Supplemental Information

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*]

Affiliations:

[1]Department of Research and Development, Blood Transfusion Service Zurich, Swiss Red Cross, 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, Swiss Red Cross, Schlieren, Switzerland

[5]Blood Transfusion Service Zurich, Swiss Red Cross, Schlieren, Switzerland

[6]Institute for Translational Medicine, Private University in the Principality of Liechtenstein, Triesen, Liechtenstein

[7]Institute for Transfusion Medicine, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany

[§]These authors contributed equally to this work

*Correspondence to: m.mattle@zhbsd.ch

Table of contents

Section 1: Sample selection

1.1 Overview of sample set

We selected 77 samples (Tables S1 and S2) from a large, well-characterized ABO genotype dataset (n=25,200) of serologically-typed blood donors from the greater Zurich area in Switzerland. This data had been generated previously using MALDI-TOF mass spectrometry (MALDI-TOF MS)[1]. We aimed to sequence at least 15 haplotypes for each of the six main *ABO* groups, i.e. *ABO*A1*, *A2*, *B*, *O.01.01*, *O.01.02*, and *O.02* (Table 1). We selected a mix of (i) *ABO* homozygous samples at pre-typed variants (i.e. same bases inherited from the mother and father, n=43), and (ii) *ABO* group heterozygous samples (i.e. different bases inherited, n=34), see SI Section 1.2. The putatively *ABO* allele homozygous individuals were included to support haplotype resolving after sequencing. A detailed list of sequenced *ABO* haplotypes including GenBank accession numbers is provided in Table S2. 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 no subject to ethical authorization.

1.2 Details of the sample selection process

The large ABO genotype dataset from which we selected samples had been generated previously in the course of a research project aiming at identifying rare blood donors by high-throughput genotyping of antigens across blood group systems using MALDI-TOF MS[1,2]. This dataset included information on the three causative variants designating *ABO*B* (c.803G>C, rs8176747), *ABO*O.01* (c.261delG, rs8176719), and *ABO*O.02* (c.802G>A, rs41302905), as well as detailed ABO serology. According to common practice, the allele group *ABO*A* was deduced from the information on the other designating variants (i.e. absence of *ABO*B*, *ABO*O.01* and *ABO*O.02* designating variants).

From this dataset, we randomly selected 378 individuals being homozygous for either of the variants designating *ABO*A* (n=142), *ABO*B* (n=73), *ABO*O.01* (n=142), and *ABO*O.02* (n=21). We genotyped these samples at 11 additional variants along the *ABO* gene to detect further potential gene heterozygosities (Table S3). The aim of this extended heterozygosity analysis was to identify individuals being potentially homozygous at the entire *ABO* gene locus. We aimed to include such individuals in our study to support *ABO* haplotype resolving after sequencing.

The 11 additional variants included two variants designating *ABO*A2* (c.467C>T and c.1061delC) as well as three single nucleotide variants (SNVs) separating *ABO*O.01.01* and *ABO*O.01.02* (c.106G>T, c.188G>A, and c.646T>A) (Table S3). Genotyping of the additional variants was done again using MALDI-TOF MS as described previously[1,2]. In total, 258 individuals remained homozygous at all genotyped variants. We selected 43 samples for the study with sufficient genomic DNA stock solution at appropriate quality. In order to increase the total sample size, we supplemented our sample set with randomly selected *ABO* group heterozygous samples (n=34; Table S1). For these *ABO* group heterozygous samples, the *ABO*O.01* subgroup information was not available a priori and was obtained from the sequencing data in retrospect. For all included study samples, we used stored pre-extracted genomic DNA from whole blood samples described in Gassner et al.[2].

Table S1. Detailed information on study samples and Oxford Nanopore sequencing data.

