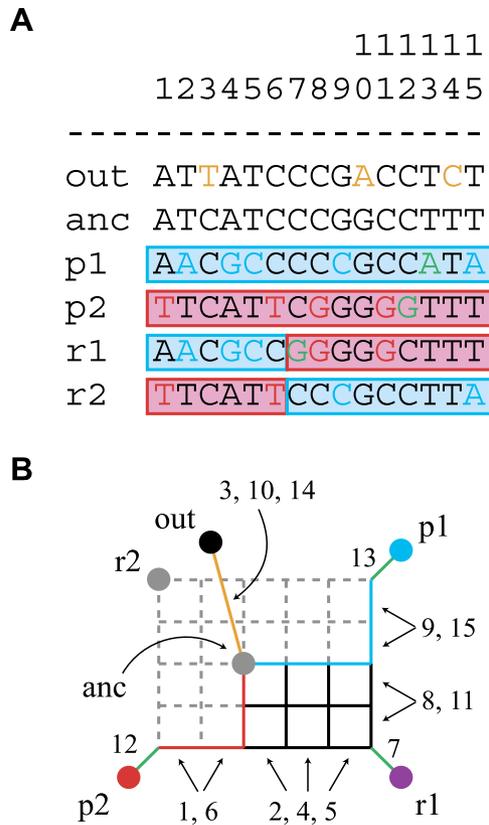


Fig. 3. Explanation of a recombination in a phylogenetic network by using model data. A nucleotide sequence data (A) and a corresponding phylogenetic network (B) are shown. See text for the details.

the upstream region of A101 is similar to B101, whereas its downstream region is similar to O01 (fig. 4A). The recombination occurred immediately after $\Delta 261$ or site 160 in figure 4A. The phylogenetic trees of the upstream and the downstream regions (fig. 5) clearly show different

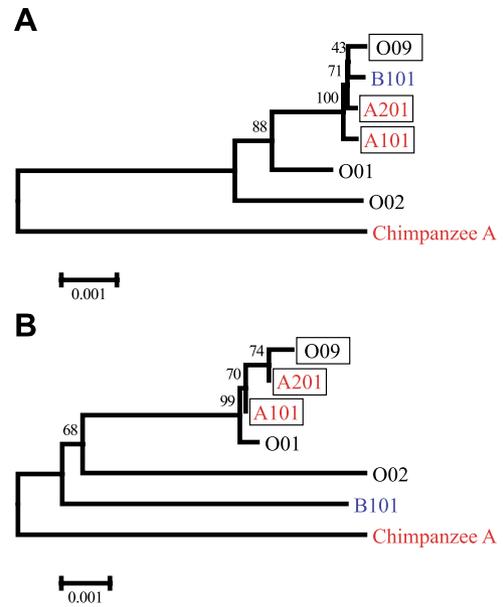


Fig. 5. Phylogenetic trees of the upstream region (A) and the downstream region (B) of the data by using the neighbor joining method. These two regions are divided by the midpoint between sites 160 and 161. Compared sites are 19,068 bp and 2,086 bp for each tree construction. A101, A201, and O09 are enclosed by a black square. The chimpanzee A allele is used as an outgroup, and midpoint rooting was used. Percent of bootstrap values are shown on each branch.

branching patterns, confirming that allele A was derived from a recombination event between B101 and O01 lineages. As for allele O01, it looks like a mosaic of B101 and O02 (fig. 4B). In addition, we detected three gene conversions among O02, O01, and B101 (supplementary fig. S2, Supplementary Material online) using GENECONV (Sawyer 1989).

The figure 6 describes an evolutionary scheme of the six major alleles of the human ABO gene. We assume A as

