REGULAR ARTICLE

blood advances

Haplotype sequence collection of *ABO* blood group alleles by long-read sequencing reveals putative *A1*-diagnostic variants

Morgan Gueuning,[1,*] Gian Andri Thun,[1,*] Michael Wittig,[2] Anna-Lena Galati,[3] Stefan Meyer,[4] Nadine Trost,[4] Elise Gourri,[1,4] Janina Fuss,[2] Sonja Sigurdardottir,[4] Yvonne Merki,[4] Kathrin Neuenschwander,[4] Yannik Busch,[3] Peter Trojok,[3] Marco Schäfer,[3] Jochen Gottschalk,[5] Andre Franke,[2] Christoph Gassner,[2,6] Wolfgang Peter,[3,7] Beat M. Frey,[1,4,5] and Maja P. Mattle-Greminger[1]

[1]Department of Research and Development, Blood Transfusion Service Zurich, Swiss Red Cross (SRC), Schlieren, Switzerland; [2]Institute of Clinical Molecular Biology, Christian Albrechts University of Kiel, Kiel, Germany; [3]Stefan Morsch Foundation, Birkenfeld, Germany; [4]Department of Molecular Diagnostics and Cytometry, Blood Transfusion Service Zurich, SRC, Schlieren, Switzerland; [5]Department of Pathogen Screening, Blood Transfusion Service Zurich, SRC, Schlieren, Switzerland; [6]Institute for Translational Medicine, Private University in the Principality of Liechtenstein, Triesen, Liechtenstein; and [7]Institute for Transfusion Medicine, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany

Key Points

- The first comprehensive collection of full-length haplotype sequences for all 6 main *ABO* allele groups will support *ABO* genetic analyses.
- *ABO* genetic diversity patterns revealed putatively *ABO*A1*-diagnostic variants, which could finally enable direct genetic typing of *A1*.

In the era of blood group genomics, reference collections of complete and fully resolved blood group gene alleles have gained high importance. For most blood groups, however, such collections are currently lacking, as resolving full-length gene sequences as haplotypes (ie, separated maternal/paternal origin) remains exceedingly difficult with both Sanger and short-read next-generation sequencing. Using the latest third-generation long-read sequencing, we generated a collection of fully resolved sequences for all 6 main *ABO* allele groups: *ABO*A1/A2/B/O.01.01/O.01.02/O.02*. We selected 77 samples from an ABO genotype data set (n = 25 200) of serologically typed Swiss blood donors. The entire *ABO* gene was amplified in 2 overlapping long-range polymerase chain reactions (covering ~23.6 kb) and sequenced by long-read Oxford Nanopore sequencing. For quality validation, 2 samples per *ABO* group were resequenced using Illumina and Pacific Biosciences technology. All 154 full-length *ABO* sequences were resolved as haplotypes. We observed novel, distinct sequence patterns for each *ABO* group. Most genetic diversity was found between, not within, *ABO* groups. Phylogenetic tree and haplotype network analyses highlighted distinct clades of each *ABO* group. Strikingly, our data uncovered 4 genetic variants putatively specific for *ABO*A1*, for which direct diagnostic targets are currently lacking. We validated *A1*-diagnostic potential using whole-genome data (n = 4872) of a multiethnic cohort. Overall, our sequencing strategy proved powerful for producing high-quality *ABO* haplotypes and holds promise for generating similar collections for other blood groups. The publicly available collection of 154 haplotypes will serve as a valuable resource for molecular analyses of *ABO*, as well as studies about the function and evolutionary history of *ABO*.

Submitted 11 February 2022; accepted 3 September 2022; prepublished online on *Blood Advances* First Edition 21 September 2022. https://doi.org/10.1182/bloodadvances.2022007133.

*M.G. and G.A.T. contributed equally to this study.

The data reported in this article have been deposited in the National Center for Biotechnology Information GenBank sequence database (accession numbers OM283861-OM284014). A detailed list of sequence accession numbers is provided in supplemental Table 2.

Sequence alignments are available from the Dryad Digital Repository, https://doi.org/10.5061/dryad.q573n5tkj.

Data are available on request from the corresponding author, Maja P. Mattle-Greminger (m.mattle@zhbsd.ch).

The full-text version of this article contains a data supplement.

Introduction

Generating reference sequence collections of blood group gene alleles has gained importance, especially as genomic technologies are being more widely used in molecular diagnostics of blood groups.[1-7] Such comprehensive collections are, for instance, essential for designing and validating blood group genotyping assays to reduce risks of unnoticed allelic dropout,[8,9] for imputing genotype data in microarray analyses,[10] as references for analyzing next-generation sequencing data to increase diagnostic reliability, for resolving complex genotype-phenotype discrepancies in routine diagnostics, and for disentangling the evolutionary history of the respective gene.[11-13]

Importantly, reference sequences for blood group gene alleles should (1) span the complete gene region, including introns and appropriate parts of the adjacent flanking regions; (2) have a fully resolved haplotype; (3) offer confirmed serology; and (4) be well accessible in a public sequence database.[14] Generating such sequences, however, is technically difficult. Particularly challenging is resolving the haplotype, that is, determining which variants were inherited together from the mother or father, and thus, lie on the same haplotype.

Classical Sanger sequencing is not useful to resolve haplotypes for technical reasons. It can only be used in combination with laborious allele-specific polymerase chain reactions (PCRs), as shown for short or highly conserved blood group genes with little variation.[15,16] Haplotype reconstruction with short-read next-generation sequencing is restricted by read length. It must, therefore, mainly rely on statistical methods, which are predictive and dependent on well-suited population-based data.[17] Owing to these technical challenges, full-length gene haplotype sequences remain rare for most of the 43 described blood group systems,[18] even for the ones deemed clinically most relevant. This, even though hundreds of thousands of genomes have already been sequenced by now.[19,20]

Latest third-generation long-read sequencing finally opened new avenues for the haplotype reconstruction issue. The key advantage over short-read sequencing is the power to sequence very long DNA fragments as haplotype. Pacific Biosciences' (PacBio) HiFi technology, in which fragments are circularly sequenced several times to directly obtain high-quality consensus sequences, provides reads of 10 to 25 kb.[21] Nanopore sequencing by Oxford Nanopore Technologies (ONT), using changes in ionic currents when molecules pass through nanopores, can even sequence much longer fragments.[22] One of the main advantages of ONT, however, is its scalability, which also allows cost-effectiveness in low-throughput settings. Long-read sequencing has great potential in the quest for comprehensive haplotype sequence collections of blood group gene alleles. Currently, reports of its application are still limited to the very short *ACKR1* gene encoding the Duffy blood group.[1,23]

Even for ABO, the first discovered,[24] and clinically most important blood group system,[25] haplotype sequence collections are still lacking. The most comprehensive collection of *ABO* sequences by short-read sequencing only covers a small exonic part of the gene.[14] A recent approach to tackling the entire gene was hindered by the high degree of repetitive elements in introns, which were finally omitted. This resulted in published sequences with incomplete haplotype information.[26] Consequently, only 7 complete human *ABO* gene sequences have been deposited in the National Center for Biotechnology Information (NCBI) nucleotide database to date (accessed 15 July 2022).

The human *ABO* gene is ~19.5 kb long and located on chromosome 9 on the reverse strand (Figure 1). The reference transcript for the ABO blood group (NM_020469.3) contains 7 exons. Individuals with codominant *ABO*A* or *B* alleles, defined by single-nucleotide variants (SNVs) in exons 6 and 7, express glycosyltransferase activities that convert the H antigen present on the red blood cells into the A or B antigen. The recessive O alleles (phenotypic ABO-null alleles) are mainly caused by either a frameshift (c.261delG) in exon 6, leading to a premature stop codon (*ABO*O.01*), or a SNV (c.802G>A) in exon 7 that causes inactivation of glycosyltransferase activity (*ABO*O.02*).

