SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
EVOLUTIONARY BIOLOGY
GLYCOBIOLOGY
IMMUNOGENETICS

Received
17 April 2014

Accepted
19 September 2014

Published
13 October 2014

Correspondence and requests for materials should be addressed to F.Y. (fyamamoto@imppc.org)

An integrative evolution theory of histo-blood group *ABO* and related genes

Fumiichiro Yamamoto[1], Emili Cid[1], Miyako Yamamoto[1], Naruya Saitou[2], Jaume Bertranpetit[3] & Antoine Blancher[4]

[1]ABO Histo-blood Groups and Cancer Laboratory, Cancer Genetics and Epigenetics Program, Institut de Medicina Predictiva i Personalitzada del Càncer (IMPPC), Campus Can Ruti, Badalona, Catalonia, Spain, [2]Division of Population Genetics, National Institute of Genetics, Mishima, Japan, [3]IBE - Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona, Catalonia, Spain, [4]Laboratoire d'Immunogénétique Moléculaire (LIMT, EA3034), Faculté de Médecine Purpan, Université Paul Sabatier (Université de Toulouse III), Toulouse, France.

The ABO system is one of the most important blood group systems in transfusion/transplantation medicine. However, the evolutionary significance of the *ABO* gene and its polymorphism remained unknown. We took an integrative approach to gain insights into the significance of the evolutionary process of *ABO* genes, including those related not only phylogenetically but also functionally. We experimentally created a code table correlating amino acid sequence motifs of the *ABO* gene-encoded glycosyltransferases with GalNAc (A)/galactose (B) specificity, and assigned A/B specificity to individual *ABO* genes from various species thus going beyond the simple sequence comparison. Together with genome information and phylogenetic analyses, this assignment revealed early appearance of *A* and *B* gene sequences in evolution and potentially non-allelic presence of both gene sequences in some animal species. We argue: Evolution may have suppressed the establishment of two independent, functional *A* and *B* genes in most vertebrates and promoted A/B conversion through amino acid substitutions and/or recombination; *A/B* allelism should have existed in common ancestors of primates; and bacterial *ABO* genes evolved through horizontal and vertical gene transmission into 2 separate groups encoding glycosyltransferases with distinct sugar specificities.

The human histo-blood group ABO system is crucial in safe blood transfusion and cell/tissue/organ transplantation[1,2]. This system consists in A and B oligosaccharide antigens expressed on red blood cells (RBCs) as glycoproteins and glycolipids and antibodies against those antigens in serum. A and B antigens are also expressed by epithelial and endothelial cells, and in secretor type individuals they are also expressed on mucins secreted by exocrine glands. The immuno-dominant structures of A and B antigens are GalNAcα1->3(Fucα1->2)Gal- and Galα1->3(Fucα1->2)Gal-, respectively. *A* and *B* alleles of the *ABO* genetic locus encode A and B transferases, which respectively transfer an N-acetyl-D-galactosamine (GalNAc) or a D-galactose (Gal) to H substances with an α1,3-glycosidic linkage. H substances with the Fucα1->2Gal- structure are synthesized by fucosylation catalyzed by α1,2-fucosyltransferases (α1,2-FTs) encoded by *FUT1/FUT2/SEC1* genes. *FUT1*-encoded α1,2-FTs and *FUT2/SEC1*-encoded α1,2-FTs exhibit distinct acceptor substrate specificity, and are differentially expressed amongst tissues. In humans *SEC1* is a pseudogene and *FUT2* gene presents frequent null alleles so that about 20% of individuals are incapable of expressing either H, A, or B antigens in secretions (non-secretor type). In the absence of α1,2-FTs no H antigens are produced. Therefore, A/B transferases function only when at least one active α1,2-FT is simultaneously present.

In 1990 we correlated the nucleotide sequences of *A*, *B*, and *O* allelic cDNAs and the expression of A and B antigens, and elucidated the molecular genetic basis of human histo-blood group ABO system[3,4]. Four amino acid substitutions (Arg176Gly, Gly235Ser, Leu266Met, and Gly268Ala) discriminate A and B transferases. A single nucleotide deletion (261delG) was found in O alleles. We later identified mutations in A/B subgroup alleles (A^2, A^3, A^x, and B^3) and mutations in *cis-AB* and $B^{(A)}$ alleles specifying dual expression of A and B antigens[5–7]. Another type of *O* allele, which lacks 261delG but contains a Gly268Arg substitution, was found afterward[8]. *ABO* alleles registered in the Blood Group Antigen Gene Mutation Database exceed 250, and *ABO* has become one of the most studied human genetic loci for its polymorphism[9].

ABO genes exist not only in humans but also in many other vertebrate species although ABH antigen expression patterns may be different. In addition to A and B transferases, there are additional enzymes transferring a GalNAc/galactose by α1,3-glycosidic linkage: α1,3-galactosyltransferase and isogloboside 3 synthase (both of galactose specificity), and Forssman glycolipid synthase (GalNAc specificity). These enzymes catalyze the last synthetic steps of α1,3-galactosyl epitope (Galα1->3Galß1->4GlcNAcß-), isogloboside 3 (Galα1->3Galß1->4Glcß1-Ceramide), and Forssman glycolipid antigen (GalNAcα1->3GalNAcß1->3Galα1->4Galß1->4Glcß1-Ceramide), respectively. It should be noted that these enzymes utilize other acceptor substrates than H substances as the chemical structures of their reaction products indicate. Genes encoding these α1,3-Gal(NAc) transferases (α1,3-Gal(NAc)Ts) (*GGTA1*, *A3GALT2*, and *GBGT1* genes, respectively) are paralogous to the *ABO* gene, and they are evolutionarily related[10–13]. Although transferase activity remains to be demonstrated for its encoded protein, another paralogous genetic locus, *GLT6D1* (glycosyltransferase 6 domain containing 1), was associated to periodontitis susceptibility[14]. Based on the nucleotide and deduced amino acid sequences of *ABO* and related genes, a birth-and-death evolution model was proposed[15,16]. Several theories have been proposed on the evolution of the primate ABO polymorphism[17–22]. And the dynamics of the human *ABO* gene evolution have been extensively studied[23,24]. A brief summary of prior knowledge about ABO evolution will be presented in each individual sub-section in the Results section. Indisputably, sequences, single nucleotide polymorphisms (SNPs), and mutations are critical to investigate gene evolution. However, the analyses based solely on sequences are insufficient especially because of genetic recombination. To interpret gene evolution properly knowledge of the gene-encoded proteins is fundamental. What is the protein function, which portion(s) of the protein are important for that function, where is the protein located, does the protein form multimers, how does the protein interact with other molecules, etc., all provide valuable information. Especially, in order to investigate the *ABO* gene evolution the understanding of the sugar specificity of A and B transferases is essential. As in many other areas of genetic studies, functional assays are of critical importance.

In the present work, we analyzed many homologous genes and sequences that had been identified in various species through genome sequencing efforts. In addition to the sequences, we also utilized additional data and information available: gene structure to determine whether a gene is partial or complete; chromosomal organization to deduce duplication(s), deletion(s), inversion(s), and translocation(s) that have occurred; and information on A/B transferases and A/B oligosaccharides to obtain clues on functionality. Data were interpreted with caution because of the incompleteness of genome sequence databases, wrong annotations, and differences among individuals within a species, and errors in genome assemblies. Based on mostly relevant, but not entirely accurate, data, we have delineated a potential scenario of the *ABO* gene evolution. Taking advantage of our expertise, we also prepared several dozens of amino acid substitution constructs of the human A transferase in an expression vector by *in vitro* mutagenesis, determined their GalNAc/galactose specificity, and generated a code table correlating amino acid sequence motif with A/B specificity. Utilizing this table, we decoded the A/B specificity of the *ABO* genes annotated from a variety of species, which in turn has allowed us to uniquely evaluate several critical hypotheses on the evolution of the *ABO* and related genes and their functional impact.

Results

Gene duplications and changes in substrate specificity of the encoded glycosyltransferases created ABO family of genes in animals. All the α1,3-Gal(NAc)T genes in genome databases that were analyzed are listed in Fig. 1. Species were aligned based on their evolutionary relationship (human at top and lamprey at bottom)[25]. A phylogenetic tree was constructed for the 104 protein sequences that are likely to encode functional α1,3-Gal(NAc)Ts, and is shown in Fig. 2. *GBGT1*, *A3GALT2*, *GGTA1*, and *GLT6D1* genes formed separate clusters, whereas both *A* and *B* genes were clustered into a single *ABO* gene cluster. Except that many nonfunctional genes are omitted, these results obtained from amino acid sequence analysis coincided well with the nucleotide sequence-based Ensembl gene tree ENSGT00400000022032 and a previous report[15].

The genes neighboring those glycosyltransferase genes are conserved well in many species and the consensus organizations are shown in Table 1. There is a wide variation in the repertoire of those genes among different species, and the model of birth-and-death evolution[26] fits well with the α1,3-Gal(NAc)T family of genes as previously reported[15]. For instance, amphibian *Xenopus tropicalis* frog has *ABO* genes but lacks any other α1,3-Gal(NAc)T genes whereas all the bird species examined have *GBGT1* but lack *A3GALT2*, *GGTA1*, and *GLT6D1* genes.

Emergence of α1,2-fucosyltransferase genes preceded A/B transferase gene appearance in amphibians. Phylogenetic analyses and their chromosomal locations were used to separate *FUT1*, *FUT2*, and *SEC1* genes, and they are shown in 3 different columns in Fig. 1. The distributions of these genes suggest that *FUT2* gene was the oldest α1,2-FT gene. *FUT1* gene later appeared from *FUT2* lineage after gene duplication followed by acquisition of novel expressional/enzymatic characteristics. *SEC1* gene emerged much later after duplication of *FUT2* gene and following divergence from it, confirming the evolutionary theory previously proposed of α1,2-FT family of genes[27]. The chromosomal region containing α1,2-FT genes has remained stable in many species, and the consensus is shown in Table 1.

A/B antigen expression was previously reported in frog species[28,29]. As shown in Fig. 1, neither *FUT1/FUT2/SEC1* genes nor *ABO* genes are present in fish genomes. Contrastingly, amphibian *Xenopus tropicalis* frog has 4 *FUT2* gene sequences, several of which seem to encode active α1,2-FTs. This frog species also contains multiple *ABO* gene sequences, including a few with possible functionality. Chinese softshell turtle and many mammalian genomes also possess potentially functional α1,2-FT and A/B transferase genes. Therefore, it is logical to hypothesize that A/B antigen(s) appeared after the separation of fish and amphibian lineages.

A code table was generated to correlate amino acid sequence motif with A/B specificity. Progresses have been made in understanding A/B transferases over the last decade. Among the 4 amino acid substitutions at codons 176, 235, 266, and 268 between the human A and B transferases, the third and fourth substitutions were shown to be crucial for different donor nucleotide-sugar substrate specificity whereas the second is influential and the first is not so important[4]. Our *in vitro* mutagenesis study[30] and the determination of the three-dimensional structures of A/B transferases by others[31] confirmed the critical roles of amino acids at codons 266 and 268.