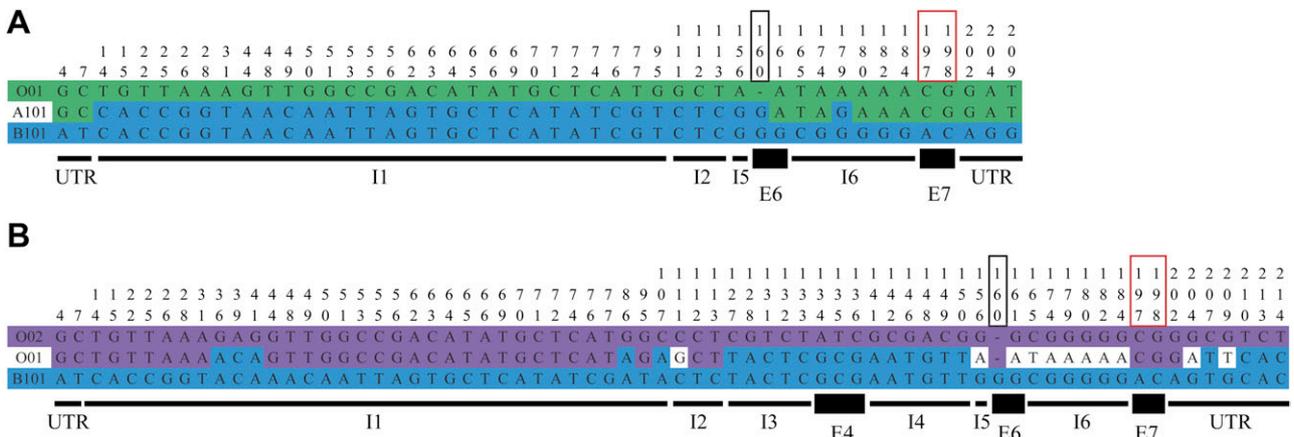


Fig. 4. Parsimony informative sites among some representative haplotypes. (A) Parsimony informative sites among O01, A101, and B101, extracted from figure 1. Site numbers followed figure 2 of Calafell et al. (2008). Untranslated regions (UTRs), some exons (E), and introns (I) are indicated. O01 (green background) and B101 (blue background) are putative parental alleles for the recombinant allele A101. Two sites enclosed by a red square are two critical sites (codons 266 and 268), which determine A/B activity. The site enclosed by a black square is $\Delta 261$. (B) Parsimony informative sites among O02, O01, and B101, extracted from figure 1. O02 and B101 are shown by purple and blue backgrounds, respectively.

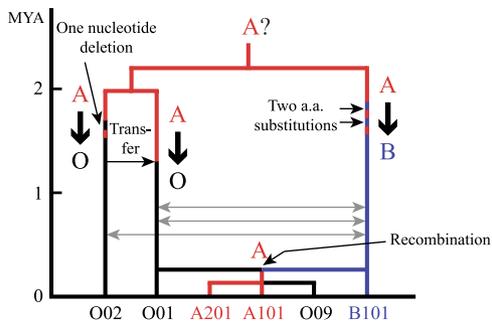


FIG. 6. A scheme of the evolutionary pathway of major alleles of the human ABO genes. A, B, and O lineages are shown by red, blue, and black lines, respectively. Three gray horizontal lines with arrowheads on both ends indicate gene conversions. One nucleotide deletion ($\Delta 261$) for O02, possible transfer of $\Delta 261$ for O01, and two amino acid substitutions for B101 are shown. Because times of each gene conversion event, the $\Delta 261$ deletion, transfer of $\Delta 261$, and two amino acid substitutions for B101 cannot be estimated, these positions are located arbitrarily.

ancestral type in humans because chimpanzees mainly have A alleles (Blancher and Socha 1997; Kermarrec et al. 1999; Sumiyama et al. 2000). Under the assumption of the evolutionary rate ($\lambda = d/2T$) of 1.1×10^{-9} (with “ d ” the genome-wide distance value between humans and chimpanzees being 0.0127, The Chimpanzee Sequencing and Analysis Consortium (2005), and “ T ” the human/chimpanzee divergence time being 6 Ma), the divergence time of the lineage leading to B was estimated to be approximately 2.08 Ma from “ m ” the mean pairwise distance between B101_H19 and O02_H29/O01_H49 ($m = 0.0044$). The divergence time between O01 and O02 lineages was approximately 1.98 Ma from the mean distance between O01_H49 and O02_H29 ($m = 0.0042$). According to the estimated divergence times of haplotype lineages, the lineage leading to B first diverged more than 2 Ma, followed by the divergence between the lineages leading to O01 and O02 alleles.

Probably the point deletion ($\Delta 261$) for the major O alleles appeared only once in early human evolution because it is highly improbable that such a deletion occurred at exactly the same position more than once during this period (Roubinet et al. 2004). Roubinet et al. (2004) also mentioned that point deletions are rarely observed among the human ABO alleles. Human O alleles appeared about 2 Ma in our estimation. Because O alleles are nonfunctional, it may be possible to accumulate additional deteriorating mutations over the 2 myr. But no stop codon or indels other than $\Delta 261$ observed in the O alleles. To examine how likely it is for deteriorating mutations to occur in O alleles, we carried out the following calculations to obtain a rough estimate of numbers of mutations causing additional stop codon(s) and frameshift(s) on O alleles. The ABO functional glycosyltransferases have 354 amino acids, and the proportion of potential codons that can be a stop codon by a single nucleotide substitution (such as UGU, UGC, UGG, UUA, UCA, CGA, GGA, and AGA for a stop codon UGA) is about 0.3 [= $18/(64-3)$]. Because O