ABO is highly polymorphic, which is apparent in over 200 different alleles, covering predominantly exonic variation listed by the

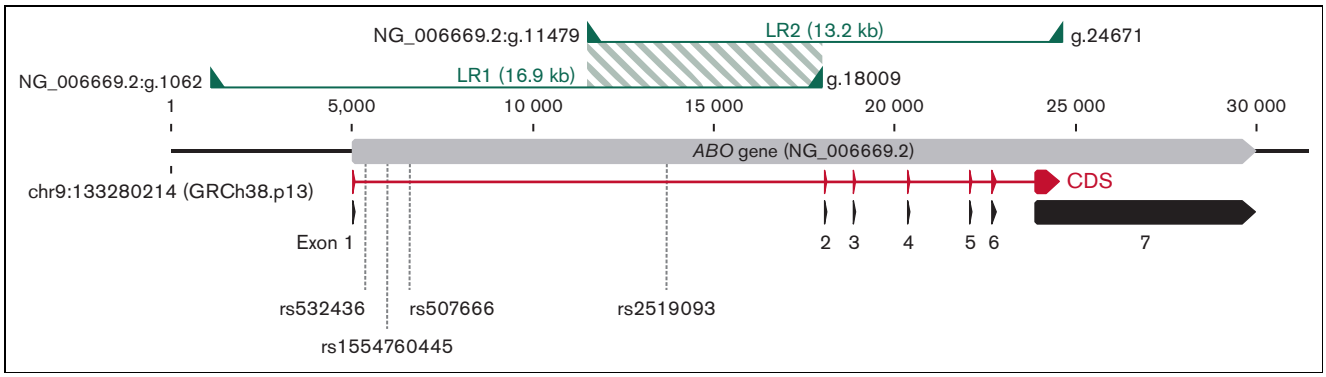

Figure 1. Genetic structure of the *ABO* gene locus, including long-range PCR (LR-PCR) amplicon locations and positions of the 4 putative *ABO*A1*-diagnostic variants. The coordinates above the gene and next to the PCR primers correspond to the reference sequence NG_006669.2 (LRG_792). Chromosomal coordinates of the first base pair of *ABO* reference gene on the current human genome reference (GRCh38.p13) is provided underneath the gene. Exons 1 to 7 are represented by black arrows, and the coding DNA sequence (CDS; reference transcript NM_020469.3) in red. The locations of the 2 overlapping LR-PCR amplicons (LR1 and LR2) used to amplify the *ABO* gene are highlighted by the striped area between both amplicons. The positions of the 4 putative *ABO*A1*-specific variants found in this study are indicated by their respective rs numbers. For graphical clearness, the *ABO* gene is shown as reverse complement.

Table 1. Overview number of haplotype sequences per *ABO* group

ABO group	No. of haplotypes	Estimated allele frequencies in the Zurich region of Switzerland† (%)
*ABO*A1*	39	17.41
*ABO*A2*	21	9.16
*ABO*B*	20	7.64
*ABO*O.01.01*	27	40.95
*ABO*O.01.02*	31	22.65
*ABO*O.02*	16	2.11
Total	**154**	**99.92**

A complete list of study samples (n = 77) is given in supplemental Table 1. Details on sequenced *ABO* haplotypes (n = 154) including GenBank accession numbers are provided in supplemental Table 2.

†Estimated *ABO* allele frequencies in the region of Zurich in Switzerland based on MALDI-TOF mass spectrometry genotyping data (see supplemental Information Section 1.3).

International Society of Blood Transfusion (ISBT).[27] Genetic variation at *ABO* can be divided into 6 main allele groups: *ABO*A1*, *ABO*A2*, *ABO*B*, *ABO*O.01.01*, *ABO*O.01.02*, and *ABO*O.02*, which make up over 99.9% of the genetic diversity at *ABO* (Table 1).

Here, we aimed to generate a collection of high-quality reference sequences for the 6 main *ABO* allele groups, taking advantage of long-read nanopore sequencing for resolving haplotypes. We ensured sequence accuracy by validation with complementing sequencing technologies (ie, Sanger, Illumina, and PacBio sequencing). The simple and reliable protocol established in this study can be adapted to any other blood group system or essentially any gene, and, therefore, holds the promise of generating analogous collections to the one presented here.

Methods

Sample selection and *ABO* allele groups

Details on the sample set and the selection process are provided in the supplemental Information Section 1. Briefly, we selected 77 samples (supplemental Tables 1 and 2) from a large, well-characterized ABO genotype data set (n = 25 200) of serologically typed blood donors from the Zurich region in Switzerland. These data had been generated previously using MALDI-TOF mass spectrometry.[28] We aimed to sequence at least 15 haplotypes for each of the 6 main *ABO* allele groups, that is, *ABO*A1*, *A2*, *B*, *O.01.01*, *O.01.02*, and *O.02* (Table 1). The 2 *O.01* subgroups, *ABO*O.01.01* and *ABO*O.01.02* (formally known as O^{1v}),[29] were considered as 2 separate groups in this study because, apart from sharing the c.261delG causative O-phenotype variant, they are not closely related in evolutionary terms.[13,30]

Based on pretyped variants (supplemental Table 3), we selected a mix of (1) *ABO* homozygous samples (ie, same base inherited from the mother and father; n = 43) and (2) *ABO* group heterozygous samples (ie, 2 different bases inherited; n = 34). The putatively *ABO* gene homozygous individuals were included to support haplotype resolving after sequencing. Estimates of genotype frequencies for the whole population are given in supplemental Table 4. All donors gave their written informed consent for molecular blood group analyses. According to the cantonal and national Swiss legislation, molecular blood group analyses are not subject to ethical authorization.

LR-PCRs of *ABO* and nanopore sequencing

We established generic LR-PCRs amplifying the entire *ABO* gene, including flanking regulatory regions (~23.6 kb; exact length dependent on haplotype) in 2 overlapping fragments (Figure 1). Fragment LR1 (16.9 kb) covered the enhancer region (~4.1 kb upstream of exon 1) up to the end of intron 1. Fragment LR2 (13.2 kb) amplified half of intron 1 up to ~100 bp after the stop codon in exon 7. Both fragments overlapped by ~6.5 kb. Details on LR-PCRs are provided in the supplemental Information Section 2.1.

Nanopore sequencing libraries were prepared following ONT's protocol for native (ie, PCR-free) barcoding of amplicons (see supplemental Information Section 2.2). The final library was sequenced for 72 hours on 2 MinION Mk1B (R9.4.1) flow cells.

Bioinformatic analysis of nanopore sequencing data

Our workflow for processing nanopore sequencing data is depicted in supplemental Figure 1 and described in detail in supplemental Information Section 3. In short, raw reads were demultiplexed and base-called using ONT's Guppy (version 4.4.1) based on a high-accuracy model. The quality-filtered raw reads were then size-selected based on the expected length of the 2 LR-PCR fragments and the observed read length distribution. To reduce computational time for downstream analysis, we set a cutoff at 1000 reads per amplicon by random downsampling with seqtk (version 1.3).

To circumvent potential biases of allelic dropout in classical single-reference–based read mapping,[31,32] we assembled for each sample its own consensus sequence from both PCR amplicons de novo (ie, reference-free). We then used the tool medaka_variant of ONT's Medaka (version 1.2.2) for variant calling and phasing (ie, resolving haplotypes) of called variants in a multistep procedure. Finally, we used BCFtools (version 1.11)[33] to generate haplotype FASTA sequences for all study samples. To achieve the best accuracy of generated sequences, we validated several sites in repetitive regions by Sanger sequencing.

For downstream analyses, we created an alignment of all 154 haplotype sequences, hereafter referred to as the "analysis sequence alignment." According to standard procedure, we masked out repetitive sequence motives as they have a different underlying evolutionary model with much higher mutation rates than SNVs, which would lead to an overestimation of genetic diversity. Furthermore, sequence quality in such regions, in particular when containing long homopolymers (ie, repetitive stretches of the same nucleotide), is reduced as they still pose a challenge for long-read sequencing technologies,[21] in particular for ONT.[34] The alignment was trimmed to the CDS start and end of the ABO blood group reference transcript NM_020469.3.

Illumina and PacBio HiFi sequencing

For quality validation of obtained ONT sequences, a subset of 12 samples (n = 2 for each *ABO* group; supplemental Table 5), which were *ABO* homozygous at pretyped variants, was additionally sequenced using both long-read PacBio HiFi sequencing (Sequel II system) and short-read Illumina sequencing (MiSeq instrument). Reads from both platforms were mapped against the *ABO*

reference sequence NG_006669.2, followed by a combined variant calling step using the Genome Analysis Toolkit (version 4.1.4.1).[35] Further details on sequencing library preparation and bioinformatics are provided in supplemental Information Section 4. Hereafter, we will refer to this approach as "Illumina/PacBio hybrid approach."

Genetic diversity analyses

To investigate genetic diversity patterns within and between the 6 *ABO* groups, we calculated several diversity statistics based on the analysis sequence alignment using DNAsp (version 6).[36] Detailed information is provided in supplemental Information Section 5.

Phylogenetic analyses

We built a median-joining haplotype network based on the analysis sequence alignment of all 154 *ABO* sequences using PopART (version 1.7).[37] The network was redrawn using 999 iterations. We further constructed a maximum-likelihood phylogenetic tree using IQ-TREE (version 2.1.2)[38] in 2 consecutive steps. We first determined the most suitable nucleotide substitution model with ModelFinder implemented in IQ-TREE, using the –mtree option and –m set to ModelFinder (MF). Model selection was computed 10 times independently using the –run option. The best-fit substitution model according to the Bayesian information criterion was K2P + R2. We then built a maximum-likelihood tree with 1000 standard nonparametric bootstraps and 10 independent runs using IQ-TREE. Following best practice, the tree was rooted with a sequence of the human's closest living relative species, that is, with the central chimpanzee NCBI reference sequence (NC_036888.1; gene ID: 450164), which corresponds to an *A*-like allele.

To detect potential recombination events between *ABO* allele groups, we ran the recombination detection program RDP4.[39] Using 7 different methods implemented in RDP4, we tested for the presence of overall recombination signals in the data set, and investigated where on the gene (ie, breakpoints) and between which haplotypes recombination events might have occurred (details see supplemental Information Section 6). Only events depicted by at least 5 different methods were considered reliable. To investigate the influence of the recombination events on the topology of the phylogenetic tree, we additionally reconstructed a tree excluding the recombination sites found by RDP4 from the sequences. The tree was built using the same approach as outlined above.