Sample ID	Assigned _ABO_ group for the two haplotypes[a]	ABO serology	_ABO_ genotype by MALDI-TOF MS[b]	Coverage LR1[c]	Coverage LR2[c]
s04	_ABO*A1_ | _ABO*A1_	A_1	A1 | A1	2680	9074
s05	_ABO*A1_ | _ABO*A1_	A_1	A1 | A1	347	5787
s30	_ABO*A1_ | _ABO*A1_	A_1	A1 | A1	1015	5330
s40	_ABO*A1_ | _ABO*A1_	A_1	A1 | A1	649	242
s41	_ABO*A1_ | _ABO*A1_	A_1	A1 | A1	4373	12478
s46	_ABO*A1_ | _ABO*A1_	A_1	A1 | A1	2615	452
s58	_ABO*A1_ | _ABO*A1_	A_1	A1 | A1	5626	10175
s64	_ABO*A1_ | _ABO*A1_	A_1	A1 | A1	1270	121
s70	_ABO*A1_ | _ABO*A1_	A_1	A1 | A1	197	154
s53	_ABO*A2_ | _ABO*A2_	A_2	A2 | A2	3598	7544
s89	_ABO*A2_ | _ABO*A2_	A_2	A2 | A2	269	4250
s90	_ABO*A2_ | _ABO*A2_	A_2	A2 | A2	400	283
s104	_ABO*A2_ | _ABO*A2_	A_2	A2 | A2	1481	2959
s07	_ABO*B_ | _ABO*B_	B	B | B	414	8084
s31	_ABO*B_ | _ABO*B_	B	B | B	585	5060
s43	_ABO*B_ | _ABO*B_	B	B | B	3372	8910
s55	_ABO*B_ | _ABO*B_	B	B | B	4008	20778
s13	_ABO*O.01.01_ | _ABO*O.01.01_	O	O1.01 | O1.01	424	399
s14	_ABO*O.01.01_ | _ABO*O.01.01_	O	O1.01 | O1.01	3922	6165
s20	_ABO*O.01.01_ | _ABO*O.01.01_	O	O1.01 | O1.01	61	1253
s26	_ABO*O.01.01_ | _ABO*O.01.01_	O	O1.01 | O1.01	3887	4319
s37	_ABO*O.01.01_ | _ABO*O.01.01_	O	O1.01 | O1.01	1329	4726
s38	_ABO*O.01.01_ | _ABO*O.01.01_	O	O1.01 | O1.01	3260	4575
s49	_ABO*O.01.01_ | _ABO*O.01.01_	O	O1.01 | O1.01	1377	1142
s61	_ABO*O.01.01_ | _ABO*O.01.01_	O	O1.01 | O1.01	1225	3203
s73	_ABO*O.01.01_ | _ABO*O.01.01_	O	O1.01 | O1.01	1190	3609
s85	_ABO*O.01.01_ | _ABO*O.01.01_	O	O1.01 | O1.01	453	5029
s107	_ABO*O.01.01_ | _ABO*O.01.01_	O	O1.01 | O1.01	786	1707
s03	_ABO*O.01.02_ | _ABO*O.01.02_	O	O1.02 | O1.02	660	1828
s15	_ABO*O.01.02_ | _ABO*O.01.02_	O	O1.02 | O1.02	869	52
s50	_ABO*O.01.02_ | _ABO*O.01.02_	O	O1.02 | O1.02	1427	6991
s51	_ABO*O.01.02_ | _ABO*O.01.02_	O	O1.02 | O1.02	1170	5816
s62	_ABO*O.01.02_ | _ABO*O.01.02_	O	O1.02 | O1.02	1559	92
s63	_ABO*O.01.02_ | _ABO*O.01.02_	O	O1.02 | O1.02	3151	8859
s74	_ABO*O.01.02_ | _ABO*O.01.02_	O	O1.02 | O1.02	1929	8548
s75	_ABO*O.01.02_ | _ABO*O.01.02_	O	O1.02 | O1.02	1500	3812
s86	_ABO*O.01.02_ | _ABO*O.01.02_	O	O1.02 | O1.02	1639	271
s87	_ABO*O.01.02_ | _ABO*O.01.02_	O	O1.02 | O1.02	1848	6280
s109	_ABO*O.01.02_ | _ABO*O.01.02_	O	O1.02 | O1.02	1533	3411
s110	_ABO*O.01.02_ | _ABO*O.01.02_	O	O1.02 | O1.02	216	1631
s57	_ABO*O.02_ | _ABO*O.02_	O	O2 | O2	2200	889
s68	_ABO*O.02_ | _ABO*O.02_	O	O2 | O2	54	182
s111	_ABO*O.02_ | _ABO*O.02_	O	O2 | O2	718	1742
s16	_ABO*A1_ | _ABO*A2_	A_1	A1 | A2	1304	318
s17	_ABO*A1_ | _ABO*A2_	A_1	A1 | A2	1540	1600
s27	_ABO*A1_ | _ABO*A2_	A_1	A1 | A2	1304	219
s28	_ABO*A1_ | _ABO*A2_	A_1	A1 | A2	710	5357
s29	_ABO*A1_ | _ABO*A2_	A_1	A1 | A2	1038	473
s34	_ABO*A1_ | _ABO*A2_	A_1	A1 | A2	75	126