In this study we prepared a library of 40 amino acid substitution constructs of human A transferase, which contained any one of potential 20 amino acid residues at codon 266 in combination with either glycine of A transferase or alanine of B transferase at codon 268. Furthermore, we also prepared additional constructs at codons 263–268 that contained deduced amino acids present in annotated *ABO* and related α1,3-Gal(NAc)T genes in the genome databases but were not represented in the library. DNA from those constructs was transfected to HeLa cells expressing cell-surface H substances, and the expression of A/B antigens was examined immunologically, using antibodies against blood group A/B antigens, respectively. A code table was generated that correlates amino acid sequence motifs and A/B specificity of the enzymes encoded by the various constructs (Table 2). The activity is shown semi-quantitatively in a 4-fold expo-

Species	Ensembl	α1,2-FT Genes			α1,3-Ga(NAc)T Genes																			
	Database	SEC1	FUT2	FUT1	ABO		ABO		GBGT1		ABO/GBGT1		A3GALT2		GGTA1(-1)		GGTA1(-2)		GLT6D1(-1)		GLT6D1(-2)	
					Pseudo	Ancient					Ancient													
Human	ENSG	232871	176920	174951	270624	IGA	175164	LGG	148288	GGA			184389	HAA	204136	HAA					204007	GNL		
							N/A	MGA																
Chimpanzee	ENSPTRG	N/A	11251	11256			21513	LGG	21510	GGA											29002	GSL		
Gorilla	ENSGGOG	232871	1236	1245			11353	MGA	3448	GGA			3596	HAA							7867	GSL		
Orangutan	ENSPPYG	10210		10213			N/A	LGG	19724	GGA			1574	HAA							19761	GSL		
							N/A	MGA																
Gibbon	ENSNLEG	4592	4601	18767			N/A	LGG													8830	GSL		
							N/A	MGA																
Macaque	ENSMMUG	19738		4020	3828	TGA	15100	MGA	15277	GGK			3268	HAA							21493	GSL		
							N/A	LGG																
Marmoset	ENSCJAG		23164				17944	LGG	N/A	GGA			1190	HAA	20186	HAA								
															7103	HKV								
Tarsier	ENSTSYG			5965																	13615	DGS		
Mouse Lemur	ENSMICG			13254			12209	MGA	10686	GGA					16109	HAA					14590	DGS		
							17926	SGA																
Bushbaby	ENSOGAG	29739	3567	24452	27171	TGA	29319	LGG	29130	GAA			34743	HAA	1891	HAA			29560	HGA	30414	DGS		
					33703	TGA																		
Tree Shrew	ENSTBEG								10196	GGA			11627	HAA	3735	HAA			12485	HST				
Rabbit	ENSOCUG	27963	27542	17168	27379	TEA	8654	LGG					12461	HAP	3469	HAA	23397	HAA	29536	HGA				
					21195	TEA	17014	MGA																
					26524	IGA	17844	MGA																
					27881	IGA	7815	IGA																
Pika	ENSOPRG				3647	TGA	15984	IGA					6196	HAA					1107	---				
Mouse	ENSMUSG	40364	55978	8461			15787	GGA	26829	GGA			28794	HAA	35778	HAA	79421	HAA	26882	HRA	36401	AGA		
Rat	ENSRNOG	21014	21011	20995			39906	AGG					5935	HAA	19179	HAA	39210	HAA	42373	HSA	27900	AGA		
							46958	MGA																
							50908	MGA																
							45801	MGA																
Kangaroo Rat	ENSDORG			1233			6089	MGA	11384	GGA			12439	HAA							7879	GSS		
							14859	MGA																
Guinea Pig	ENSCPOG	20316	7337	6070					4874	GGA			3018	HAA	14621	HAA					2951	GDT		
Squirrel	ENSSTOG	23913	12296	12334					N/A	GGA			24594	HAA	5448	HAA					25373	GGS		
Dolphin	ENSTTRG		8204	14561					14438	GGA			956	HAA										
Cow	ENSBTAG	14514	21557	23374			12525	AGG	30319	GRA			44093	HAA	12090	HAA	39186	HAA	39201	HSA	20249	DGA		
Pig	ENSSSCG	27364	3145	3141			N/A	AGG							5518	HAA					5758	DGA		
		23756																						
Horse	ENSECAG	5681	5862	6058	4966	TGA	14463	MGA	12442	GGA					22868	HAA	22656	HAA	21425	HAA	8592	DSA		
			5993		5049	TGA	20130	AGG																
Cat	ENSFCAG		22344	27868			25873	AGG	24387	GGA			3576	HTA	30534	HAA			22798	QSA	11754	DGS		
Panda	ENSAMEG	14807	19834	19826	19214	IGA	7867	MGP					7288	HAA	845	HAA	865	HAA			3338	DGA		
							1005	MGA																
							1485	---																
Ferret	ENSMPUG	3823	19365	3721			12419	AGG	12426	GGA			15083	HAA	7373	HAA	7461	HAA	7470	---				
							13012	IGA																
							1985	MEA																
Dog	ENSCAFG	23807	32038		7214	IGA	19757	AGG	19864	GGA			10375	HAA	20295	HAA	3697	HAA			29005	DGS		
					24235	AGA	24143	SGG																
Microbat	ENSMLUG	6325	23144	6349			29891	LGG					23219	HAA	12248	HAA	26642	HAA	29606	HSA				
			27047				26173	MGA																
							9519	MGA																
Megabat	ENSPVAG								1029	GGA			15028	HAA							13151	DGA		
Hedgehog	ENSEEUG			11062			6261	TGG	14255	GGA			13948	HAA	13477	HAA					12715	DGA		
Shrew	ENSSARG		2891	3064											5276	HAA					14033	DGA		
Elephant	ENSLAFG	28837	2910	12789	29580	TGF	10293	AGG	5668	GGA			5495	HAA	28653	HAA	27324	HAA	15846	HSA	22081	DSS		
Hyrax	ENSPCAG		1114	1320			12074	AGG	15568	GGA									7432	HSA				
L. Hedgehog Tenrec	ENSETEG		17392				9122	TGS	11084	GGA			5305	HAA										
Armadillo	ENSDNOG			5189					898	GGA			139	HAA					18207	GNG	10340	DGA		
Sloth	ENSCHOG		7931						2651	GGA											2110	DGS		
Opossum	ENSMODG		4120	4363			12702	MGG	12666	GGA					19662	HAA			25786	HAA				
			6773																					
			23793																					
			23794																					
Wallaby	ENSMEUG		1419	15264																				
Tasmanian Devil	ENSSHAG			5251					8459	GGA					18242	HAA			18125	HVA				
Platypus	ENSOANG		10021	30729			335	---	8304	GGA														
							10669	MGA																
							12223	LGA																
							15238	AGG																
Chicken	ENSGALG								3340	GGA														
Duck	ENSAPLG								4843	GGA														
Turkey	ENSMGAG								6307	GGA														
Zebra Finch	ENSTGUG				2987	TAS			5430	GGA														
Flycatcher	ENSFALG				887	TAS			1494	GGA														
Anole Lizard	ENSACAG		16777																					
C. Softshell Turtle	ENSPSIG		5700		4564	AAS	9003	AAA	14027	GGA			13541	HAA										
					5387	AAN	16341	AAA																
							10508	AAA																
							9261	MGA																
Xenopus Frog	ENSXETG		2881		10713	MAA	5013	AGG																
			4564		22822	MAA	30623	AGG																
			30323				31740	AGG																
			33054				25598	AGG																
							5033	TGC																
							31253	TGC																
							5032	TGC																
							5025	AGG																
							5024	AGG																
Coelacanth	ENSLACG		469						483	GGA	10082	TSE	7593	HAA										
Tilapia	ENSONIG								9223	CGG	N/A	TSE	4079	HAA										
													3914	HAA										
													5927	HAA										
													7169	HAA										
													7166	HAA										
Tetraodon	ENSTNIG										10664	TSE	14786	HAA										
													788	---										
Fugu	ENSTRUG										N/A	TSE	14665	HAA										
													16170	HAA										
Stickleback	ENSGACG								14813	CGG	N/A	TSG	7442	HAA										
									537	TAA			17105	HAA										
									16064	TAA														
Platyfish	ENSXMAG								12659	CGG	5560	TSE	17120	HAA										
									14479	TAA														
Medaka	ENSORLG								10140	TAA	8616	SSE	4045	HAA										
													893	HAA										
													895	HAA										
Cod	ENSGMOG								12375	SAA	19862	TSE	1877	HAA										
													1884	HAA										
Zebrafish	ENSDARG								91969	GGA														
									35555	GGA														
									91936	---														
									91944	GGA														
									68503	GGA														
									11283	TAA														
									92718	GGA														
									5257	TAA														
									25275	TAA														
									19207	CGA														
Lamprey	ENSPMAG										5316	TSE	9878	HAA										
													4546	HAA										

Figure 1 | Species-dependent distribution of *FUT1/FUT2/SEC1* α1,2-fucosyltransferase genes and *ABO/GBGT1/A3GALT2/GGTA1/GLT6D1* α1,3-Gal(NAc) transferase genes. This table shows the distribution of α1,2-FT genes and α1,3-Gal(NAc)T genes in a variety of organisms. Ensembl gene identifiers are listed only with the meaningful digits, excluding 0 s on the left from their IDs. Genes were categorized into groups based on Ensembl gene trees, chromosomal locations, and our own analyses, and they are aligned in different columns and shown highlighted in different colors. Amino acid sequences corresponding to the codons 266–268 of human A/B transferases are also shown. The symbol "---" indicates the absence of sequence motif, and "N/A" means not annotated in databases. A single column of "Pseudo/Ancient" was used to list two types of annotated gene sequences: The *ABO* retropseudogene sequences that were originally derived from an intronless cDNA are highlighted in tan color (Pseudo) and the sequences that formed a cluster next to the *ABO* gene in the phylogenetic analysis are highlighted in yellow (Ancient). The gene sequences that formed a cluster outside of the *ABO/GBGT1* genes are highlighted in orange, and they are shown separately in the "ABO/GBGT1 Ancient" column. The annotated genes may or may not be functional, the latter of which may also be called as *O* genes or pseudogenes. Note that genome sequences were not complete for many species, and therefore, errors may exist. In addition, there are numerous homologous sequences that have yet to be annotated and mapped on chromosomes. Furthermore, polymorphism may also exist.