Validation of putative *ABO*A1*-diagnostic variants in a multiethnic cohort

To study the diagnostic accuracy of our discovered putatively *ABO*A1*-specific variants in a larger and ethnically more diverse cohort, we used whole-genome sequencing data of 4872 individuals participating in the MESA project (Multi-Ethnic Study of Atherosclerosis).[40-42] Because of missing serological information, *ABO*A1* alleles were predicted by the absence of diagnostic variants of complimentary alleles. Hence, we extracted the allele-defining variants for *ABO*A2* (c.1061delC), *B* (c.803G>C), *O.01* (c.261delG), and *O.02* (c.802G>A) as well as the 4 *ABO*A1* candidate variants from the whole-genome sequencing data, and deduced estimates for *ABO*A1* sensitivity and specificity. Further details are provided in supplemental Information Section 7.

Results

ABO haplotypes and ONT sequencing

We successfully resolved both the maternal and paternal full-length (~23.6 kb) *ABO* haplotype sequences for all 77 samples. A detailed list of sequenced *ABO* haplotypes including NCBI GenBank accession numbers (OM283861–OM284014) is provided in supplemental Table 2. Sequence alignments of all 154 haplotypes are available from the Dryad Digital Repository (https://doi.org/10.5061/dryad.q573n5tkj).

The median depth of nanopore sequencing was 1454x per LR-PCR amplicon. Detailed sequencing depth per sample and amplicon is provided in supplemental Table 1. Nanopore sequence data proved highly accurate with 100% agreement to the Illumina/PacBio hybrid approach data based on the analysis sequence alignment.

Exonic variation in *ABO* sequences

An overview of genetic variation in *ABO* exons among haplotypes is given in Table 2. Details for each sequence are provided in supplemental Table 2. According to their exonic variation, most sequences (n = 135) corresponded to main ISBT alleles,[27] that is *ABO*A1.01*, *A2.01*, *B.01*, *O.01.01, O.01.02*, and *O.02.01*. Intronic variation is currently largely ignored in ISBT nomenclature.[27] We also observed less common alleles within *ABO* groups, that is, *ABO*O.01.26* (n = 2), *ABO*O.01.67* (n = 4), *ABO*O.01.68* (n = 1), *ABO*O.01.75* (n = 2), *ABO*O.02.02* (n = 4), *ABO*O.02.03* (n = 1), and *ABO*O.02.04* (n = 1). These alleles differed by only 1 or 2 exonic SNVs from their *ABO* group background allele (Table 2). Because they still belong to one of the 6 *ABO* groups defined in this study, alleles were kept in the sample set for genetic diversity and phylogenetic analyses.

Four sequenced haplotypes were not yet listed in the official ABO (ISBT 001) blood group allele table (version 1.1 171023).[27] Three of them had silent mutations that did not alter the protein. The fourth one represented a novel *ABO*B.01* allele (sample s07_h2) with an additional SNV c.122G>A in exon 3, replacing serine by asparagine. All SNVs were verified by Sanger sequencing.

Genetic diversity patterns among and within *ABO* groups

An alignment of all *ABO* haplotype sequences revealed yet undescribed distinct sequence patterns of each *ABO* allele group across the entire gene locus. Figure 2 shows this pattern for a random subset of 6 haplotype sequences per *ABO* group. Specific sequence patterns were found in both exonic and intronic regions.

In total, we found 230 SNVs and 16 sites with insertions/deletions (indels) among all the sequences (Table 3; supplemental Table 6). The 154 haplotype sequences represented 47 unique haplotypes, with on average 66.4 nucleotide differences between 2 haplotypes. Genetic diversity was much higher between *ABO* groups than within groups (Tables 3 and 4). Within *ABO* groups (Table 3; supplemental Table 6), genetic diversity was particularly low for *ABO*A1* and *B*. The group of *ABO*O.01.01* showed the highest within-group diversity. This appeared to be linked to deep within-group substructure into 2 phylogenetic entities (Figures 3 and 4), which is inflating diversity measurements. This effect is also present

Table 2. List of genetic variation in *ABO* exons among the 154 haplotype sequences

Phenotype	Allele name	Nucleotide change†	Exon	Predicted amino acid change	No. of sequences	Comments
A_1	***ABO*A1.01***				39	ISBT reference allele
A_2	***ABO*A2.01***	c.467C>T;	7	p.Pro156Leu;	20	
		c.1061delC	7	p.Pro354Argfs*23		
B	***ABO*B.01***	c.297A>G;	6		19	
		c.526C>G;	7	p.Arg176Gly;		
		c.657C>T;	7			
		c.703G>A;	7	p.Gly235Ser;		
		c.796C>A;	7	p.Leu266Met;		
		c.803G>C;	7	p.Gly268Ala		
		c.930G>A	7			
O	***ABO*O.01.01***	c.261delG	6	p.Thr88Profs*31	25	
O	***ABO*O.01.26***	c.261delG	6	p.Thr88Profs*31	2	*ABO*O.01.01* background with additional c.768C>A
		c.768C>A	7			
O	***ABO*O.01.02***	c.106G>T;	3	p.Val36Phe;	22	
		c.188G>A;	4	p.Arg63His;		
		c.189C>T;	4			
		c.220C>T;	5	p.Pro74Ser;		
		c.261delG;	6	p.Thr88Profs*31		
		c.297A>G;	6			
		c.646T>A;	7			
		c.681G>A;	7			
		c.771C>T;	7			
		c.829G>A	7			
O	***ABO*O.01.67***	c.103G>A;	3	p.Gly35Arg;	4	*ABO*O.01.02* background with additional c.103G>A
		c.106G>T;	3	p.Val36Phe;		
		c.188G>A;	4	p.Arg63His;		
		c.189C>T;	4			
		c.220C>T;	5	p.Pro74Ser;		
		c.261delG;	6	p.Thr88Profs*31		
		c.297A>G;	6			
		c.646T>A;	7			
		c.681G>A;	7			
		c.771C>T;	7			
		c.829G>A	7			
O	***ABO*O.01.68***	c.106G>T;	3	p.Val36Phe;	1	*ABO*O.01.02* background without c.220C>T
		c.188G>A;	4	p.Arg63His;		
		c.189C>T;	4			
		c.261delG;	6	p.Thr88Profs*31		
		c.297A>G;	6			
		c.646T>A;	7			
		c.681G>A;	7			
		c.771C>T;	7			
		c.829G>A	7			

Details on all sequences are provided in supplemental Table 2. Four observed alleles are not yet listed in the official ABO (ISBT 001) blood group allele table (version 1.1 171023).[27] The corresponding novel nucleotide changes are highlighted in bold.

†Positions of nucleotide changes relate to reference transcript NM_020469.3.

Table 2 (continued)

Phenotype	Allele name	Nucleotide change†	Exon	Predicted amino acid change	No. of sequences	Comments
O	***ABO*O.01.75***	c.106G>T;	3	p.Val36Phe;	2	*ABO*O.01.02* background with additional c.542G>A
		c.188G>A;	4	p.Arg63His;		
		c.189C>T;	4			
		c.220C>T;	5	p.Pro74Ser;		
		c.261delG;	6	p.Thr88Profs*31		
		c.297A>G;	6			
		c.542G>A;	7			
		c.646T>A;	7			
		c.681G>A;	7			
		c.771C>T;	7			
		c.829G>A	7			
O	***ABO*O.02.01***	c.53G>T;	2	p.Arg18Leu;	10	
		c.220C>T;	5	p.Pro74Ser;		
		c.297A>G;	6			
		c.526C>G;	7	p.Arg176Gly;		
		c.802G>A	7	p.Gly268Arg		
O	***ABO*O.02.02***	c.53G>T;	2	p.Arg18Leu;	4	*ABO*O.02* background with additional c.649C>T and c.689G>A
		c.220C>T;	5	p.Pro74Ser;		
		c.297A>G;	6			
		c.526C>G;	7	p.Arg176Gly;		
		c.649C>T;	7	p.Arg217Cys;		
		c.689G>A;	7	p.Gly230Asp;		
		c.802G>A	7	p.Gly268Arg		
O	***ABO*O.02.03***	c.53G>T;	2	p.Arg18Leu;	1	*ABO*O.02* background with additional c.689G>A
		c.220C>T;	5	p.Pro74Ser;		
		c.297A>G;	6			
		c.526C>G;	7	p.Arg176Gly;		
		c.689G>A;	7	p.Gly230Asp;		
		c.802G>A	7	p.Gly268Arg		
O	***ABO*O.02.04***	c.53G>T;	2	p.Arg18Leu;	1	*ABO*O.02* background with additional c.488C>T
		c.220C>T;	5	p.Pro74Ser;		
		c.297A>G;	6			
		c.488C>T;	7	p.Thr163Met;		
		c.526C>G;	7	p.Arg176Gly;		
		c.802G>A	7	p.Gly268Arg		
New alleles not listed in the ABO (ISBT 001) blood group allele table (version 1.1 171023)						
A_2	***ABO*A2.01(c.1032G>A)***	c.467C>T;	7	p.Pro156Leu;	1	*ABO*A2.01* background with additional c.1032G>A
		c.1032G>A;	7			
		c.1061delC;	7	p.Pro354Argfs*23		
B	***ABO*B.01(c.122G>A)***	**c.122G>A**;	3	**p.Ser41Asn**	1	*ABO*B.01* background with additional c.122G>A
		c.297A>G;	6			
		c.526C>G;	7	p.Arg176Gly;		
		c.657C>T;	7			
		c.703G>A;	7	p.Gly235Ser;		
		c.796C>A;	7	p.Leu266Met;		

Details on all sequences are provided in supplemental Table 2. Four observed alleles are not yet listed in the official ABO (ISBT 001) blood group allele table (version 1.1 171023).[27] The corresponding novel nucleotide changes are highlighted in bold.