Table S1 continued.

s39	*ABO*A1* | *ABO*A2*	A_1	A1 | A2	566	50
s52	*ABO*A1* | *ABO*A2*	A_1	A1 | A2	1480	2494
s65	*ABO*A1* | *ABO*A2*	A_1	A1 | A2	595	5737
s69	*ABO*A1* | *ABO*A2*	A_1	A1 | A2	194	793
s76	*ABO*A1* | *ABO*A2*	A_1	A1 | A2	959	3621
s77	*ABO*A1* | *ABO*A2*	A_1	A1 | A2	1338	7318
s10	*ABO*A1* | *ABO*B*	A_1B	A | B	450	6442
s18	*ABO*A1* | *ABO*B*	A_1B	A | B	85	5715
s22	*ABO*A1* | *ABO*B*	A_1B	A | B	430	642
s45	*ABO*A1* | *ABO*B*	A_1B	A | B	1857	4948
s81	*ABO*A1* | *ABO*B*	A_1B	A | B	553	300
s06	*ABO*A2* | *ABO*B*	A_2B	A | B	167	6543
s01	*ABO*O.01.01* | *ABO*B*	B	O1 | B	104	3017
s25	*ABO*O.01.02* | *ABO*B*	B	O1 | B	900	81
s42	*ABO*O.01.02* | *ABO*B*	B	O1 | B	2744	7022
s54	*ABO*O.01.01* | *ABO*B*	B	O1 | B	935	6495
s56	*ABO*O.01.02* | *ABO*B*	B	O1 | B	50	219
s93	*ABO*O.01.02* | *ABO*B*	B	O1 | B	524	3690
s11	*ABO*O.01.01* | *ABO*O.02*	O	O1 | O2	238	59
s19	*ABO*O.01.01* | *ABO*O.02*	O	O1 | O2	3025	7183
s21	*ABO*O.01.02* | *ABO*O.02*	O	O1 | O2	861	1076
s66	*ABO*O.01.02* | *ABO*O.02*	O	O1 | O2	99	3593
s67	*ABO*O.01.02* | *ABO*O.02*	O	O1 | O2	1219	3408
s78	*ABO*O.01.01* | *ABO*O.02*	O	O1 | O2	1807	1578
s33	*ABO*O.02* | *ABO*A1*	A_1	O2 | A	1083	2752
s79	*ABO*O.02* | *ABO*A1*	A_1	O2 | A	1974	7512
s88	*ABO*O.02* | *ABO*A1*	A_1	O2 | A	780	284
s91	*ABO*O.02* | *ABO*A1*	A_1	O2 | A	219	3157

[a] *ABO* group assignment of both haplotypes (maternal and paternal allele) for the purpose of this study based on serological and genetical (i.e. MALDI-TOF MS genotype data) prevalues. For *ABO* group heterozygous samples, the *ABO*O.01* subgroup information was obtained from the sequencing data in retrospect; [b] Genotype of both haplotypes obtained by MALDI-TOF MS genotyping (see SI Section 1.2); [c] Sequence read coverage per long-range PCR fragment (LR1 and LR2) of pooled Oxford Nanopore sequencing.

Table S2. List of sequenced *ABO* haplotypes according to current ISBT nomenclature based on nucleotide changes in exons. Provided as separate Excel file.

Table S3. Genetic variants included in the extended heterozygosity analysis using MALDI-TOF mass spectrometry.

rs number	SNV	Nucleotide change[a]	Exon/Intron location
rs2073828	[C/T]	c.98+362C>T	Intron 2
rs688976	[G/T][b]	c.106G>T	Exon 3
rs8176702	[C/T]	c.155+575C>T	Intron 3
rs549446	[G/A][b]	c.188G>A	Exon 4
rs638756	[T/G]	c.203+738T>G	Intron 4
rs514708	[G/A]	c.204-220G>A	Intron 4
rs4962040	[C/T]	c.204-9C>T	Intron 4
rs7873416	[A/G]	c.374-103A>G	Intron 6
rs1053878	[C/T]	c.467C>T	Exon 7
rs8176740	[T/A][b]	c.646T>A	Exon 7
rs56392308	[C/-]	c.1061delC	Exon7

[a] *ABO* transcript: NM_020469.3; [b] SNVs discriminating the two subgroups *ABO*O.01.01* and *ABO*O.01.02*.