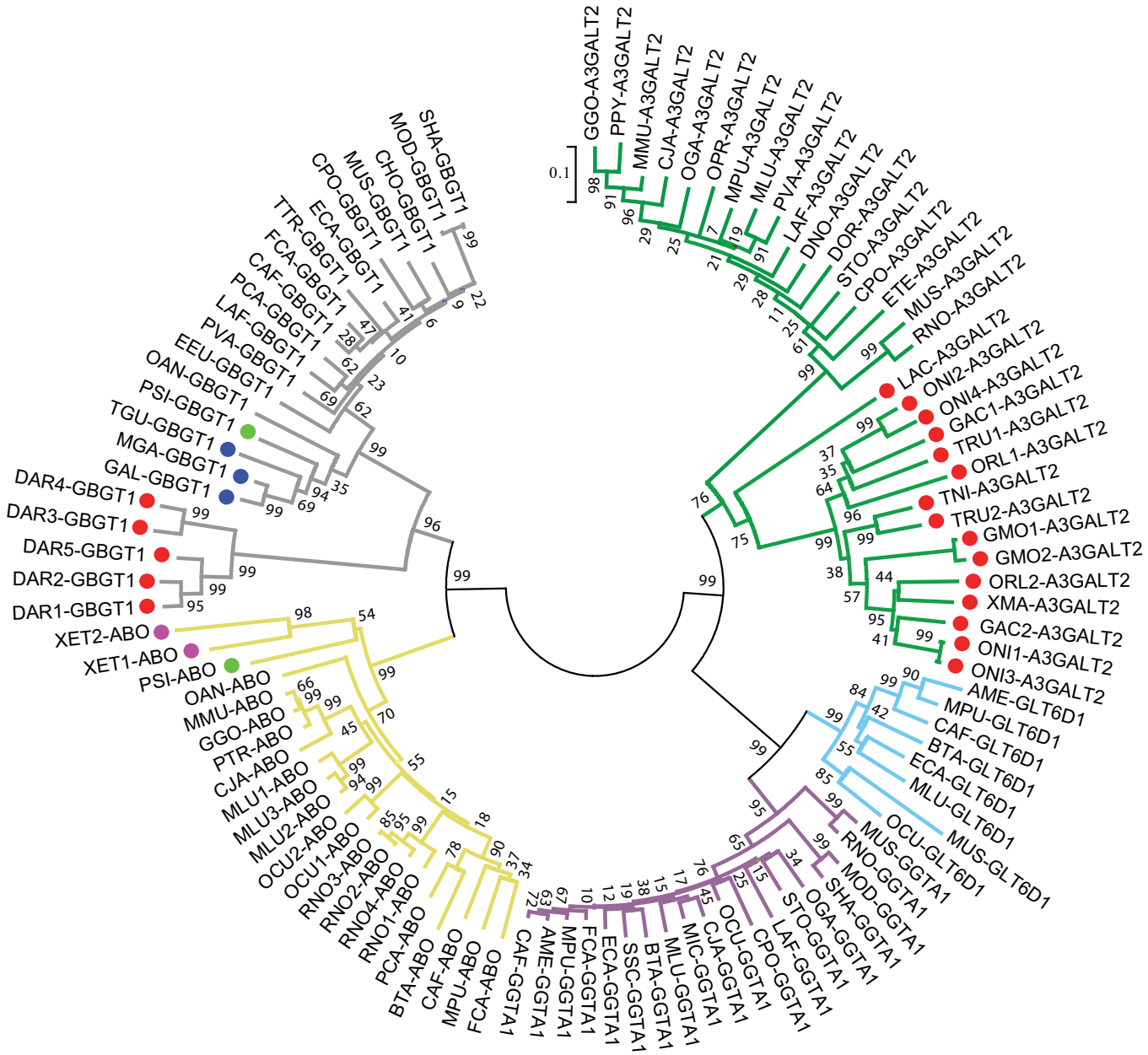

Figure 2 | Evolution of α1,3-Gal(NAc) transferase family of genes. The MEGA5 software was used to analyze 104 amino acid sequences potentially encoding intact ABO proteins. The amino acid sequences corresponding to codons 69–354 of the human A transferase were examined. 1,000 bootstrap replications were computed. Branches leading to *ABO*, *GBGT1*, *A3GALT2*, *GGTA1*, and *GLT6D1* genes are colored in yellow, grey, green, purple, and blue, respectively. The bootstrap frequencies are shown on the branching points. Fishes, amphibians, reptiles, and birds are marked with closed circles in red, purple, green, and dark blue whereas mammals are unmarked. The species code names correspond to the names shown in the "Ensembl Database" column in Fig. 1. For instance, PTR for chimpanzee (*Pan troglodytes*) is obtainable by removing ENS and G from the database name (ENSPTRG).

nential scale with 5+ highest and - none. The motifs observed in *ABO* genes *in natura* are shown in bold type.

The control constructs exhibited the anticipated specificity: AGG motif at codons 266–268 in pig *A* gene, LGG and MGA in human *A* and *B* alleles, and GGA in mouse *cis-AB* gene for A, A, B, and AB specificity, respectively. The results clearly demonstrated that the amino acid residue at codon 266 is crucial to determine the sugar specificity and activity of the encoded transferase. Some constructs possessing glycine at codon 268 exhibited different specificity/activity from those possessing alanine, suggesting that codon 268 is also important. A tendency of preferential use of galactose over GalNAc was observed by the Gly268Ala substitution, possibly because increased size in side chain at that position hinders larger GalNAc access whereas facilitating smaller galactose access. Several constructs with the amino acid sequence motifs that were overlapped with our previous study[30] exhibited the same specificity/activity in spite of the differences in the A/B transferase backbone.

In addition to the constructs expressing either A or B transferase activity, several constructs exhibited both A and B transferase activities whereas several others showed none. For instance, human A transferase constructs containing AAA, CGG, or SGG motif exhibited A specificity, whereas those with IGA, MAA, MGS, or QGC exhibited B specificity. The constructs with MGG, SGA, TGA, or AAS showed both A and B specificity whereas those with AAN, TEA, or TGF showed neither. An unexpected finding was that glycine at codon 267 is not an absolute pre-requisite for A/B transferase activity. We next applied the codes to uniquely assign potential A/B

Table 1 | Consensus organization of genes surrounding α1,3-Gal(NAc) transferase and α1,2-fucosyltransferase genes

α1,3-Gal(NAc) transferase genes

***ABO* and *GBGT1* genes**

REXO4->, <-C9ORF96, SURF4->, <-SURF2, SURF1->, <-RPL7A, MED22->, SURF6->, ***ABO***->, LCN1->, OBP2B->, ***GBGT1***->, RALGDS->, <-CEL, <-GTF3C5, <-GFI1B

***A3GALT2* gene**

Mammals

<-ZSCAN20, PHC2->, ***A3GALT2***->, <-ZNF362, TRIM62->

Fish

<-FAM83E, EMP3->, <-***A3GALT2***, <-ZNF362, <-TRIM62

GUCA1B->, MAPK8IP1->, ***A3GALT2***->, LRP4->, <-NELL1

***GGTA1* and *GLT6D1* genes**

TTLL11->, <-DAB2IP, ***GGTA1*** (-1)->, ***GGTA1*** (-2)->, ***GLT6D1*** (-1)->, STOM->, <-GSN

OBP2A->, PAEP->, <-***GLT6D1*** (-2), LCN9->, <-SOHLH1, KCNT1->

α1,2-fucosyltransferase genes

***FUT1/FUT2/SEC1* genes**

SULT2B->, <-FAM83E, SPACA4->, <-RPL18, SPHK2->, <-DBP, <-CA11, <-NTN5, ***SEC1***->, ***FUT2***->, *<-MAMSTR, <-RASIP1, <-IZUMO1, <-* ***FUT1****, FGF21->, <-BCAT2, <-HSD17B14, <-PLEKHA4, PPP1R15A->, <-TULP2, NUCB1->*

Chromosomal regions containing α1,3-Gal(NAc)T and α1,2-FT genes have remained stable in many species with the consensus organization shown. The arrows indicate the direction of transcription.

specificity of the annotated *ABO* genes and critically evaluated several hypotheses on the evolution of the *ABO* genes.

***A* and *B* gene sequences appeared early in evolution and are potentially present in a non-allelic manner in some species.** The first evidence of genomes with multiple copies of *ABO* gene sequences came from the Southern hybridization experiments showing multiple bands of hybridization in dog, rabbit, and rat genomic DNA using a human probe[32]. Later studies demonstrated multiple genes in rat[33]. As shown in Fig. 1, additional species also seem to possess multiple *ABO* gene sequences. They are *Xenopus tropicalis* frog, Chinese softshell turtle, platypus, microbat, dog, ferret, panda, horse, Kangaroo rat, rat, and rabbit species. Genes flanking full/partial *ABO* genes are shown for each individual species in Table 3, together with the amino acid sequences corresponding to codons 266–268 of the human A/B transferases.

We applied Table 2 to decode A/B specificity of individual *ABO* gene sequences annotated in various vertebrate species. It was found that several species not only contain multiple copies of *ABO* gene sequences but also they may have both *A*-specific and *B*-specific gene

Table 2 | Specificity and activity of human A transferase expression constructs containing various amino acids at codons 263–268

(I). G at codon 268				(II). A at codon 268				(III). Additional			
Codons	A	B	A/B	Codons	A	B	A/B	Codons	A	B	A/B
(266–268)	Activity	Activity	Specificity	(266–268)	Activity	Activity	Specificity	(266–268)	Activity	Activity	Specificity
AGG	+++++	−	A	**AGA**	+++++	++	AB	**AAA**	+++++	−	A
CGG	+++++	−	A	CGA	++++	+++	AB	**AAN**	−	−	−
DGG	++	++	AB	DGA	−	+++	B	**AAS**	++++	+++	AB
EGG	++++	−	A	EGA	−	++++	B	**MAA**	−	+++++	B
FGG	−	++++	B	FGA	−	++++	B	**MGP**	−	+++	B
GGG	++++	−	A	**GGA**	++++	+++	AB	**MGS**	−	+++++	B
HGG	−	++++	B	HGA	−	++++	B	**QGC**	−	+++++	B
IGG	++++	++++	AB	**IGA**	−	++++	B	SSE	−	−	−
KGG	−	−	−	KGA	−	+++	B	TAS	−	−	−
LGG	+++++	−	A	**LGA**	+++++	+	AB	**TEA**	−	−	−
MGG	++++	++++	AB	**MGA**	−	++++	B	**TGC**	++++	−	A
NGG	+++++	+	AB	NGA	++++	++	AB	**TGF**	−	−	−
PGG	+++++	−	A	PGA	++++	−	A	TSE	−	−	−
QGG	++++	+++	AB	QGA	−	+++++	B				
RGG	−	−	−	RGA	−	−	−	(263–268)			
SGG	++++	−	A	**SGA**	++++	+++	AB	**AYVYGS**	−	−	−
TGG	+++++	−	A	**TGA**	++++	+++	AB	FYFTSE	−	−	−
VGG	++++	−	A	VGA	++++	+++	AB	HYYMGG	++++	++++	AB
WGG	++	+	AB	WGA	−	++++	B	YYYAGG	+++++	−	A
YGG	−	++++	B	YGA	−	++	B	YYYMGG	+++++	+++	AB
								YYYTGS	+++++	−	A
								YYYTSE	−	−	−
								YYYTSG	+++++	−	A

The left 2 sets show the results of a library of human A transferase expression constructs containing any of 20 potential amino acid residues at codon 266 with glycine of A transferase or alanine of B transferase at codon 268. The right set shows the results of additional constructs that were not included in the library. The results of immunostaining with anti-A or anti-B antibodies were adjusted by transfection efficiency using co-transfected GFP-positive cell percentages. The activity is shown in a semi-quantitative manner on a 4-fold exponential scale with 5+ highest and − none. The letter size in A/B Specificity reflects the activity strength whereas "−" indicates no activity. The constructs shown in bold type are mentioned in the text.

sequences in their genomes. For instance, *Xenopus tropicalis* frog has *A* gene sequences with AGG or TGC motif and *B* gene sequences with MAA motif. Other species identified are: Chinese softshell turtle (AAA for *A* and MGA for *B*), platypus (AGG for *A*, MGA for *B*, and LGA for *AB*), horse and rat (AGG for *A* and MGA for *B*), microbat (LGG for *A* and MGA for *B*), and rabbit (LGG for *A* and MGA and IGA for *B*). These results suggest that functional differentiation between *A* and *B* gene sequences appeared early in evolution, possibly just after the *ABO* gene emergence in amphibians.