†Positions of nucleotide changes relate to reference transcript NM_020469.3.

Table 2 (continued)

Phenotype	Allele name	Nucleotide change†	Exon	Predicted amino acid change	No. of sequences	Comments
		c.803G>C;	7	p.Gly268Ala		
		c.930G>A	7			
O	***ABO*O.01.02(c.6C>T)***	**c.6C>T**;	1		1	*ABO*O.01.02* background with additional c.6C>T
		c.106G>T;	3	p.Val36Phe;		
		c.188G>A;	4	p.Arg63His;		
		c.189C>T;	4			
		c.220C>T;	5	p.Pro74Ser;		
		c.261delG;	6	p.Thr88Profs*31		
		c.297A>G;	6			
		c.646T>A;	7			
		c.681G>A;	7			
		c.771C>T;	7			
		c.829G>A	7			
O	***ABO*O.01.68(c.595C>T)***	c.106G>T;	3	p.Val36Phe;	1	*ABO*O.01.68* background with additional c.595C>T
		c.188G>A;	4	p.Arg63His;		
		c.189C>T;	4			
		c.261delG;	6	p.Thr88Profs*31		
		c.297A>G;	6			
		c.595C>T;	7			
		c.646T>A;	7			
		c.681G>A;	7			
		c.771C>T;	7			
		c.829G>A	7			

Details on all sequences are provided in supplemental Table 2. Four observed alleles are not yet listed in the official ABO (ISBT 001) blood group allele table (version 1.1 171023).[27] The corresponding novel nucleotide changes are highlighted in bold.

†Positions of nucleotide changes relate to reference transcript NM_020469.3.

when looking at all *ABO*O.01* haplotypes combined without separating the 2 subgroups, *ABO*O.01.01* and *ABO*O.01.02*, as we have done in this study (Table 3).

Most of the nucleotide differences found between *ABO* groups were fixed, that is, SNVs for which 1 group has 1 allele and the other group the other allele (Table 4). The 2 *ABO*O.01* subgroups (ie, *ABO*O.01.01* and *ABO*O.01.02*) had on average 84 nucleotide differences between 2 random sequences, of which 73 were fixed between *ABO*O.01* subgroups. This was in the same order as, for example, comparing *ABO*O.01.01* and *ABO*B,* highlighting the deep divergence of the 2 *ABO*O.01* subgroups.

We also observed 4 fixed SNVs, one 1-bp and one 12-bp indel in all *ABO*A1.01* haplotypes sequenced in this study compared with the ISBT reference sequence NG_006669.2 (LRG_792) for the ABO blood group, also an *ABO*A1.01* allele. This reference sequence is an artificially assembled *ABO*A1.01* allele on an *ABO*O.01* background. Hence, it may be possible that the observed differences are relicts to be corrected of the old *ABO*O.01* background sequence in the current ISBT reference sequence.

Phylogenetic analyses

In line with the observed genetic diversity patterns, phylogenetic analyses revealed distinct clades (ie, phylogenetic groups sharing a common ancestor) for all *ABO* groups. The rooted phylogenetic tree constructed using the full-length *ABO* sequences showed well-supported monophyletic lineages (defined groups containing all descendants of the respective ancestor) for *ABO*A1*, *A2*, *B*, *O.01.02*, and *O.02* (Figure 3). The *ABO*O.01.01* haplotypes were paraphyletic (originated from the same ancestor but not all descendants from this ancestor were contained in 1 group) and split into 2 subgroups (g1 and g2), which was also observed in the haplotype network (Figure 4). All SNVs separating the 2 subgroups were located in intronic regions.

The 2 major O-phenotype groups (ie, *ABO*O.01* and *ABO*O.02*) were paraphyletic to each other with deep (ie, ancient) splits, showing that these groups are not closely related in evolutionary terms, despite sharing a null phenotype. The *ABO*O.02* lineage split off first from all other human *ABO* haplotypes. The group *ABO*O.01.01* appeared evolutionary closer to the cluster of *ABO*A1*, *A2*, and *B* than *ABO*O.01.02*.

The haplotype network inferred from all 154 *ABO* sequences (Figure 4), which is a different way to investigate evolutionary relationships among sequences, was in congruence with the phylogenetic tree.

Our analysis for detecting overall patterns of recombination among *ABO* haplotypes did find strong evidence for recombination events

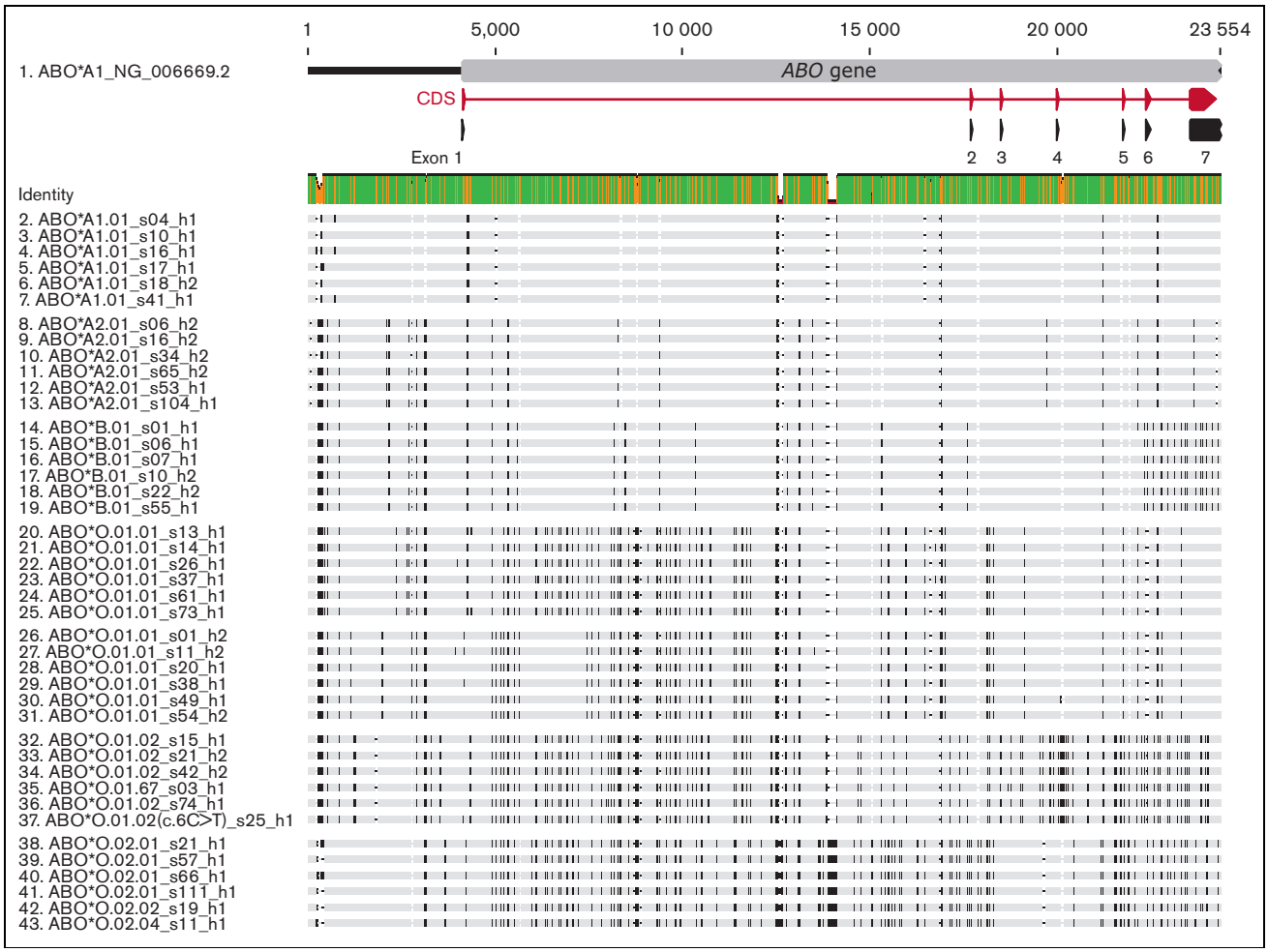

Figure 2. Alignment of a random subset of haplotype sequences highlighting distinct sequence patterns of *ABO* groups. For each *ABO* group, 6 haplotype sequences were randomly picked. For the subgroup *ABO*O.01.01*, we show 6 sequences for both subgroups (g1 and g2) observed in the phylogenetic tree (Figure 3) and haplotype network (Figure 4). Black bars on the haplotype sequences highlight positions that are different to the *ABO* reference sequence (NG_006669.2). The identity graph (Identity) above the sequences indicates the mean pairwise identity over all sequence pairs by gene position; green represents 100% identity, orange identity between 30% and 99%, and red identity <30%. The *ABO* gene structure is provided at the top of the graph for orientation purpose.