1.3 *ABO* allele frequency estimation based on the MALDI-TOF MS genotype dataset

We performed data mining from the ABO genotype dataset (n=25,200) generated by MALDI-TOF MS[1] (SI Sections 1.1. and 1.2) to estimate frequencies of each of the six main *ABO* allele groups, i.e. *ABO*A1*, *A2*, *B*, *O.01.01*, *O.01.02*, and *O.02* in the region of Zurich (Table 1). For each ABO phenotype (i.e. O, A, B, and AB), we calculated the frequencies of the genotypes underlying the phenotype. Due to favored sampling of O phenotypes[1], calculated genotype frequencies solely represented genotype distributions within respective phenotype groups, and could not be compared among phenotype groups. Hence, we standardized frequencies using ABO phenotype data from 1000 consecutive first-time donors from the same donor population (i.e. Blood Transfusion Service Zurich, Switzerland). Based on the observed actual distribution of O, A, B, and AB phenotypes in the donor population, we estimated standardized genotype frequencies (Table S4). From this data, respective allele frequencies for the six main *ABO* allele groups of this study (*ABO*A1*, *A2*, *B*, *O.01.01*, *O.01.02*, and *O.02*) were summed-up (Table 1). Allele frequencies for the two subgroups *O.01.01* and *O.01.02* were split according to their observed genotype distribution in the extended heterozygosity analysis (see SI Section 1.2). Separate frequencies for the alleles *ABO*A1* and *ABO*A2* were calculated using the actual distribution of A_1 and A_2 phenotypes in the subset of the MALDI-TOF MS genotype dataset [1] with the respective detailed serological information (n=2,442).

Table S4. Estimated *ABO* genotype frequencies in the region of Zurich in Switzerland.

Phenotype	Genotype	Estimated frequency
O	*O1O1*	41.18%
	O1O2	2.80%
	O2O2	0.02%
A	*O1A*	32.57%
	O2A	1.19%
	AA	7.44%
B	*O1B*	9.66%
	O2B	0.18%
	BB	0.47%
AB	*AB*	4.50%

Section 2: Oxford Nanopore sequencing

2.1 Long-range PCRs of the *ABO* gene locus

We established generic long-range PCRs (LR-PCR) amplifying the entire *ABO* gene including flanking regulatory regions (~23.6 kb; exact length dependent on haplotype) in two overlapping fragments (Figure 1). Fragment LR1 (16.9 kb) covered the enhancer region 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.

LR1 was amplified using the PCR primer pair ABO_K13_MG03_F [5'-TCCTTCTCTCACCTGCCCCACTTTA-3'] and ABO_K13_MG03_R [5'-TAAGCTCTTGCTCCTAGATGATAAAGAAGAAC-3']; LR2 was amplified using the PCR primer pair ABO-K3-F [5'-AGTCTGACGTTAGCATTTCTCCTCAAG-3'] and ABO-K3-R [5'-CTAGGCTTCAGTTACTCACAACAG-3'].

Prior to PCR amplification, DNA concentration of all study samples was measured using a NanoDrop-3000 spectrophotometer (Thermo Scientific) and samples were diluted to 20 ng/µl with sterile H_2O. LR-PCR amplifications were performed in duplicates using a PrimeSTAR GXL DNA polymerase (TaKaRa Bio) according to the manufacturer's instructions. In brief, PCR reactions were carried out in a total volume of 50 µl and were composed of 200 ng DNA template, 1x PrimeSTAR GXL Buffer (TaKaRa Bio), 1.25 U PrimerSTAR GXL DNA polymerase (TaKaRa Bio), 200 µM of each dNTP, 1 M Betaine enhancer (VWR), and 0.2 µM of each PCR primer. As suggested by the manufacturer, we used a two-step amplification profile with a 10 second denaturation step at 95°C and a 10 minutes extension step at 68°C for 30 cycles. Amplification success was verified by agarose gel electrophoresis using 0.8% agarose gels stained with GelRed Nucleic Acid Gel Stains (Biotium). For each sample, PCR replicates were pooled prior to purification with 1x Agencourt AMPure XP magnetic beads (Beckman Coulter) according to the manufacturer's instructions for PCR clean up. Purified PCR products were eluted in 20 µl sterile H_2O and quantified using a dsDNA broad range assay kit on a Qubit fluorometer 3.0 (Invitrogen).