As shown in Table 3, horse *A* and *B* gene sequences are closely located in tandem on the same chromosome. Therefore, if horse genome assembly is correct, those sequences may not be unigenic alleles. Microbat *A* and *B* gene sequences have not yet been mapped on chromosomes, however, at least one *A* and one *B* gene sequences of the three present in the genome were aligned side-by-side within a single contig (ENSMLUG00000029891 with LGG and ENSMLUG00000026173 with MGA in Scaffold GL431842: 18,186-26,341). Accordingly, they are not allelic, either. The rat genome in the Ensembl database lists 4 *ABO* gene sequences: 1 *A* (AGG) and 3 *Bs* (MGA). The surrounding chromosomal organization in Table 3 shows that those sequences are not alleles. Rat *A* and *B* gene sequences located tandemly in a *cis*-manner contrast to mouse gene (GGA) encoding a transferase with dual specificity (*cis*-AB enzyme)[34].

However, heterogeneity seems to exist among different strains of rats. The Ensembl genome is from the BN/SsNHsdMCW strain. In addition to this strain, GenBank database also houses the genome sequence from another strain, the BN/Sprague-Dawley strain (Rn_Celera). 1 *A* (AGG) and 2 *B* (MGA) gene sequences, rather than 1 *A* and 3 *B*, were mapped for this strain. In another strain, Wistar, 3 *A* and 1 *B* gene sequences were cloned although they have not been mapped[33]. Different cloning results were obtained from inbred GOTW strain[35] and the BDIX strain[36], further complicating the understanding of rat *ABO* genes.

In spite of potential errors and problems that are frequently associated with the sequences and genome assemblies of polymorphic genes and multi-gene families, the presence of multiple copies of non-allelic *A* and *B* gene sequences in rat and other species cannot be all attributed to bioinformatics mistakes. Even if sequence alignment all failed from the same caveats, the case still stands with rats at least. Because three different *A* and one *B* gene sequences were cloned from a single Wistar rat, they cannot be allelic at a single genetic locus[33,37].

Many of non-allelic ABO protein sequences were clustered within species in phylogenetic analyses. Phylogenetic trees of ABO proteins/peptides were constructed from species having more than 1 annotated *ABO* gene (Fig. 3a). For comparison, the human A and B transferase sequences were included in the analysis although human sequences are allelic. Proteins corresponding to full genes with initiation and termination codons are marked with circles, whereas peptides corresponding to partial genes are marked with triangles. The symbols' colors indicate deduced potential A/B specificity (GalNAc, galactose, both, none, and uncharacterized specificity are shown in red, green, yellow, blue, and black, respectively). The amino acids corresponding to codons 266–268 of the human A transferase are shown in parentheses.

The majority of ABO protein sequences were clustered in species-specific groups, including platypus, microbat, rabbit, and rat. However, several protein sequences from two distant species are on a common phylogenetic branch. Among them, two frog (both with MAA motif) and two turtle (with AAN and AAS motifs) sequences clustered together. However, those sequences were deduced to be nonfunctional, having aberrant gene organizations such as the absence of *N*-terminal exons or missing initiation/termination codons. Two ferret (IGA or MEA) and three panda (MGP, MGA, and ---) protein sequences corresponding partial genes with aberrations in codon reading frame and gene structure, clustered on a common branch, apart from the ferret protein from a full gene with AGG motif. In horse species two genes (MGA and AGG) that are located side-by-side on the same chromosome were separated in the phylogenetic tree, possibly due to frameshift mutations deleting a serine close to MGA motif (MGAFFGGSV) and the accelerated accumulation of mutations after inactivation.

An intronless *ABO* gene cDNA was integrated into the mammalian genome. In addition to full/partial genes, *ABO* retropseudogenes also exist, originally derived from an intronless *ABO* gene cDNA that was integrated into the genome during the mammalian evolution (Fig. 1). Those retropseudogenes clustered separately from full/partial *ABO* genes in phylogenetic analyses, and a phylogenetic tree of *ABO* retropseudogene products is shown in Fig. 3b. This tree suggests that the original sequence may have contained a TGA motif, which is present in some bacterial *ABO* genes (see below), but is missing in animal *ABO* genes that were analyzed other than the retropseudogenes. The implication and potential significance are unknown.

Several different molecular mechanisms may be responsible for animal AO polymorphism. Generation of enzymes with novel specificity and/or creation of genes with differential expression patterns must suffice special conditions and requirements. On the contrary, inactivation of gene function or annulment of transferase activity may be relatively easily achieved. Diverse inactivation mechanisms, including frameshift and missense mutations, have been identified in human *O* alleles[4,8,16,23,38,39]. Additionally, species-specific *O* alleles, which possibly resulted from independent silencing mutations, are known to exist in non-human primates[40–42]. In non-primate animal species unigenic AO polymorphism has been reported of pig, dog, rat, cow, and rabbit[43]. The molecular mechanism of the porcine AO polymorphism was previously elucidated[44,45]. A major portion of the structural gene, including the entire coding sequence in the last coding exon, was found missing in *O* alleles from various pig strains.

Assignment of A/B specificity to individual *ABO* gene sequences has allowed us to investigate the molecular mechanisms that established AO polymorphism in other species. Two genes are annotated in dog species (with AGG or SGG). The AGG sequence is located in the consensus chromosomal region, but the SGG sequence is located on a different chromosome and seems to be nonfunctional as judged by abnormal gene structure with the last coding exon indel-disrupted. Therefore, AO polymorphism is suspected at the AGG gene locus. The examination of the coding sequence identified two interesting SNPs: rs9240920 [897G->A] and rs9240927 [701delG]. The former is a nonsense mutation (Trp299Ter) and the latter is a frameshift mutation. Therefore, the genes with either of these SNPs may account for some of the *O* alleles in the dog AO polymorphism.

An interesting finding was made when the chromosomal organization surrounding the *ABO* genes was compared between rat and mouse species. The mouse genome is of very high quality, and many duplicated regions have been properly solved. Therefore, it provides a useful control. The gene organizations are similar except that a DNA fragment containing 3 *ABO* (1 *A* and 2 *B*) and several additional genes is present in rat between *ABO* and *FAM69B* genes (Table 3). The genes present specifically in this chromosomal region in the rat genome are shown in bold type. If the insertion occurred at the population level, the genome without the insert may be regarded as *O* allele. Alternatively, *O* alleles may have arisen from the genome with *A* gene by deletion/unequal crossover. The cow and rabbit genomes list one (*A* gene sequence with AGG motif) and four (1 *A* gene sequence with LGG motif, 1 *B* gene sequence with IGA, and 2 *B* gene sequences with MGA, in addition to 4 retropseudogene sequences), respectively. The information on the *ABO* genes in those

Table 3 | Genes adjacent to *ABO* genes

Species	Gene order*
Primates	
Human (*Homo sapiens*)	1->, <-2, <-3, 4->, <-6, 7->, 8->, *ABO* (LGG)->, 9->, 10->, *GBGT1* (GGA)->, 11->, <-12, <-13, <-14, <-15
Chimpanzee (*Pan troglodytes*)	1->, <-2, <-3, <-16, 4->, 7->, 8->, *ABO* (LGG)->, 10->, *GBGT1* (GGA)->, 11->, <-13, <-14, <-15
Gorilla (*Gorilla gorilla gorilla*)	1->, <-2, <-3, 4->, <-6, 7->, 8->, *ABO* (MGA)->, 9->, 10->, *GBGT1* (GGA)-> 11->, 17->, <-13, <-14, <-15
Orangutan (*Pongo abelii*)	1->, <-2, <-3, 5->, 4->, <-6, 7->, 8->, <-18, 9->, 10->, *GBGT1* (GGA)->, 11->, <-19, 20->, <-13, <-14, <-15
Rhesus macaque (*Macaca mulatta*)	1->, <-2, 5->, <-3, 4->, <-6, 7->, 8->, 21->, *ABO* (MGA)->, 9->, 10->, *GBGT1* (GGK)->, 11->, <-13, <-14, <-15
Marmoset (*Callithrix jacchus*)	1->, <-2, 5->, <-3, 4->, <-*ABO* (LGG), <-8, <-7, 9->, 10->, *GBGT1* (GGA)->, 11->, <-13, <-14, <-15
Bushbaby (*Otolemur garnettii*)	1->, <-2, <-3, 4->, <-6, 7->, 8->, *ABO* (LGG)->, <-22, 23->, <-24, 25-> // 26->, *GBGT1* (GAA)->
Other Mammals	
Mouse (*Mus musculus*)	1->, <-27, 5->, <-3, 4->, <-6, 7->, 8->, *ABO* (GGA)->, 28->, <-22, 23->, <-24, 25-> // 29->, <-30, 31->, *GBGT1* (GGA)->, 11->, <-13, <-14, <-15
Rat (*Rattus norvegicus*)	1->, <-32, 5->, <-3, 4->, <-33, 7->, 8->, *ABO* (MGA)->, 10->, <-28, <-***ABO* (AGG), 34-**>, **4-**>, <-**3,** <-**35, 7-**>, ***ABO* (MGA)-**>, ***ABO* (MGA)-**>, <-**10, 36-**>, <-**37,** <-**38,** <-**39, 40-**>, <-22, 23->, <-24, 25->
Rabbit (*Oryctolagus cuniculus*)	*ABO* (LGG)-> // *ABO* (MGA)-> // *ABO* (MGA)-> // 42->, 43->, *ABO* (IGA)->, 44->, 45->
Dog (*Canis lupus familiaris*)	5->, <-3, <-6, 7->, 8->, *ABO* (AGG)->, 9->, 46->, 46->, <-*GLT6D1* (DGS), 47->, 48->, 49->, <-50, 51->, <-52 // 29->, <-53, <-54, 31->, <-55, *GBGT1* (GGA)->, 11->, <-13, <-14, <-15 // 56->, 57->, 58->, 59->, <-60, <-61, 62->, 63->, *ABO* (SGG)->, 64->, <-65
Ferret (*Mustela putorius furo*)	1->, <-2, 5->, 4->, 7->, 8->, <-*ABO* (AGG), *GBGT1* (GGA)->, 11->, <-13, <-14, <-15 // <-66, 67->, <-68, 69->, 70->, <-71, *ABO* (IGA)->, <-72, 73-> // <-74, 75->, 76->, <-77, 78->, <-79, <-79, *ABO* (MEA)->, <-80
Horse (*Equus caballus*)	1->, <-2, <-3, 4->, <-6, 7->, 8->, *ABO* (AGG)->, *ABO* (MGA)->, *GBGT1* (GGA)->, 11->, <-13, <-14, <-15
Cow (*Bos taurus*)	1->, <-2, <-3, 4->, <-6, 7->, 8->, *ABO* (AGG)->, <-22, 23->, <-24, 25-> // <-51, 50->, <-47, *GLT6D1* (DGA)->, <-81, <-46, <-10, *GBGT1* (GRA)->, 11->, <-13, <-14, <-15
Microbat (*Myotis lucifugus*)	*ABO* (LGG)->, <-*ABO* (MGA) // <-*ABO* (MGA)
Elephant (*Loxodonta africana*)	1->, <-2, 4->, <-6, 7->, 8->, *ABO* (AGG)->, 82->, <-83 // <-84, <-13, <-85, *GBGT1* (GGA)->, 11->, <-14, <-15
Opossum (*Monodelphis domestica*)	<-3, <-6, 4->, 7->, 8->, <-*ABO* (MGG), 86->, *GBGT1* (GGA)->, 11->, <-13, <-14, <-15
Platypus (*Ornithorhynchus anatinus*)	*ABO* (AGG)-> // *ABO* (–)->, *ABO* (MGA)-> // *ABO* (LGA)->, 86-> // *GBGT1* (GGA)->
Birds	
Flycatcher (*Ficedula albicollis*)	1->, <-2, 5->, <-3, 4->, <-6, 7->, 8->, *GBGT1* (GGA)->, 11->, <-14, <-13, <-15 // 87->, 88->, 89->, <-*ABO* (TAS), 90->, 91->
Zebra finch (*Taeniopygia guttata*)	1->, <-2, <-2, 5->, <-3, 4->, <-6, 7->, 8->, 8->, *GBGT1* (GGA) ->, 11->, <-14, <-13, <-15 // <-92, <-93, <-94, 95->, <-96, <-*ABO* (TAS), 97->, 98->
Turkey (*Meleagris gallopavo*)	1->, <-2, 5->, <-3, 4->, <-6, 7->, 8->, 86->, *GBGT1* (GGA)->, 11->, <-14, <-13, <-15
Duck (*Anas platyrhynchos*)	1->, <-2, 5->, <-3, 4->, <-6, 7->, 8->, 86->, *GBGT1* (GGA)->, 11->, <-14, <-13, <-15
Chicken (*Gallus gallus*)	1->, <-2, 5->, <-3, 4->, <-6, 7->, 8->, 86->, *GBGT1* (GGA)->, 11->, <-14, <-13, <-15
Reptiles:	
Softshell turtle (*Pelodiscus sinensis*)	99->, <-100, <-101, 1->, <-2, <-*ABO* (AAA), 5->, <-3, 4->, <-6, 7->, 8-> // *ABO* (AAA)-> // *ABO* (AAA)-> // *ABO* (MGA)-> // 86->, *GBGT1* (GGA)->, 11->, <-13, <-14, <-15 // 102->, 103->, <-*ABO* (AAN), 104->, 104-> // <-*ABO* (AAS)
Amphibians:	
Xenopus frog (*Xenopus tropicalis*)	<-6, 7->, <-105, 8->, *ABO* (AGG)->, *ABO* (AGG)->, *ABO* (AGG)->, *ABO* (AGG)-> // *ABO* (AGG)->, *ABO* (AGG)->, *ABO* (TGC)->, *ABO* (TGC)->, *ABO* (TGC)->, 86-> // <-106, <-107, <-108, 109->, <-110, *ABO* (MAA)->, <-41 // *ABO* (MAA)->