($P < 10^{-5}$). Five recombination events (supplemental Information Section 6; supplemental Figure 2) were identified by at least 5 different methods and, therefore, wereregarded as credible. One event was detected by only 2 methods and, therefore, was discarded from the analysis. Among the 5 retained events, 1 involved recombination between ancestral sequences of the contemporary *ABO*A2* and *ABO*O.02* alleles. The 4 other recombination events all resulted in null alleles.

Table 3. Overview of genetic diversity among the 154 *ABO* haplotype sequences

		Within *ABO* groups						
	All samples	***ABO*A1***	***ABO*A2***	***ABO*B***	***ABO*O.01***	***ABO*O.01.01***	***ABO*O.01.02***	***ABO*O.02***
No. of sequences	154	39	21	20	58	27	31	16
No. of SNVs	230	7	7	6	110	23	14	18
No. of indels	16	0	0	0	2	0	0	1
No. of unique haplotypes	47	5	5	6	25	14	11	6
Average no. of differences between haplotypes†	66.4	0.4	1.1	0.8	44.8	8.5	1.9	1.8

For comparison, statistics are also provided for *ABO*O.01* without separating the 2 subgroups *ABO*O.01.01* and *ABO*O.01.02*. An extended version of the table with more detailed statistics is provided in the supplemental Information (supplemental Table 6).

†Average number of nucleotide differences between 2 sequences within the ABO group.[43]

Table 4. High degree of fixed differences between *ABO* groups

	*ABO*A2*	*ABO*B*	*ABO*O.01.01*	*ABO*O.01.02*	*ABO*O.02*
ABO*A1	8 (6)	33 (32)	60 (52)	112 (107)	114 (109)
ABO*A2		30 (28)	60 (51)	111 (107)	114 (108)
ABO*B			83 (75)	122 (118)	109 (104)
ABO*O.01.01				84 (73)	94 (82)
ABO*O.01.02					114 (110)

Listed are average number of actual nucleotide differences of 2 sequences between *ABO* groups. The number in parentheses shows the number of nucleotide differences that were fixed between groups (ie, sites for which one group has one allele and the other group the other allele).

The inferred phylogenetic tree excluding the identified recombination regions from the sequences (supplemental Figure 3) showed a very similar topology to the tree constructed using the full-length haplotype sequences (Figure 3). The only difference between both trees was in the relationship between the 2 *ABO*O1* subgroups, *ABO*O.01.01* and *ABO*O.01.02*. Although these 2 subgroups were paraphyletic in the tree constructed on the entire alignment, they were monophyletic in the tree constructed excluding recombinant regions.

Putative *ABO*A1*-diagnostic variants and their validation in a multiethnic cohort

Strikingly, among all the variants that were fixed between *ABO* groups, we identified 4 variants in intron 1 that were exclusively present in all *ABO*A1* haplotypes. These 4 variants (Figure 1) were composed of the SNVs rs532436 (NG_006669.2:g.5801T>C), rs507666 (g.6232T>C), and rs2519093 (g.13759A>G), as well as of the dinucleotide variant rs1554760445, which merges the SNV rs115478735 (g.5920T>A) with the adjacent indel rs8176643 (g.5921delG).

As the major limitation of our haplotype collection is its sole representation of European ancestry, we validated the diagnostic potential of the 4 variants in the multiethnic MESA cohort.[40-42] Detailed results are provided in supplemental Information Section 7 and supplemental Table 7. In short, *ABO*A1*-specificities for the 3 SNV-based variants were 99.41% (rs532436), 99.44% (rs507666), and 99.60% (rs2519093), whereas sensitivity was estimated between 97.55% (rs507666) and 97.99% (rs2519093). These 3 variants showed very high linkage disequilibrium (pairwise $r^2 > 0.97$). Consequently, combining them did not increase specificity and only marginally increased sensitivity (to 98.43%). Specificity for the dinucleotide variant rs1554760445 was high (99.72%), but this variant was much rarer in the cohort (found in 21.7% of individuals) than the 3 SNVs (29.5% to 29.7%). Accordingly, the sensitivity was substantially reduced (71.93%). Limiting the analysis to samples of European ancestry, sensitivity increased to the scale of the other variants (97.10%), implying that rs1554760445 may specifically tag *ABO*A1.01*, the predominant *ABO*A1* allele subgroup in Europe. Although the variant was indeed completely absent in *ABO*A1.02* predicted alleles (n = 319), *ABO*A1.01*-sensitivity across the whole cohort reached only 90.0% owing to the absence of the variant in ~100 individuals of African descent with predicted *ABO*A1.01* alleles.

Discussion

Taking advantage of third-generation sequencing, we have generated a comprehensive collection of full-length haplotype sequences (n = 154) for all 6 main *ABO* allele groups (*ABO*A1*, *A2*, *B*, *O.01.01*, *O.01.02*, and *O.02*). Together, these groups cover 99.9% of the genetic diversity of *ABO* in Switzerland.

Characteristic sequence patterns among *ABO* groups

Our haplotype collection uncovered hitherto unknown distinct sequence patterns among *ABO* groups, which had not yet been unveiled because of the very few complete human *ABO* gene sequences available so far. As for almost all blood group genes, sequencing efforts of *ABO* have largely been limited to exons. In particular, the large intron 1 (~13.0 kb) has rarely been successfully sequenced owing to technical difficulties,[26] although it is known to harbor major genetic variation of interest.[44,45] Our established generic LR-PCRs of the *ABO* gene will facilitate haplotype-based sequencing of *ABO* in molecular diagnostics in transfusion and transplantation medicine. The distinct sequence patterns of each *ABO* group allow unambiguous assignment of sequences to a particular *ABO* group. This will, for instance, greatly support identification and exact breakpoint determination of *ABO* hybrid alleles and other structural variation in routine diagnostics; information that ultimately helps resolving serotype-genotype discrepancies.

Phylogenetic analyses and genetic diversity

In agreement with the distinct sequence patterns, all *ABO* groups formed separate evolutionary clades, as revealed congruently by the phylogenetic tree and haplotype network. Our data showed that *ABO*O* groups are not closely related in evolutionary terms, even though they share the null phenotype. Despite the presence of several recombination events between null alleles, we observed in the phylogenetic tree deep splits separating the *ABO*O* clades. The split between *ABO*O.02* and *ABO*O.01* alleles was even more pronounced in the phylogeny excluding the identified recombinant regions (supplemental Figure 3). Contrary to the ancestral *ABO*A* and *ABO*B* alleles, which originated early in the evolution of animals (ie, after speciation between fish and amphibian lineages), null alleles are most often species-specific.[46,47] For *ABO*O.01* and *ABO*O.02*, it has recently been hypothesized that they originated from independent Neanderthal to modern human introgression events.[48]