2.2 ONT library preparation and sequencing

Oxford Nanopore Technologies (ONT) sequencing libraries were prepared following ONT's protocol for native (i.e. PCR-free) barcoding of amplicons (protocol name: 'Amplicon barcoding with Native Barcoding Expansion 96; version: NBA_9102_v109_revF_09Jul2020'). Briefly, per sample 50 fmol of both LR-PCR fragments were pooled and end-repaired using a NEBNext Ultra II End repair / dA-tailing enzyme (New England Biolabs). End prepared amplicons were then uniquely barcoded using the 'Native Barcoding Expansion 96 (EXP-NBD196)' kit (ONT) and a Blunt/TA Ligase Master Mix (New England Biolabs). Barcoded libraries were pooled and purified using 0.4x Agencourt AMPure XP magnetic beads (Beckman Coulter). ONT-specific sequencing adapters were then ligated to the amplicons with a Quick T4 DNA Ligase (New England Biolabs). The final library was sequenced on two MinION Mk1B (R9.4.1) flow cells.

Section 3: Bioinformatic analysis of Oxford Nanopore sequencing data

3.1 Basecalling and read processing

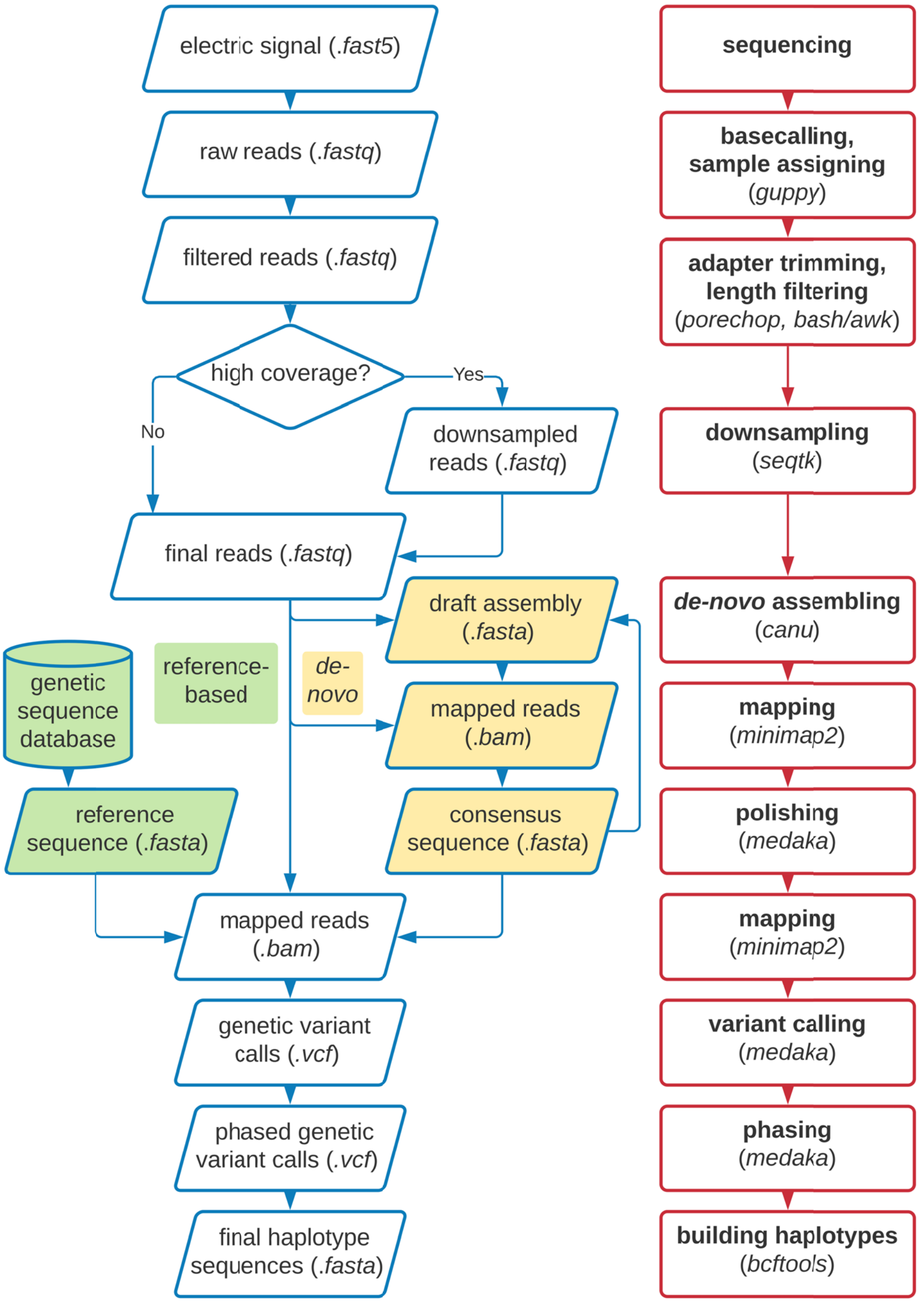
electric signal (*.fast5*)
raw reads (*.fastq*)
filtered reads (*.fastq*)
high coverage?
Yes
No
downsampled reads (*.fastq*)
final reads (*.fastq*)
draft assembly (*.fasta*)
genetic sequence database
reference-based
de-novo
mapped reads (*.bam*)
reference sequence (*.fasta*)
consensus sequence (*.fasta*)
mapped reads (*.bam*)
genetic variant calls (*.vcf*)
phased genetic variant calls (*.vcf*)
final haplotype sequences (*.fasta*)
sequencing
basecalling, sample assigning (*guppy*)
adapter trimming, length filtering (*porechop, bash/awk*)
downsampling (*seqtk*)
de-novo assembling (*canu*)
mapping (*minimap2*)
polishing (*medaka*)
mapping (*minimap2*)
variant calling (*medaka*)
phasing (*medaka*)
building haplotypes (*bcftools*)