When there is long gap, double slash (//) is given. Three key amino acid sequences are shown in parentheses for *ABO*, *GBGT1*, and *GLT6D1* genes. The genes in the inserted chromosomal region that is specifically present in the rat genome and is absent in the mouse genome are shown in bold type. Other genes are abbreviated as follows.

1: *REXO4*	2: *C9ORF96*	3: *SURF2*	4: *SURF1*	5: *SURF4*	6: *RPL7A*
7: *MED22*	8: *SURF6*	9: *LCN1P1* = *LCN1*	10: *OBP2B*	11: *RALGDS*	12: *CELP*
13: *CEL*	14: *GTF3C5*	15: *GFI1B*	16: *ENSPTRG039599*	17: *ENSGGOG027486*	18: *ENSPPYG019727*
19: *ENSPPYG019722*	20: *ENSPPYG019721*	21: *ENSMMUG032079*	22: *FAM69B*	23: *AGPAT2*	24: *EGFL7*
25: *NOTCH1*	26: *MUS81*	27: *GM711*	28: *LCN4*	29: *PPP1R26*	30: *C9ORF116*
31: *MRPS2*	32: *RGD1307355*	33: *RGD1560194*	34: *GOT2*	35: *RGD1560194*	36: *RPS13*
37: *OBP2A*	38: *RPL9*	39: *VEGP1*	40: *VEGP2*	41: *FAM5B*	42: *TRIB1*
43: *MTPN*	44: *ENSOCUG029177*	45: *ARHGAP20*	46: *PAEP*	47: *LCN9*	48: *ENSCAFG019749*
49: *ENSCAFG019747*	50: *SOHLH1*	51: *KCNT1*	52: *CAMSAP1*	53: *ENSCAFG032138*	54: *ENSCAFG031986*
55: *EEF1A1*	56: *IFIT2*	57: *IFIT3*	58: *IFIT1*	59: *IFIT5*	60: *ZNF248*
61: *ENSCAFG029179*	62: *ZNF487*	63: *ZNF33A*	64: *ZNF37A*	65: *CHRM3*	66: *POLR1C*
67: *YIPF3*	68: *TJAP1*	69: *LRRC73*	70: *DLK2*	71: *ABCC10*	72: *SAP18*
73: *ZNF318*	74: *PDS5B*	75: *N4BP2L2*	76: *N4BP2L1*	77: *BRCA2*	78: *ZAR1L*
79: *FRY*	80: *RXFP2*	81: *LGB*	82: *INSL6*	83: *JAK2*	84: *TJP2*
85: *FXN*	86: *CCDC180*	87: *NDC80*	88: *USP17L23*	89: *OR10AG1*	90: *OR6Y1*
91: *OR9K2*	92: *SOST*	93: *DUSP3*	94: *MPP3*	95: *KCNJ3*	96: *ACR*
97: *SDR39U1*	98: *DAD1*	99: *SLC2A6*	100: *CACFD11*	101: *ADAMTS13*	102: *OR5AP2*
103: *OR14I1*	104: *OR11A1*	105: *A4GNT*	106: *TOR3A*	107: *FAM20B*	108: *RALGPS2*
109: *ANGPTL1*	110: *RASAL2*				

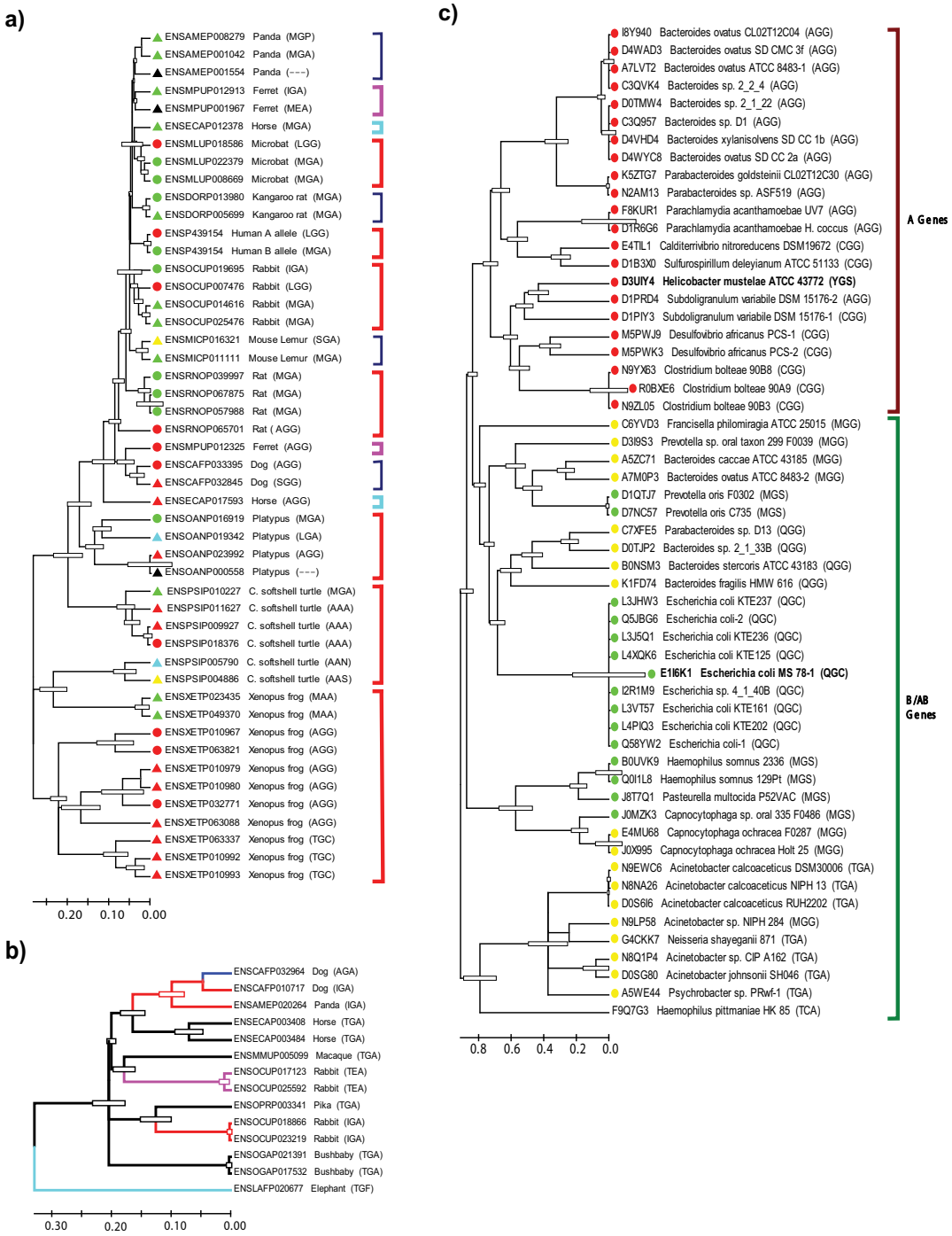

Figure 3 | (*a*): A phylogenetic tree of ABO proteins/peptides from species possessing multiple copies of *ABO* gene. Phylogenetic analyses were performed with protein/peptide sequences from species that contain more than one *ABO* genes in their genomes. Processed intronless retropseudogenes were excluded from analysis. The amino acid sequences were analyzed in its entirety. Potentially functional proteins from full genes with the initiation and termination codons and peptides from partial genes without them are marked with circles and triangles, respectively. The symbol's color indicates potential sugar specificity (GalNAc, galactose, GalNAc/galactose, none, and unknown for red, green, yellow, blue, and black, respectively). Amino acid sequences corresponding to the codons 266–268 of human A/B transferases are also shown in parentheses. Genes in the same species are bracketed. When potential *A* and *B* gene sequences are both present in a single species, the bracket was colored in red. Horse genes and ferret genes in 2 separate clusters are bracketed in blue and purple, respectively. Other species are bracketed in dark blue. (*b*): A phylogenetic tree of originally intronless *ABO* retropseudogene products. The entire protein sequences of processed retropseudogenes were analyzed. Branches leading to different amino acid sequences at the important positions are coded in different colors. (*c*): *ABO* gene evolution in bacteria. EMBL-EBI InterPro database listed 57 bacterial proteins within the GT6 family. 56 proteins/peptides, excluding 1 short one, were aligned to construct a phylogenetic tree. *A* gene from *Helicobacter mustelae* and *B* gene from *Escherichia coli* O86 strain were included in the study, and their results are shown in bold type. The *B* gene-encoded protein (E1I6K1) consists of 234 amino acids, and the bacterial protein sequences corresponding to codons 2–219 of this protein were analyzed. The amino acid sequence motifs corresponding to the codons 266–268 of human A/B transferases are also shown in parentheses. In E1I6K1 these correspond to codons 145–147. The symbols' color indicates sugar specificity of transferases: red, green, and yellow for GalNAc, galactose, and both, respectively, assuming that they are functional.

species is currently fragmental, and their inactivating mechanisms of *O* alleles remain to be determined.