alleles are nonfunctional, let us use the rate of synonymous substitutions. Kitano et al. (2004) compared 103 protein-coding genes for hominoids. The average number of synonymous substitutions per synonymous site between human and chimpanzee was 0.0157 from their data. If we assume the average coalescence time of gene between human and chimpanzee to be 6 Ma, the average rate of synonymous substitutions in the human and chimpanzee lineage becomes 1.3×10^{-9} . If we use this estimate, the expected number of the newly created stop codons in the locus is estimated to be about 0.27 (= $0.3 \times 354 \times 1.3 \times 10^{-9} \times 2 \times 10^6$). On the other hand, the expected number of the newly created indels (insertion and deletion) in the locus becomes about 0.42 (= $354 \times 3 \times 2.0 \times 10^{-10} \times 2 \times 10^6$) since the indel rate for nuclear noncoding nucleotide sequences in the primate lineage was estimated by up to 2.0×10^{-10} (Saitou and Ueda 1994). Therefore, it is not surprising that no additional stop codon and/or frameshift appeared during the only 2 myr history of the nonfunctional O allele.

There are some distinct O lineages, such as O01, O02, and O09, and these appearances might be independent events by acquiring $\Delta 261$ by interallelic exchanges (Calafell et al. 2008). This is exactly the case for the allele O09 because it is almost similar with A101 except for $\Delta 261$. Roubinet et al. (2004) also suggested that the two major O alleles, O01 and O02, have persisted by genetic drift in the human population because in all human populations studied so far, O01 and O02 are present, though in variable proportions, and they are nonfunctional and therefore have identical selective values.

Although modern humans have three allele types, several primates have fewer ones (Blancher and Socha 1997). For example, chimpanzees have only A and O alleles and all gorillas’ alleles are B. Therefore, it is possible that human ancestors only had B and O alleles during a certain period.

The A type allele was resurrected by the recombination, that is, the intact exon 6 from B101 and A type two critical sites in exon 7 from O01 had been jointed to form the recombinant resurrected functional A allele as the A101. To estimate the recombination time between B101 and O01 lineages, we used the mean pairwise distance in the upstream region (1–19,176) between B101_H19 and A101_H11/A201_H41/O09_H07 ($m = 0.0005$) and the mean distance in the downstream region (19,177–22,328) between O01_H49 and A101_H11/A201_H41/O09_H07 ($m = 0.0006$). These regions are same as those used in figure 5. The estimates of the recombination time between B101 and O01 from the two regions were very close (ca. 0.24 and 0.28 Ma). Thus, we conclude that the recombination time between B101 and O01, which resulted in the resurrected A lineage, was around 260,000 years ago.

It may be speculated that *Homo erectus* had only B and O alleles until its later stage. O01 type alleles were found in two Neandertal individuals (Lalueza-Fox et al. 2008). This suggests that the frequency of O allele was quite high among Neandertals. The resurrection of the A allele was before the emergence of anatomically modern humans,

and it possibly occurred in Africa. Nowadays, the A101 is distributed throughout the world as one of the major types of the human ABO gene. It is reasonable to assume that the rapid increase of A101 frequency could have been favored by some kind of positive selection. Meanwhile, even nowadays recombinations are at work to produce some rare variant alleles (e.g., Suzuki et al. 1997; Olsson and Chester 1998; Ogasawara et al. 2001; Yip 2002).

Various mechanisms were invoked to explain natural selection that has possibly been operating on the ABO polymorphism; 1) the ABO antigens expressed by epithelial cells are targeted by various infectious agents, which use them to stick the cell surface, 2) in individuals of the secretor type (they express ABO antigens in secretions), mucins bear the ABO substances and inhibit the binding of aggressive infectious agents to epithelial cells, and 3) the ABO polymorphism influences the fate of endogenous glycoproteins, such as von Willebrand factor and intestinal alkaline phosphatase (see Yamamoto et al. 2012). Another mechanism involves the anti-A and anti-B natural antibodies, which belongs to a first line of the immune defense system against infectious agents expressing the A or B substances (see Ochsenbein and Zinkernagel 2000; Seymour et al. 2004). If so, the homozygous O individuals, who express neither A nor B substances and who produce anti-A and anti-B antibodies, could appear as highly favored. However, they intensely express the H antigen, the precursor of A and B substances, which is targeted by various infectious agents. In this context, the mechanisms that have favored the maintenance of the ABO polymorphism could have varied a long time to face the evolution of the infectious agents, leading to an arms race or a red queen type of evolution, or a frequency-dependent selection of alleles.