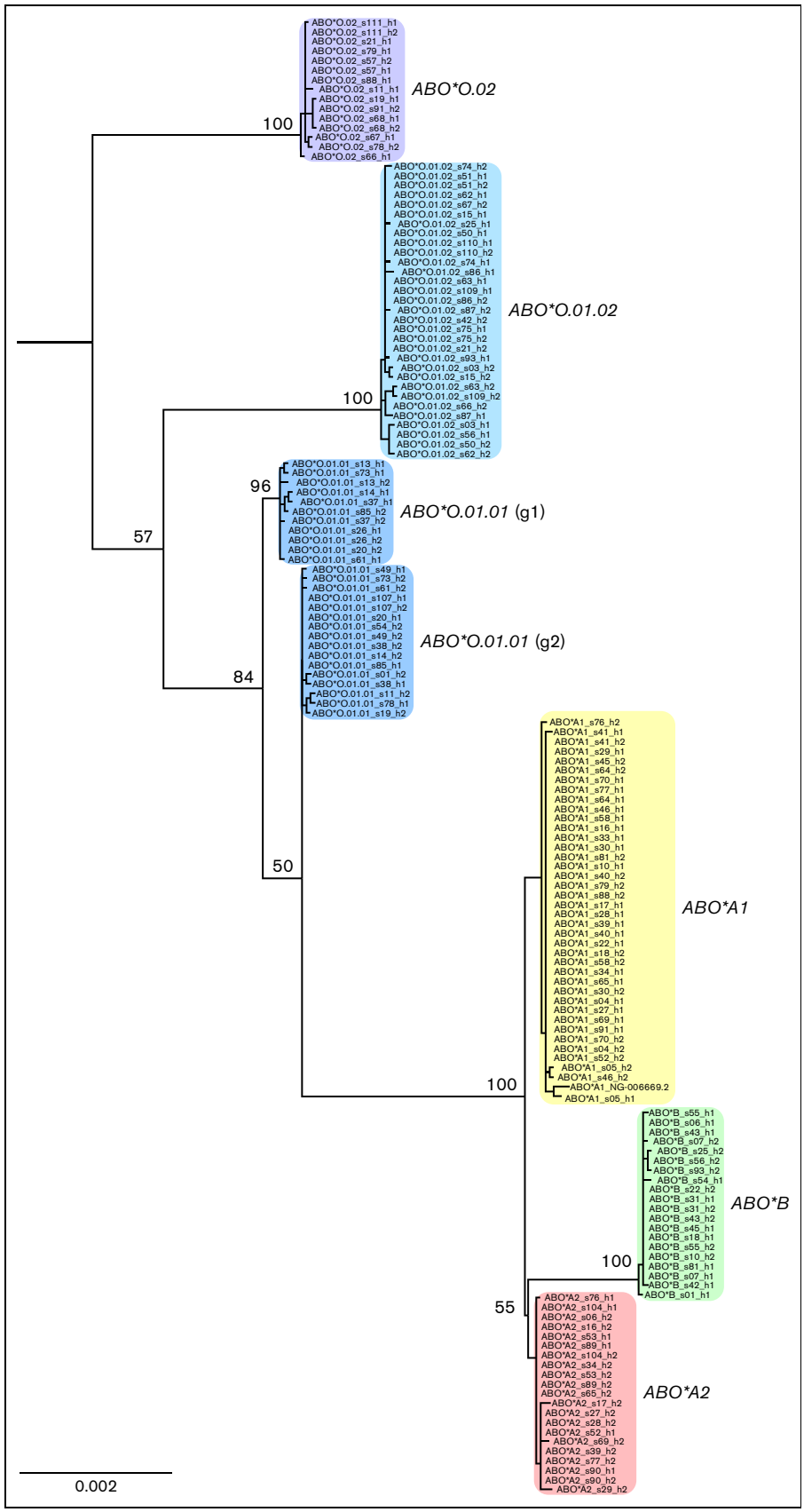

Figure 3. Maximum-likelihood phylogenetic tree based on the entire *ABO* gene locus. All *ABO* groups form distinct evolutionary clades with an additional split of *ABO*O.01.01* into 2 subgroups (g1 and g2). Bootstrap support is provided for main branching points. The tree was rooted with central chimpanzee sequence (not shown).

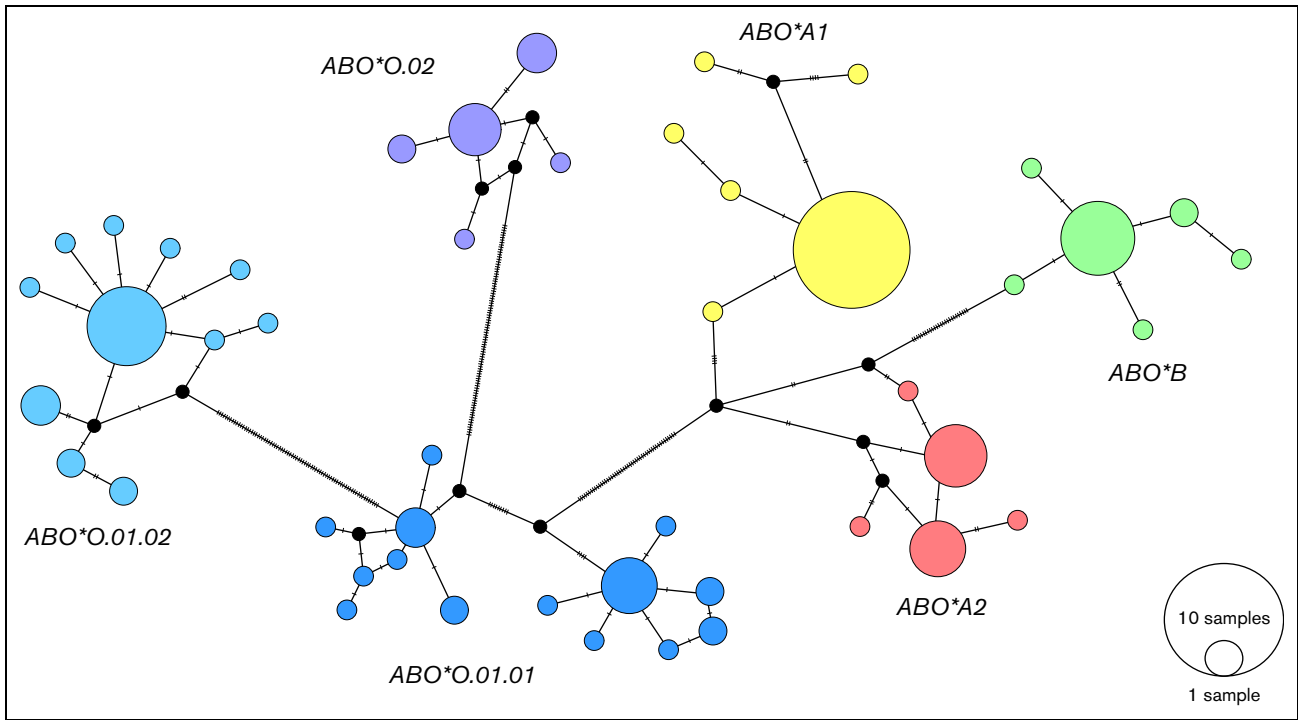

Figure 4. Median-joining haplotype network inferred from all 154 *ABO* sequences. Phylogenetic network showing the evolutionary relationships among *ABO* haplotypes. Each circle represents a unique haplotype with the size being proportional to the number of sequences represented. Haplotype circles are colored according to the *ABO* allele groups; black dots represent missing intermediate haplotypes (ie, unsampled, likely ancient haplotypes). Mutational steps between haplotypes are displayed as hatch marks along the connection lines. The length of the connection lines is not scaled by phylogenetic distance.

The early splits of the *ABO*O* lineages in the rooted phylogenetic tree, with *ABO*O.02* splitting off first from all other human *ABO* haplotypes, seems surprising considering that usually loss-of-function rather than gain-of-function changes are observed along evolutionary lineages. Kitano et al[12] raised the hypothesis that the supposedly ancestral *A*-like allele[13,49] once became extinct in the human lineage, and that the present-day *A1* allele was resurrected by a recombination event between *B* and *O.01* alleles around 260 000 years ago. They hypothesized a breakpoint region around the *ABO*O.01*-specific deletion, c.261delG. While testing for recombination events, we found evidence of recombination in the same gene region as described by Kitano et al.[12] Our analysis, however, identified *ABO*B* as being the resulting allele from the recombination between ancestral alleles of *ABO*A2* and *ABO*O.O2*. Disentangling the parental from recombinant sequences is very challenging, particularly as analyses are based on present-day alleles and lost ancestral alleles can only be inferred. Also, outcomes from such analyses are highly dependent on the alleles contained in the data set. Therefore, we advocate that our results are not antagonistic to Kitano et al[12] findings, but rather provide supplementary evidence that recombination has happened in this gene region between *ABO*A*, *ABO*B*, and *ABO*O* ancestral alleles.

The complex evolutionary history of *ABO* is also highlighted by the high genetic diversity found across *ABO* groups. The unusually high diversity at *ABO* is likely maintained by balancing selection,[11] a form of adaptation that maintains diversity in a species in the face of random genetic drift.[50] *ABO* genetic variation has been associated with predispositions to a large number of diseases (reviewed in Liumbruno et al),[51] including infectious diseases,[11,25,52,53] gastric[54] and pancreatic cancers,[55] and cardiovascular diseases.[56] The phenotypes under natural selection, to which the observed genetic diversity contributes, however, remain less clear.[57] Overall, although there has been considerable research on the complex evolutionary history of *ABO*,[11-13,25,46,58-60] many aspects remain obscure and need further in-depth studies, which will hopefully be supported by our novel *ABO* haplotype collection and workflow.

Our results of high, distinct genetic diversity among *ABO* groups exemplifies the importance of having comprehensive haplotype sequence collections for designing primers for genotyping assays and sequencing in diagnostics. Ignored genetic variation at primer-binding sites may lead to unnoticed allelic dropout (ie, only 1 of the 2 alleles is detected), and thus, spurious homozygosity.[61] This may, for instance, have severe clinical consequences in transfusion medicine owing to potentially lethal incompatible transfusions and alloimmunization reactions, in particular, in cases where serological confirmation is not possible. Therefore, we deem it important to establish similar collections as the one presented in this study for other highly diverse blood group systems (eg, RhD/RhCE and MNS).