***A/B* allelism should have existed in primate ancestors, and later inactivation at population level resulted in ABO polymorphism.** Several primates exhibit ABO polymorphism, and the repertoire of types are species-dependent[40]. The inter-species sharing of the ABO polymorphism led Landsteiner and Wiener to conceive the theory of trans-species evolution of polymorphism. In this concept the allele coalescence time of the most recent common allele ancestor predates the speciation time. We previously determined partial nucleotide sequences of the *ABO* genes from several primate species and demonstrated that amino acid residues corresponding to codons 266 and 268 of human A/B transferases are conserved in all the species examined, depending on *A* or *B* allele[32]. Later evolutionary analyses led to the hypotheses of trans-species inheritance[17,22], convergent gene evolution[18–20], and a combination of those[21]. Because the *ABO* gene inheritance in primates was still controversial[46], we re-visited the topic for further evaluation, with additional experimental data on sugar specificity and activity of A/B transferases summarized in the code table.

Genome sequences in databases do not cover ABO polymorphism. Human reference and non-reference genes (both with LGG motif) in Ensemble database represent *O* and *A* alleles, respectively. The chimpanzee, gorilla, and macaque genes with LGG, MGA, and MGA, respectively, represent *A*, *B*, and *B* alleles from those species. In all the primate species the chromosomal region containing *ABO* gene is similar to the consensus with minor differences (Table 3). The current EMBL-EBI InterPro database hosts non-overlapping 65 ABO protein/peptide sequences, including several proteins with MGG, MGS, or LGA motif.

The phylogenetic trees of primate *ABO* genes are complex[22]. However, A and B specificity may be ascribed to amino acid residues corresponding to human codons 266 and 268 and their neighbors, by narrowing down the scanning window. In this investigation we, instead, evaluated the convergent evolution theory from an enzymological point of view. As shown in Table 2, the A to B conversion of sugar specificity may be achieved not only by the change from LGG to MGA motif, but also by other amino acid substitutions and even with single amino acid substitutions. Note that only one base change may be sufficient for the conversion to FGG, HGG, or YGG motif with B specificity. The B to A conversion is also possible by changing to other amino acids than LGG. However, the conversion from MGA to an A specific motif may need at least 2 nucleotide changes, even for the single amino acid substitution to PGA.

Therefore, it is difficult to assume that the same LGG <-> MGA conversion occurred in so many different occasions during the evolution period of primates. Selection after random mutation(s) does not explain the convergent evolution hypothesis because other motifs than LGG and MGA are also enzymatically functional (see Table 2). Rather, current distribution may be easily explained by assuming that functional *A* and *B* alleles were both present in the common ancestors of primates.

Bacterial *ABO* genes evolved into 2 separate groups with different sugar specificities through horizontal and vertical gene transmission. In addition to eukaryotes, ABO specificity also exists in prokaryotes, especially in Gram-negative bacteria, which constitute the bulk of intestinal flora[47]. The first two *ABO* genes cloned from bacteria are from O_{86} strain of *Escherichia coli* and from *Helicobacter mustelae*, which express B and A antigens, respectively[48,49]. Analyzing 19 bacterial genes, horizontal gene transfer between eukaryotes and prokaryotes and among bacteria was proposed to explain the absence of *ABO* genes in many species of invertebrates, plants, and fungi[50]. Because recent microorganism genome sequencings have identified additional bacterial *ABO* genes, we analyzed 56 bacterial proteins in EMBL-EBI InterPro database, and constructed phylogenetic trees of bacterial *ABO* genes. A tree is shown in Fig. 3c. In contrast to vertebrate *ABO* genes, all the bacterial *A* genes with GalNAc specificity segregated from the *B* or *AB* genes with galactose or GalNAc/galactose specificity, respectively. Another important finding is that the bacterial *ABO* genes have a different variation in the amino acid sequence motif from the animal genes. AGG and CGG motifs were found in the *A* gene sequences, MGS and QGC in the *B* gene sequences, and MGG, QGG, and TGA were in the *AB* gene sequences. Whereas many of the motifs found in the bacterial *ABO* genes are also present in animal genes, QGG motif seems to be unique to bacteria. TGA motif was found in animal *ABO* retropseudogenes as described above. In *Bacteroides* and *Parabacteroides* species *ABO* genes were clustered separately for possible *A* gene sequences with AGG and possible *AB* gene sequences with QGG or MGG. In other bacterial species their genes were grouped in either of the two big clusters of *A* or *B*/*AB* genes.

Discussion

What is the evolutionary significance of the *ABO* gene and its polymorphism? We tackled this question, employing an integrative approach with standard phylogenetic techniques combined with molecular enzymology. Based on gene distribution, we first concluded that A/B transferase gene appeared after the separation of fish and amphibian lineages. Requirement of A/B transferases for an α1,2-linked fucosylated substrate strongly supports preceding emergence of α1,2-FT genes over A/B transferase genes. In this context it is noteworthy that coelacanth has a *FUT2* gene sequence (although its functionality is questionable) and no *ABO* gene sequence. However, because coelacanth genome sequence is preliminary, a possibility remains that *ABO* gene may also exist in coelacanth. If this happens to be true, A/B gene appearance may be dated back to the time of lobe-finned fish appearance.

We created a code table correlating amino acid sequence motif with A/B specificity (Table 2). However, it should be noticed that having an active enzyme motif does not guarantee the gene function and sugar specificity. Mutation(s) in other position(s) may spoil the enzymatic activity[51]. Care must be taken to interpret the results because sugar specificity is based on the assumption that gene sequences encode functional glycosyltransferases, which is not always the case. *A* and *B* gene sequences can be *O*, depending on their functionality context[42]. Moreover, the table reveals one discordance, concerning the AYVYGS motif. The human A transferase construct containing this motif (in place of FYYLGG) at codons 263–268 did not exhibit A transferase activity whereas the *H. mustelae* bacterial gene having this sequence was reported to exhibit A activity. We assume that structural differences in other portions of the bacterial enzyme may have compensated for the activity variation.

We identified multiple copies of *ABO* gene sequences in a variety of species (Fig. 1), some of which possess sequence(s) with *A*-specific motif(s) and sequence(s) with *B*-specific motif(s) (Table 3). If multiple copies are found only in one species, the possibility exists that they were erroneously assembled. However, because this was observed in several different species, it seems unlikely that all those findings may be artifacts. In case of rats ABO gene duplication seems undeniably proved[33,37]. The number of species having both *A* and *B* gene sequences is expected to increase as new genome sequencing projects proceed, providing that duplicated regions are properly solved, which may be somewhat difficult in most NextGen sequencing projects. Irrespective of A/B specificity, phylogenetic analyses clustered those *ABO* gene sequences into a single cluster that was separated from the clusters of other α1,3-Gal(NAc)T genes (*GBGT1*, *A3GALT2*, *GGTA1*, and *GLT6D1*) (Fig. 2).

It is evident that animal *A* and *B* genes did not evolve into two separate genetic entities. Apparently, evolution suppressed the establishment of independent, functional *A* and *B* genes by certain

mechanism(s). However, proximity in genetic distance does not seem to be responsible for this failed separation in spite of the fact that *A* and *B* gene sequences are situated very closely on a chromosome in some species. *GGTA1(-1)*, *GGTA1(-2)*, and *GLT6D1(-1)* genes are also closely linked, as well as *SEC1* and *FUT2* genes (Table 1). These genes, however, took independent evolution paths, as opposed to *A* and *B* gene sequences which did not. As shown in Fig. 1, the majority of *GBGT1*, *A3GALT2*, and *GGTA1* genes possess conserved motifs of GGA, HAA, and HAA, respectively. This restriction strongly suggests that those motifs are vital to their glycosylation reactions. However, there are some variations in the motif with *ABO* gene and more with *GLT6D1* gene. *A* and *B* genes encode glycosyltransferases with distinct sugar specificity. However, both A and B transferases utilize the same H substances. Although this sharing of acceptor substrates may have contributed to mutual dependence of those two genes to a certain degree, it is not sufficient because *SEC1* and *FUT2* genes encoding α1,2-FTs with similar enzymatic characteristics still formed separate phylogenetic clusters.

Two modes of appearance and inheritance of *A* and *B* gene sequences in a given animal species may be contemplated to explain the results in Fig. 3a. One is that those sequences with different sugar specificity appeared recurrently after the separation from other analyzed species by convergent mutations. Another much likely possibility is that those sequences may have attained species-specific sequence homology through intergenic exchanges after A/B specificity was inherited from common ancestral genes. An examination of gene organization revealed that full genes with initiation and termination codons are rare in those species possessing multiple *ABO* gene copies. Many are partial genes that are incapable of encoding functional glycosyltransferases by themselves. We speculate that they may serve as a reservoir for genetic diversity to switch A/B specificity through gene conversion, exon shuffling, or recombination. In several species multiple *ABO* gene sequences are closely linked to one another, which facilitates recombination/gene conversion without genetic catastrophe, producing new possible adaptations at a higher rate than by nucleotide substitutions.

As mentioned above with rats, insertion/deletion/unequal crossovers/gene conversion seems to have occurred frequently at the *ABO* gene locus. It may have reduced gene number from several to one on certain occasions. Therefore, it is not too far-fetched to hypothesize that differential deletions/crossovers may have resulted in differential outcomes. Starting from tandemly linked *A* and *B* gene sequences, *A* and *B* alleles may have been created (the multigenic-to-unigenic transition hypothesis). New functional allele(s) may have been generated within partial and nonfunctional sequence(s) so far as changes in gene organization could restore their functionality to encode active enzymes that are expressed after being inserted or copied in the functional gene(s). An example of such restored function (and not merely changing it) has recently been demonstrated of human *A* allele by recombination from functional *B* allele and nonfunctional *O* allele[52]. Those events may have taken place before simians appeared. Rats and rabbits have *A* genes with AGG and LGG, respectively. Therefore, prosimians and simians may have inherited an *A* gene with LGG similar to Lagomorpha genes, rather than Rodentia genes, because no genes with AGG motif are found in primates[22]. An alternative explanation would be the unigenic-to-multigenic transition hypothesis: *A/B* allelism appeared first and then natural selection favored duplication events in many species to separate both alleles whereas this separation did not occur in primates. This is an interesting hypothesis because it may easily explain the absence of separate evolution of *A* and *B* genes. However, it seems to be less likely because all the other species than primates, which are known to have unigenic polymorphism, exhibit AO, and not AB, polymorphism[43].