Supplementary Material

Supplementary figures S1–S2 and table S1 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgments

We thank the Associate editor Naoko Takezaki and three anonymous reviewers for their valuable comments. This study was supported by the National Institute of Genetics Cooperative Research Program (2009-A68, 2010-A23, and 2011-A14).

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.
- Bandelt HJ. 1994. Phylogenetic networks. *Verh Natwiss Ver Hambg.* 34:51–71.
- Blancher A, Socha WW. 1997. The ABO, Hh and Lewis blood group in humans and nonhuman primates. In: Blancher A, Klein J, Socha WW, editors. *Molecular biology and evolution of blood group and MHC antigens in primates*. New York: Springer-Verlag. p. 30–92.
- Bryant D, Moulton V. 2004. Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol.* 21:255–265.
- Calafell F, Roubinet F, Ramírez-Soriano A, Saitou N, Bertranpetit J, Blancher A. 2008. Evolutionary dynamics of the human ABO gene. *Hum Genet.* 124:123–135.
- The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87.
- Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol.* 23:254–267.
- 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.
- Kitano T, Liu Y-H, Ueda S, Saitou N. 2004. Human-specific amino acid changes found in 103 protein-coding genes. *Mol Biol Evol.* 21:936–944.
- 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.* 51:465–471.
- Lalueza-Fox C, Gigli E, de la Rasilla M, Fortea J, Rosas A, Bertranpetit J, Krause J. 2008. Genetic characterization of the ABO blood group in Neandertals. *BMC Evol Biol.* 8:342.
- Ochsenbein AF, Zinkernagel RM. 2000. Natural antibodies and complement link innate and acquired immunity. *Immunol Today.* 21:624–630.
- 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. 1998. Heterogeneity of the blood group Ax allele: genetic recombination of common alleles can result in the Ax phenotype. *Transfus Med.* 8:231–238.
- Roubinet F, Despiaud S, Calafell F, Jin F, Bertranpetit J, Saitou N, Blancher A. 2004. Evolution of the O alleles of the human ABO blood group gene. *Transfusion* 44:707–715.
- Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 4:406–425.
- Saitou N, Ueda S. 1994. Evolutionary rates of insertion and deletion in noncoding nucleotide sequences of primates. *Mol Biol Evol.* 11:504–512.
- Saitou N, Yamamoto F. 1997. Evolution of primate ABO blood group genes and their homologous genes. *Mol Biol Evol.* 14:399–411.
- Sawyer SA. 1989. Statistical tests for detecting gene conversion. *Mol Biol Evol.* 6:526–538.
- Seltsam A, Hallensleben M, Kollmann A, Blasczyk R. 2003. The nature of diversity and diversification at the ABO locus. *Blood* 102:3035–3042.
- Seymour RM, Allan MJ, Pomiankowski A, Gustafsson K. 2004. Evolution of the human ABO polymorphism by two complementary selective pressures. *Proc R Soc Lond B.* 271:1065–1072.
- Sumiyama K, Kitano T, Noda R, Ferrell RE, Saitou N. 2000. Gene diversity of chimpanzee ABO blood group genes elucidated from exon 7 sequences. *Gene* 259:75–79.
- Suzuki K, Iwata M, Tsuji H, Takagi T, Tamura A, Ishimoto G, Ito S, Matsui K, Miyazaki T. 1997. A de novo recombination in the ABO blood group gene and evidence for the occurrence of recombination products. *Hum Genet.* 99:454–461.
- Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol.* 24:1596–1599.
- Yamamoto F. 2000. Molecular genetics of ABO. *Vox Sang.* 78:91–103.

- Yamamoto F, Cid E, Yamamoto M, Blancher A. Forthcoming 2012. ABO research in the modern era of genomics. *Transfus Med Rev.*
- Yamamoto F, Clausen H, White T, Hakomori S. 1990. Molecular genetic basis of the histo-blood group ABO system. *Nature* 345:229–233.
- 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.* 265:19257–19262.
- Yamamoto F, McNeill PD, Hakomori S. 1992. Human histo-blood group A2 transferase coded by A2 allele, one of the A subtypes, is characterized by a single base deletion in the coding sequence, which results in an additional domain at the carboxyl terminal. *Biochem Biophys Res Commun.* 187:366–374.
- Yip SP. 2002. Sequence variation at the human ABO locus. *Ann Hum Genet.* 66:1–27.