Nanopore sequencing

Our collection of *ABO* haplotypes could be generated thanks to the technical advances of long-read sequencing technologies. ONT produced consensus sequences equivalent in accuracy to

the consensus sequences obtained from the combination of Illumina and PacBio HiFi sequencing, except in highly repetitive sequence motifs, which we finally adopted from the Illumina/PacBio hybrid approach data. This is attributed to the major recent developments of ONT's sequencing technology and machine-learning algorithms.[62-64]

Putative diagnostic *ABO*A1* variants

Thanks to our full-length *ABO* haplotype data encompassing all main *ABO* groups, 4 variants in intron 1 could be uncovered with putative diagnostic specificity for *ABO*A1*. Such diagnostic markers are currently lacking, although *ABO*A1* is the official ISBT reference allele for the ABO blood group. In molecular diagnostics, *ABO*A1* can only be inferred indirectly by the method of exclusion, that is, by targeting causative variants defining *ABO*A2*, *B*, *O.01*, and *O.02*.[65] The A_1 antigen can be solely determined serologically using an anti-A_1 lectin, which is rarely done routinely.

*ABO*A1*-candidate variants have been reported as lead SNVs in large genetic association studies on cardiovascular diseases[66,67] and inflammatory markers.[68] Although these phenotypes are well known to be influenced by ABO blood group,[51] rigorous analyses linking the lead SNVs to *ABO* allele groups have so far not been undertaken.

In a validation approach based on whole-genome data of 4872 individuals of a multiethnic cohort, we observed high diagnostic potential for 3 of the 4 *ABO*A1*-candidate variants across ethnicities. Hence, our haplotype analyses provided support for previous hypotheses raised in the context of genetic association and risk score studies of rs507666[68] and rs2519093[69] being surrogates for *ABO*A1*. Furthermore, we found evidence that the compound *ABO*A1*-candidate variant rs1554760445 specifically tags the *ABO*A1.01* allele (instead of generically also *ABO*A1.02*), unless in populations of African descent.

Importantly, sensitivity and specificity values computed in this study are overall very conservative as estimated solely at the allele level. Phenotype prediction from genetic data in a diagnostic setting would, however, first focus on identifying the number of *ABO*O* alleles (as they impair the allele's function) and only subsequently incorporate variants linked to non-*ABO*O* alleles.[4,70] Such a procedure would significantly increase specificity to over 99.89%, as our candidate variants coincided most frequently with *ABO*O* alleles and as few as only 3 times with *ABO*A2* or *ABO*B* alleles (in case of rs2519093). Notably, accuracy estimates may generally be an underestimation, given that some of the incongruences may be attributed to unrecognized hybrid alleles, lack of structural variation information, improper statistical phasing, or sequencing errors in the MESA data.

In summary, the dinucleotide candidate variant rs1554760445 showed promising diagnostic potential to specifically tag the *ABO*A1.01* allele outside Africa, whereas the 3 SNV-based *ABO*A1*-candidates performed very well in generally tagging *ABO*A1* alleles across ethnicities. We are currently validating diagnostic *ABO*A1* specificity and sensitivity of the 4 variants in more detail by using a large sample set of blood donor populations with available serological data around the world. If the variants are confirmed to be diagnostically accurate, they will finally allow for direct genetic typing of *ABO*A1* in routine molecular diagnostics.

Conclusions

The discovery of intronic markers that accurately represent a main allele (*ABO*A1*) in the most relevant blood group exemplifies the importance of including non-exonic regions in the definition of reference sequences. Alternative haplotype-resolving technologies, such as sequencing of complementary DNA or direct RNA sequencing are incomprehensive, and, therefore, inappropriate strategies for collecting reference haplotype sequences.

Overall, our long-read sequencing strategy proved powerful for generating a comprehensive haplotype collection for the clinically most important blood group system, ABO. As a proof of principle, our strategy holds promise for generating similar collections for other blood group systems. Our publicly available haplotype collection revealed new insights into genetic diversity patterns at *ABO*, including uncovering putatively *ABO*A1*-diagnostic variants, and will serve as a valuable reference resource for molecular diagnostic analyses of *ABO* and future studies of evolutionary history.

Acknowledgments

The authors thank all laboratory staff at the Stefan Morsch Foundation (Germany) and the Institute of Clinical Molecular Biology of the Christian Albrechts University of Kiel (Germany) involved in providing the Illumina/PacBio sequencing data. The authors are grateful to Valentina Donà for contributing sequencing know-how. Furthermore, the authors are indebted to all individuals involved in the Multi-Ethnic Study of Atherosclerosis, in particular, Jerome Rotter (Lundquist Institute), Stephen S. Rich (University of Virginia), and W. Craig Johnson (University of Washington). Finally, the authors thank the anonymous reviewers for carefully reviewing the manuscript.

This work was financially supported by the Blood Transfusion Service Zurich, Swiss Red Cross (Switzerland), the Stefan Morsch Foundation, and the Institute of Clinical Molecular Biology of the Christian Albrechts University of Kiel.

Authorship

Contribution: W.P., C.G., B.M.F., and M.P.M.-G. initiated the study and contributed ideas; M.P.M.-G. conceived and coordinated the study; B.M.F., C.G., and W.P. provided their input; M.P.M.-G., M.G., and G.A.T. designed the study and experiments; S.M., N.T., E.G., S.S., Y.M., K.N., J.G., C.G., and M.P.M.-G. contributed samples and MALDI-TOF mass spectrometry genotype data; G.A.T., M.G., and M.P.M.-G. performed experiments and analyzed data; A.-L.G., M.S., and W.P. contributed to long-range polymerase chain reaction design; W.P., M.W., A.-L.G., J.F., Y.B., P.T., M.S., and A.F. provided Illumina/PacBio sequencing data; M.P.M.-G., G.A.T., and M.G. wrote the manuscript; M.W. and W.P. contributed to the supplemental information; and all authors commented on the manuscript and approved the final version.

Conflict-of-interest disclosure: C.G. acts as a consultant for inno-train GmbH, Kronberg im Taunus, Germany. The remaining authors declare no competing financial interests.

ORCID profiles: M.G., 0000-0001-8574-9640; G.A.T., 0000-0003-4436-3455; M.W., 0000-0003-1103-4196; E.G., 0000-0001-5183-6527; J.F., 0000-0002-7631-9355; Y.B., 0000-0001-6261-6853.

Correspondence: Maja P. Mattle-Greminger, Department of Research and Development, Blood Transfusion Service Zurich, Swiss Red Cross, Rütistrasse 19, 8952 Schlieren, Switzerland; email: m.mattle@zhbsd.ch.

References

1. Fichou Y, Berlivet I, Richard G, Tournamille C, Castilho L, Férec C. Defining blood group gene reference alleles by long-read sequencing: proof of concept in the ACKR1 gene encoding the duffy antigens. *Transfus Med Hemotherapy*. 2020;47(1):23-32.
2. Tounsi WA, Madgett TE, Avent ND. Complete RHD next-generation sequencing: establishment of reference RHD alleles. *Blood Adv*. 2018;2(20):2713-2723.
3. Möller M, Jöud M, Storry JR, Olsson ML. Erythrogene: a database for in-depth analysis of the extensive variation in 36 blood group systems in the 1000 Genomes Project. *Blood Adv*. 2016;1(3):240-249.
4. Lane WJ, Westhoff CM, Gleadall NS, et al. Automated typing of red blood cell and platelet antigens: a whole-genome sequencing study. *Lancet Haematol*. 2018;5(6):e241-e251.
5. Wheeler MM, Johnsen JM. The role of genomics in transfusion medicine. *Curr Opin Hematol*. 2018;25(6):509-515.
6. Fichou Y, Audrézet MP, Guéguen P, Le Maréchal C, Férec C. Next-generation sequencing is a credible strategy for blood group genotyping. *Br J Haematol*. 2014;167(4):554-562.
7. Gassner C, Olsson ML, Lane WJ, Hyland CA. Novel or not? Reference alleles, genes, and genomes to unmask the true nature of the ABO*AW.10 allele associated with weak A phenotype. *Transfusion*. 2022;62(4):721-724.
8. Gleadall NS, Veldhuisen B, Gollub J, et al. Development and validation of a universal blood donor genotyping platform: a multinational prospective study. *Blood Adv*. 2020;4(15):3495-3506.
9. Gassner C, Meyer S, Frey BM, Vollmert C. Matrix-assisted laser desorption/ionisation, time-of-flight mass spectrometry-based blood group genotyping–the alternative approach. *Transfus Med Rev*. 2013;27(1):2-9.
10. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*. 2016;48(10):1279-1283.
11. Ségurel L, Gao Z, Przeworski M. Ancestry runs deeper than blood: the evolutionary history of ABO points to cryptic variation of functional importance. *Bioessays*. 2013;35(10):862-867.
12. Kitano T, Blancher A, Saitou N. The functional A allele was resurrected via recombination in the human ABO blood group gene. *Mol Biol Evol*. 2012;29(7):1791-1796.
13. Calafell F, Roubinet F, Ramírez-Soriano A, Saitou N, Bertranpetit J, Blancher A. Evolutionary dynamics of the human ABO gene. *Hum Genet*. 2008;124(2):123-135.
14. Lang K, Wagner I, Schöne B, et al. ABO allele-level frequency estimation based on population-scale genotyping by next generation sequencing. *BMC Genom*. 2016;17(1):374.
15. Srivastava K, Almarry NS, Flegel WA. Genetic variation of the whole ICAM4 gene in Caucasians and African Americans. *Transfusion*. 2014;54(9):2315-2324.
16. Körmöczi GF, Scharberg EA, Gassner C. A novel KEL* 1, 3 allele with weak Kell antigen expression confirming the cis-modifier effect of KEL3. *Transfusion*. 2009;49(4):733-739.
17. Vergara C, Parker MM, Franco L, et al. Genotype imputation performance of three reference panels using African ancestry individuals. *Hum Genet*. 2018;137(4):281-292.
18. The International Society of Blood Transfusion. Table of blood group systems; v.10.0, 2021. Accessed 15 July 2021. https://www.isbtweb.org/resource/tableofbloodgroupsystems.html
19. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434-443.
20. Taliun D, Harris DN, Kessler MD, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*. 2021;590(7845):290-299.
21. Wenger AM, Peluso P, Rowell WJ, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*. 2019;37(10):1155-1162.
22. Jain M, Koren S, Miga KH, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol*. 2018;36(4):338-345.
23. Srivastava K, Khil PP, Sippert E, et al. ACKR1 alleles at 5.6 kb in a well-characterized renewable US Food and Drug Administration (FDA) reference panel for standardization of blood group genotyping. *J Mol Diagn*. 2020;22(10):1272-1279.
24. Landsteiner K. Über Agglutinationserscheinungen Normalen Menschlichen Blutes [On agglutination phenomena of normal human blood]. *Wiener Klinische Wochenschrift*. 1901;14:1132-1134.
25. Storry J, Olsson ML. The ABO blood group system revisited: a review and update. *Immunohematol*. 2009;25(2):48-59.