Based on the relationship between amino acid motifs and A/B specificity, we have shown that *A* and *B* alleles with LGG and MGA motifs, respectively, existed in common ancestors of primates. This suggests that they were inherited, most probably, in a trans-species manner. However, the fact that other motifs than LGG and MGA also exist in some primate species signifies that mutations/recombination also happened, of which several may be the result of convergent evolution. For instance, LGA motif is found in Ecuadorian squirrel monkeys and humans, and MGG is found in Ecuadorian squirrel monkeys, Weeper capuchins, and humans, although cases of *cis-AB* (with LGA or MGG motif) are rare in humans. These motifs may be derived from either LGG or MGA by point mutation or by recombination of those two alleles, still supporting the inheritance of an ancestral polymorphism with *A* allele (LGG) and *B* allele (MGA) as prototypic alleles. MGS motif in titi monkeys may have resulted from MGA by a single nucleotide substitution, rather than from LGG by 2 amino acid substitutions.

In addition to primates, many other animal species analyzed also maintain the prevailing motifs of LGG and MGA although AGG is also frequent in non-primate animals. Considering that additional motifs may also render the *ABO* gene-encoded proteins enzymatically active as demonstrated in the code table, those 3 motifs may be considered ancestral for those species. However, to evaluate this possibility further characterization of additional *ABO* genes from many other species, including amphibians and reptiles, will be needed. *ABO* genes seem to have evolved under more or less constant selective pressure for some polymorphism in their catalytic specificity, which in some species is achieved by carrying different gene copies (multigenic polymorphism) and in some other species through allelic polymorphism of a single gene (unigenic polymorphism). Whether the latter is limited to primate species or not needs to be determined in order to conclusively prove or disprove the multigenic-to-unigenic transition hypothesis.

The A/B antigen expression depends on the A/B genotype of individual. Although human and several other species express A/B antigens on red blood cells, the expression on RBCs is relatively rare. On the contrary, epithelial cells, including those of the gastrointestinal tract, express A/B antigens in many species. Accordingly, its significance may be better found in that cell-type. Many of cell-surface oligosaccharide structures are involved in microbial interactions, and ABH antigens are not an exception[53]. Actually, ABO polymorphism has been associated with certain infectious diseases[54–56]. The presence/absence of A/B antigens and concordant absence/presence of anti-A/B antibodies provide strong defensive lines against infection. Having *ABO* gene should be beneficial because many vertebrate species maintain this gene. However, having both functional *A* and *B* genes ubiquitously within species might not be so advantageous because they may eventually lose anti-A/B antibodies. Rather, frequent gene conversion of A/B specificity producing amino acid substitutions or recombination with nonfunctional partial genes may have conferred an adaptation against microbial attacks. Different ABO phenotypes in different species and ABO polymorphism within species may inhibit inter-species and intra-species infections, respectively. Our results conformed to the hypothesis that host organisms attained the variation utilizing those two molecular mechanisms.

We unexpectedly observed the separate clustering of bacterial *ABO* genes into 2 groups with different sugar specificities (*A* and *B*/*AB* genes) (Fig. 3c), as opposed to animal *ABO* genes, of which *A* and *B* genes did not evolve independently. Widespread presence of *A/B* genes in bacteria[47] indicates that ABO mimicry is advantageous to survival. The bacterial *ABO* genes have been transmitted horizontally to different bacteria and vertically through generations. We reason that these mixed modes of gene inheritance have allowed the segregated evolution of the bacterial *ABO* genes in 2 groups. It is evident that horizontal gene transfer has been providing bacteria with easier adaptation against host defense system. Contrastingly, interactions with infectious agents may have stimulated the host

ABO gene evolution, as intra-species polymorphism may help the survival of host species by changing allele frequency through balancing selection.

In conclusion, the systematic functional analysis correlating amino acid sequence motifs with A/B specificities opened a new venue to investigate the *ABO* gene and protein evolution. Together with phylogenetic analyses, we have gained invaluable insights into the evolutionary significance of the *ABO* gene and its polymorphism and successfully decoded several important questions.

Methods

Materials. Reagents for PCR, restriction endonucleases, T4 DNA ligase, and other enzymes were purchased from LifeTechnologies (Carlsbad, CA) and New England BioLabs (Ipswich, MA). HeLa cells, human cancer cells of uterus, were originally obtained from American Type Culture Collection (ATCC), and have been maintained in the laboratory over a decade. Cell culture media, frozen transformation-competent *E. coli* bacteria, and Lipofectamine 2000 were also purchased from LifeTechnologies. Oligodeoxynucleotides were custom-synthesized at the same company. Anti-A and anti-B murine monoclonal antibody mixtures were from OrthoDiagnostic Systems (Piscataway, NJ), and Vectastain ABC System and DAB (3, 3′-diaminobenzidine) substrate for color development were from Vector Laboratories (Burlingame, CA).

***In vitro* mutagenesis of human A transferase expression construct.** We employed a PCR-mediated *in vitro* mutagenesis approach as previously described[30]. Degenerate oligodeoxynucleotides were used to introduce amino acid substitutions at codon 266 and 268 of human A transferase. The primers originally used for a library construction were the followings:

FYV7 (T7-F): 5′-TAATACGACTCACTATAGGG
FYV1 (SV40 polyA-R): 5′-GAAATTTGTGATGCTATTGC
IMPPC235 (F): GGCGATTTCTACTACNNNGGGGSGTTCTTCGGGGGGTC
IMPPC236 (R): GACCCCCCGAAGAACSCCCCCNNNGTAGTAGAAATCGCC

The capitalized underlined letters N and S denote a mixture of 4 nucleotides (G/A/T/C) and 2 nucleotides (G/C) at those positions. Human A transferase expression construct[57] prepared in pSG-5 vector (Stratagene, La Jolla, CA) was used as a PCR template. Two consecutive rounds of PCR reactions were performed, first with FYV7 (T7-F) and IMPPC236 (R) primers and separately with IMPPC235 (F) and FYV1 (SV40 polyA-R) primers, and second by mixing both the reactions. The PCR products were cleaved with *SacII* and *BamHI* restriction enzymes, and ligated with the *SacII-BamHI* vector fragment of human A transferase expression construct. After DNA transformation of *E. coli* bacteria, plasmid DNA was prepared from transformant colonies, sequenced, and the constructs containing intended amino acid substitutions but lacking additional non-synonymous mutations were selected for DNA transfection experiments. For those constructs, which we failed to obtain by using degenerate oligodeoxynucleotide primers, and those constructs, which were not covered by the library approach, specific primers were designed for individual constructions (not shown).

DNA transfection and immunostaining. HeLa cells were used as a recipient of DNA transfection. These cells were derived from a type O individual and exhibit cell surface H substances. When functional A/B transferases are expressed by DNA transfection, H substances are converted to A/B antigens. We have used this system at various occasions to examine the specificity and activity of A/B transferase variants[30,57,58]. DNA transfection experiments were performed using 96-well plates as previously described[59]. Lipofectamine 2000 reagent was used, following the manufacturer's instructions. DNA from the *FUT2* expression construct prepared in pSG-5 and DNA from the pEGFP-N1 vector (GenBank Accession #U55762) were co-transfected: the former to increase the acceptor substrate availability and the latter to calculate the transfection efficiency for activity adjustment. Two days after DNA transfection, GFP-positive cells were counted. The next day, cells were fixed with paraformaldehyde and washed with PBS. After drying, cells were treated first with either anti-A or anti-B monoclonal antibodies, second with biotinylated anti-mouse IgM, then with Avidin/Biotinylated Peroxidase Complex (ABC), followed by color development using DAB substrate. Stained cells were counted microscopically, and A/B specificity and activity were determined after adjusting the transfection efficiency using GFP-positive cell counts. Because of variable detachment of cells from dish substratum during fixation and immunostaining procedures, data were presented in a semi-quantitative manner.

Databases, sequence alignment, and construction of phylogenetic trees. Nucleotide and amino acid sequences, exon-intron organizations, and chromosomal locations of α1,2-FT genes (*FUT1*/*FUT2*/*SEC1*) and α1,3-Gal(NAc)T genes (*ABO*/*GBGT1*/*A3GALT2*/*GGTA1*/*GLT6D1*) were retrieved from Ensembl (www.ensembl.org/index.html) and GenBank (www.ncbi.nlm.nih.gov/genbank/) genome sequence databases. Protein/peptide sequences of the *ABO* genes were retrieved from the EMBL-EBI InterPro database (www.ebi.ac.uk/interpro/).

Ensembl genome sequence database (release 73) listed 89 annotated α1,2-FT genes with 66 speciation nodes and 15 duplications in the ENSGT00390000001450 gene tree and 255 annotated α1,3-Gal(NAc)T genes with 185 speciation nodes and 65 duplications in the ENSGT00400000022032 gene tree. The phylogenetic tree in Fig. 2 was constructed by the neighbor-joining method[60]. JTT model[61] was used for estimating number of amino acid substitutions and 1,000 bootstrap replications were computed by using MEGA5[62]. The phylogenetic trees in Fig. 3 were constructed by Maximum Likelihood method, using the same software.