26. Wu PC, Lin YH, Tsai LF, Chen MH, Chen PL, Pai SC. ABO genotyping with next-generation sequencing to resolve heterogeneity in donors with serology discrepancies. *Transfusion*. 2018;58(9):2232-2242.
27. The International Society of Blood Transfusion. Names for ABO (ISBT 001) Blood Group Allele [table]; v1.1, 17 October 2023. Accessed 12 December 2021. https://www.isbtweb.org/resource/001aboalleles.html
28. Gassner C, Degenhardt F, Meyer S, et al. Low-frequency blood group antigens in Switzerland. *Transfus Med Hemotherapy*. 2018;45(4):239-250.
29. Olsson ML, Chester MA. Frequent occurrence of a variant O1 gene at the blood group ABO locus. *Vox Sang*. 1996;70(1):26-30.
30. Yazer M, Olsson M. The O2 allele: questioning the phenotypic definition of an ABO allele. *Immunohematol*. 2008;24(4):138-147.
31. Barbitoff YA, Bezdvornykh IV, Polev DE, et al. Catching hidden variation: systematic correction of reference minor allele annotation in clinical variant calling. *Genet Med*. 2018;20(3):360-364.
32. Ballouz S, Dobin A, Gillis JA. Is it time to change the reference genome? *Genome Biol*. 2019;20(1):1-9.
33. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *GigaScience*. 2021;10(2):giab008.
34. Shafin K, Pesout T, Chang P-C, et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat Methods*. 2021;18(11):1322-1332.
35. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303.
36. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*. 2009;25(11):1451-1452.
37. Leigh JW, Bryant D. POPART: full-feature software for haplotype network construction. *Methods Ecol Evol*. 2015;6(9):1110-1116.
38. Minh BQ, Schmidt HA, Chernomor O, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*. 2020;37(5):1530-1534.
39. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evol*. 2015;1(1):vev003.
40. Bild DE, Bluemke DA, Burke GL, et al. Multi-ethnic study of atherosclerosis: objectives and design. *Am J Epidemiol*. 2002;156(9):871-881.
41. Burke G, Lima J, Wong ND, Narula J. The Multiethnic Study of Atherosclerosis. *Global Heart*. 2016;11(3):267-268.
42. Olson JL, Bild DE, Kronmal RA, Burke GL. Legacy of MESA. *Global Heart*. 2016;11(3):269-274.
43. Tajima F. Evolutionary relationship of DNA sequences in finite populations. *Genetics*. 1983;105(2):437-460.
44. Kominato Y, Sano R, Takahashi Y, Hayakawa A, Ogasawara K. Human ABO gene transcriptional regulation. *Transfusion*. 2020;60(4):860-869.
45. Sano R, Nakajima T, Takahashi K, et al. Expression of ABO blood-group genes is dependent upon an erythroid cell-specific regulatory element that is deleted in persons with the B(m) phenotype. *Blood*. 2012;119(22):5301-5310.
46. Yamamoto F, Cid E, Yamamoto M, Saitou N, Bertranpetit J, Blancher A. An integrative evolution theory of histo-blood group ABO and related genes. *Sci Rep*. 2014;4(1):1-12.
47. Ségurel L, Thompson EE, Flutre T, et al. The ABO blood group is a trans-species polymorphism in primates. *Proc Natl Acad Sci USA*. 2012;109(45):18493-18498.
48. Villanea FA, Huerta-Sanchez E, Fox K. ABO genetic variation in Neanderthals and Denisovans. *Mol Biol Evol*. 2021;38(8):3373-3382.
49. Saitou N, Yamamoto F-i. Evolution of primate ABO blood group genes and their homologous genes. *Mol Biol Evol*. 1997;14(4):399-411.
50. Charlesworth D. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet*. 2006;2(4):e64.
51. Liumbruno GM, Franchini M. Beyond immunohaematology: the role of the ABO blood group in human diseases. *Blood Transfusion*. 2013;11(4):491-499.
52. Aspholm-Hurtig M, Dailide G, Lahmann M, et al. Functional adaptation of BabA, the H. pylori ABO blood group antigen binding adhesin. *Science*. 2004;305(5683):519-522.
53. Cserti CM, Dzik WH. The ABO blood group system and Plasmodium falciparum malaria. *Blood*. 2007;110(7):2250-2258.
54. Etemadi A, Kamangar F, Islami F, et al. Mortality and cancer in relation to ABO blood group phenotypes in the Golestan Cohort Study. *BMC Med*. 2015;13(1):8.
55. Rizzato C, Campa D, Pezzilli R, et al. ABO blood groups and pancreatic cancer risk and survival: results from the PANcreatic Disease ReseArch (PANDoRA) consortium. *Oncol Rep*. 2013;29(4):1637-1644.
56. Ohira T, Cushman M, Tsai M, et al. ABO blood group, other risk factors and incidence of venous thromboembolism: the Longitudinal Investigation of Thromboembolism Etiology (LITE). *J Thromb Haemostasis*. 2007;5(7):1455-1461.
57. Garratty G. Relationship of blood groups to disease: do blood group antigens have a biological role? *Rev Méd Inst Mex Seguro Soc*. 2005;43(suppl 1):113-121.
58. Franchini M, Bonfanti C. Evolutionary aspects of ABO blood group in humans. *Clin Chim Acta*. 2015;444:66-71.
59. Lalueza-Fox C, Gigli E, de la Rasilla M, et al. Genetic characterization of the ABO blood group in Neandertals. *BMC Evol Biol*. 2008;8(1):1-5.
60. Seltsam A, Hallensleben M, Kollmann A, Blasczyk R. The nature of diversity and diversification at the ABO locus. *Blood*. 2003;102(8):3035-3042.

61. Pompanon F, Bonin A, Bellemain E, Taberlet P. Genotyping errors: causes, consequences and solutions. *Nat Rev Genet*. 2005;6(11):847-859.

62. De Coster W, Van Broeckhoven C. Newest methods for detecting structural variations. *Trends Biotechnol*. 2019;37(9):973-982.

63. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol*. 2021;39(11):1348-1365.

64. Miga KH, Koren S, Rhie A, et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature*. 2020;585(7823):79-84.

65. Gassner C, Schmarda A, Nussbaumer W, Schonitzer D. ABO glycosyltransferase genotyping by polymerase chain reaction using sequence-specific primers. *Blood*. 1996;88(5):1852-1856.

66. Malik R, Traylor M, Pulit SL, et al. Low-frequency and common genetic variation in ischemic stroke: the METASTROKE collaboration. *Neurology*. 2016;86(13):1217-1226.

67. Lindström S, Wang L, Smith EN, et al. Genomic and transcriptomic association studies identify 16 novel susceptibility loci for venous thromboembolism. *Blood*. 2019;134(19):1645-1657.

68. Paré G, Chasman DI, Kellogg M, et al. Novel association of ABO histo-blood group antigen with soluble ICAM-1: results of a genome-wide association study of 6,578 women. *PLoS Genet*. 2008;4(7):e1000118.

69. Goumidi L, Thibord F, Wiggins KL, et al. Association of ABO haplotypes with the risk of venous thrombosis: impact on disease risks estimation. *Blood*. 2021;137(17):2394-2402.

70. Giollo M, Minervini G, Scalzotto M, Leonardi E, Ferrari C, Tosatto SC. BOOGIE: predicting blood groups from high throughput sequencing data. *PLoS One*. 2015;10(4):e0124579.