1. Watkins, W. M. The ABO blood group system: historical background. *Transfus Med* **11**, 243–265 (2001).
2. Daniels, G. *Human blood groups* (Blackwell Science, Oxford, 2002).
3. Yamamoto, F. *et al.* 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 (1990).
4. Yamamoto, F., Clausen, H., White, T., Marken, J. & Hakomori, S. Molecular genetic basis of the histo-blood group ABO system. *Nature* **345**, 229–233 (1990).
5. Yamamoto, F. *et al.* Molecular genetic analysis of the ABO blood group system: 1. Weak subgroups: A^3 and B^3 alleles. *Vox Sang* **64**, 116–119 (1993).
6. Yamamoto, F. *et al.* Molecular genetic analysis of the ABO blood group system: 2. *Cis*-AB alleles. *Vox Sang* **64**, 120–123 (1993).
7. Yamamoto, F., McNeill, P. D., Yamamoto, M., Hakomori, S. & Harris, T. Molecular genetic analysis of the ABO blood group system: 3. A^x and $B^{(A)}$ alleles. *Vox Sang* **64**, 171–174 (1993).
8. Yamamoto, F. *et al.* Molecular genetic analysis of the ABO blood group system: 4. Another type of O allele. *Vox Sang* **64**, 175–178 (1993).
9. Yamamoto, F., Cid, E., Yamamoto, M. & Blancher, A. ABO research in the modern era of genomics. *Transfus Med Rev* **26**, 103–118 (2012).
10. Joziasse, D. H., Shaper, J. H., Van den Eijnden, D. H., Van Tunen, A. J. & Shaper, N. L. Bovine alpha 1->3-galactosyltransferase: isolation and characterization of a cDNA clone. Identification of homologous sequences in human genomic DNA. *J Biol Chem* **264**, 14290–14297 (1989).
11. Larsen, R. D. *et al.* Isolation of a cDNA encoding a murine UDPgalactose:beta-D-galactosyl- 1,4-N-acetyl-D-glucosaminide alpha-1,3-galactosyltransferase: expression cloning by gene transfer. *Proc Natl Acad Sci U S A* **86**, 8227–8231 (1989).
12. Haslam, D. B. & Baenziger, J. U. Expression cloning of Forssman glycolipid synthetase: a novel member of the histo-blood group ABO gene family. *Proc Natl Acad Sci U S A* **93**, 10697–10702 (1996).
13. Keusch, J. J., Manzella, S. M., Nyame, K. A., Cummings, R. D. & Baenziger, J. U. Expression cloning of a new member of the ABO blood group glycosyltransferases, iGb3 synthase, that directs the synthesis of isoglobo-glycosphingolipids. *J Biol Chem* **275**, 25308–25314 (2000).
14. Schaefer, A. S. *et al.* A genome-wide association study identifies GLT6D1 as a susceptibility locus for periodontitis. *Hum Mol Genet* **19**, 553–562 (2010).
15. Turcot-Dubois, A. L. *et al.* Long-term evolution of the CAZY glycosyltransferase 6 (ABO) gene family from fishes to mammals--a birth-and-death evolution model. *Glycobiology* **17**, 516–528 (2007).
16. Casals, F. *et al.* Human pseudogenes of the ABO family show a complex evolutionary dynamics and loss of function. *Glycobiology* **19**, 583–591 (2009).
17. Martinko, J. M., Vincek, V., Klein, D. & Klein, J. Primate ABO glycosyltransferases: evidence for trans-species evolution. *Immunogenetics* **37**, 274–278 (1993).
18. Saitou, N. & Yamamoto, F. Evolution of primate ABO blood group genes and their homologous genes. *Mol Biol Evol* **14**, 399–411 (1997).
19. O'HUigin, C., Sato, A. & Klein, J. Evidence for convergent evolution of A and B blood group antigens in primates. *Hum Genet* **101**, 141–148 (1997).
20. Doxiadis, G. G. *et al.* Characterization of the ABO blood group genes in macaques: evidence for convergent evolution. *Tissue Antigens* **51**, 321–326 (1998).
21. Noda, R., Kitano, T., Takenaka, O. & Saitou, N. Evolution of the ABO blood group gene in Japanese macaque. *Genes Genet Syst* **75**, 141–147 (2000).
22. Segurel, L. *et al.* The ABO blood group is a trans-species polymorphism in primates. *Proc Natl Acad Sci U S A* **109**, 18493–18498 (2012).
23. Roubinet, F. *et al.* Evolution of the O alleles of the human ABO blood group gene. *Transfusion* **44**, 707–715 (2004).
24. Calafell, F. *et al.* Evolutionary dynamics of the human ABO gene. *Hum Genet* **124**, 123–135 (2008).
25. Huson, D. H. & Scornavacca, C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst Biol* **61**, 1061–1067 (2012).
26. Nei, M. & Rooney, A. P. Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet* **39**, 121–152 (2005).
27. Apoil, P. A. *et al.* Evolution of alpha 2-fucosyltransferase genes in primates: relation between an intronic Alu-Y element and red cell expression of ABH antigens. *Mol Biol Evol* **17**, 337–351 (2000).
28. Yamamoto, S. [Blood group substances in toads and frogs. I. Serological analysis of the red cells and gastric mucosa of *Xenopus laevis* and *Rana catesbiana*]. *Igaku To Seibutsugaku* **79**, 83–87 (1969).
29. Wrann, M., Schenkel-Brunner, H. & Kothbauer, H. Blood-group A and G specific structures in toad (*Bufo*) spawn. Comparative studies on three species (*Bufo bufo*, *Bufo viridis*, *Bufo calamita*). *Z Immunitatsforsch Immunobiol* **154**, 471–473 (1978).
30. Yamamoto, F. & McNeill, P. D. Amino acid residue at codon 268 determines both activity and nucleotide-sugar donor substrate specificity of human histo-blood

group A and B transferases: *In vitro* mutagenesis study. *J Biol Chem* **271**, 10515–10520 (1996).
31. Patenaude, S. I. *et al.* The structural basis for specificity in human ABO(H) blood group biosynthesis. *Nat Struct Biol* **9**, 685–690 (2002).
32. Kominato, Y. *et al.* Animal histo-blood group ABO genes. *Biochem Biophys Res Commun* **189**, 154–164 (1992).
33. Iwamoto, S. *et al.* Rat encodes the paralogous gene equivalent of the human histo-blood group ABO gene. Association with antigen expression by overexpression of human ABO transferase. *J Biol Chem* **277**, 46463–46469 (2002).
34. Yamamoto, M. *et al.* Murine equivalent of the human histo-blood group ABO gene is a *cis*-AB gene and encodes a glycosyltransferase with both A and B transferase activity. *J Biol Chem* **276**, 13701–13708 (2001).
35. Olson, F. J. *et al.* Blood group A glycosyltransferase occurring as alleles with high sequence difference is transiently induced during a *Nippostrongylus brasiliensis* parasite infection. *J Biol Chem* **277**, 15044–15052 (2002).
36. Cailleau-Thomas, A. *et al.* Cloning of a rat gene encoding the histo-blood group A enzyme. Tissue expression of the gene and of the A and B antigens. *Eur J Biochem* **269**, 4040–4047 (2002).
37. Turcot, A. L. *et al.* Cloning of a rat gene encoding the histo-blood group B enzyme: rats have more than one Abo gene. *Glycobiology* **13**, 919–928 (2003).
38. Ogasawara, K. *et al.* Extensive polymorphism of ABO blood group gene: three major lineages of the alleles for the common ABO phenotypes. *Hum Genet* **97**, 777–783 (1996).
39. Olsson, M. L., Guerreiro, J. F., Zago, M. A. & Chester, M. A. 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 (1997).
40. Moor-Jankowski, J., Wiener, A. S. & Rogers, C. M. Human blood group factors in non-human primates. *Nature* **202**, 663–665 (1964).
41. Moor-Jankowski, J. & Wiener, A. S. Blood group antigens in primate animals and their relation to human blood groups. *Primates in Medicine* **3**, 64–77 (1969).
42. Kermarrec, N., Roubinet, F., Apoil, P. A. & Blancher, A. Comparison of allele O sequences of the human and non-human primate ABO system. *Immunogenetics* **49**, 517–526 (1999).
43. Oriol, R. *et al.* Major carbohydrate epitopes in tissues of domestic and African wild animals of potential interest for xenotransplantation research. *Xenotransplantation* **6**, 79–89 (1999).
44. Yamamoto, F. & Yamamoto, M. Molecular genetic basis of porcine histo-blood group AO system. *Blood* **97**, 3308–3310 (2001).
45. Nguyen, D. T. *et al.* Molecular characterization of the human ABO blood group orthologus system in pigs. *Anim Genet* **42**, 325–328 (2011).
46. Blancher, A. Evolution of the ABO supergene family. *ISBT Science Series* **8**, 201–206 (2013).
47. Springer, G. F. Importance of blood-group substances in interactions between man and microbes. *Ann N Y Acad Sci* **169**, 134–152 (1970).
48. Yi, W. *et al. Escherichia coli* O_{86} O-antigen biosynthetic gene cluster and stepwise enzymatic synthesis of human blood group B antigen tetrasaccharide. *J Am Chem Soc* **127**, 2040–2041 (2005).
49. Yi, W., Shen, J., Zhou, G., Li, J. & Wang, P. G. Bacterial homologue of human blood group A transferase. *J Am Chem Soc* **130**, 14420–14421 (2008).
50. Brew, K., Tumbale, P. & Acharya, K. R. Family 6 glycosyltransferases in vertebrates and bacteria: inactivation and horizontal gene transfer may enhance mutualism between vertebrates and bacteria. *J Biol Chem* **285**, 37121–37127 (2010).
51. Marcus, S. L. *et al.* A single point mutation reverses the donor specificity of human blood group B-synthesizing galactosyltransferase. *J Biol Chem* **278**, 12403–12405 (2003).
52. Kitano, T., Blancher, A. & Saitou, N. The functional A allele was resurrected via recombination in the human ABO blood group gene. *Mol Biol Evol* **29**, 1791–1796 (2012).
53. Seymour, R. M., Allan, M. J., Pomiankowski, A. & Gustafsson, K. Evolution of the human ABO polymorphism by two complementary selective pressures. *Proc Biol Sci* **271**, 1065–1072 (2004).
54. Boren, T., Falk, P., Roth, K. A., Larson, G. & Normark, S. Attachment of *Helicobacter pylori* to human gastric epithelium mediated by blood group antigens. *Science* **262**, 1892–1895 (1993).
55. Cserti, C. M. & Dzik, W. H. The ABO blood group system and *Plasmodium falciparum* malaria. *Blood* **110**, 2250–2258 (2007).
56. Anstee, D. J. The relationship between blood groups and disease. *Blood* **115**, 4635–4643 (2010).
57. Yamamoto, F., Yamamoto, M. & Blancher, A. Generation of histo-blood group B transferase by replacing the N-acetyl-D-galactosamine recognition domain of human A transferase with the galactose-recognition domain of evolutionarily related murine alpha1,3-galactosyltransferase. *Transfusion* **50**, 622–630 (2010).
58. Yamamoto, F. & Hakomori, S. Sugar-nucleotide donor specificity of histo-blood group A and B transferases is based on amino acid substitutions. *J Biol Chem* **265**, 19257–19262 (1990).
59. Yamamoto, M., Cid, E. & Yamamoto, F. Molecular genetic basis of the human Forssman glycolipid antigen negativity. *Sci Rep* **2**, 975 (2012).
60. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406–425 (1987).
61. Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**, 275–282 (1992).
62. Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**, 2731–2739 (2011).

Acknowledgments

This work was supported by the Instituto de Salud Carlos III (grant number PI11/00454) from the Spanish Government and the fund from Institut de Medicina Predictiva i Personalitzada del Càncer (IMPPC) to FY. We thank Mrs. Masako Mizuguchi for phylogenetic construction of Fig. 2 and Ms. Patricia Barrero for technical assistance.

Author contributions

F.Y. conceived the project. M.Y., E.C. and F.Y. prepared amino acid substitution constructs of the human A transferase, performed DNA transfection experiments, and immunologically determined the A/B specificity of the individual constructs. F.Y. retrieved sequence data and other information from databases, F.Y. and N.S. prepared phylogenetic trees, and F.Y., N.S., J.B. and A.B. analyzed and interpreted results. F.Y. wrote the manuscript draft, and all the other authors participated in revision and editing.

Additional information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Yamamoto, F. *et al.* An integrative evolution theory of histo-blood group *ABO* and related genes. *Sci. Rep.* **4**, 6601; DOI:10.1038/srep06601 (2014).