Figure S1. Overview of the bioinformatic analysis pipeline of ONT sequencing data.

The workflow for processing ONT data is depicted in Figure S1. Raw electrical signals (squiggles) were stored in FAST5 format files, which were then basecalled and demultiplexed (read assignment to the respective sample according to barcode) with the ONT's stand-alone *Guppy* software (v4.4.1) based on a high-accuracy model. Reads with a Phred-scaled quality score (Q) < 10 were excluded and successful sample assignment during demultiplexing required a minimal barcode detection score of

70 at either end of the read. Such classified raw reads in FASTQ format were then filtered based on the expected length of the two LR-PCR fragments and the observed read length distribution in the sequencing report. Acceptable ranges were set to 1 kb, i.e. reads between 16.5 kb and 17.5 kb in length (LR1) and between 12.5 kb and 13.5 kb (LR2) were selected. Remaining adapter or barcode sequences that had not been detected and trimmed by *Guppy* were chopped off by *Porechop* (v0.2.4). Filtered reads represented both amplicons in all 77 samples.

Read numbers per sample (Table S1) showed a median of 4428x (interquartile range 1573x to 7208x). LR1 was sequenced to a lower depth (median number of reads 1038x; interquartile range 440x to 1723x) than LR2 (median 3408x; interquartile range 463x to 5991x), likely related to the well-known preferential sequencing of shorter fragments (i.e. LR2) as they pass the nanopores faster than longer fragments. In order to reduce computational time for downstream analysis, we set a cut-off at 1000 reads per amplicon by random downsampling with *seqtk* (v1.3), thus limiting the number of reads to maximally 2000x per sample. The lowest read number (i.e. coverage) per amplicon was 50x (LR1 of s56 and LR2 of s39). This value lies still well above the standard coverage (30x) strived for in whole-genome sequencing (WGS) projects. Furthermore, we observed in downsampling tests that variants were called very reliably within the coverage range of 50x to 1000x (see SI Section 3.3).

3.2 Sequence read mapping and *de-novo* assembly pipeline

Ubiquitous read mapping against a single reference sequence may lead to unnoticed allelic drop-out, a risk that is increasing with sequence divergence of the reference sequence and the sequenced allele[3,4]. Therefore, we used a *de-novo* assembly pipeline, i.e. assembled for each sample its own consensus sequence from both PCR-amplicons. This approach is completely independent of any reference sequence.

First, *Canu* (v2.2.1)[5] was used to create a draft assembly for each sample. We switched off the default downsampling by setting *readSamplingCoverage* to 2000 (default 200). Also, *corOutCoverage* was set to 9999 to make sure that all reads got corrected. As suggested for assemblies from amplicons, we further disabled the contig filter by setting *contigFilter* to '2 0 1.0 0.5 0' avoiding having to define a minimal coverage for generated contigs. *MhapSensitivity* was set to 'low' as recommended when high coverage data is available. In over 80% of all samples these parameter settings led to a successful draft assembly. For the remaining samples, an approach of increasing the option *correctedErrorRate* to 0.2 (allowing higher discordance within corrected reads) and/or strong downsampling to 30x was successful. Draft assemblies for each sample were then polished twice with the tool *medaka_consensus* of the ONT's *Medaka* package (v1.2.2). All filtered reads of a sample were mapped with *Minimap2* (v2.17)[6] to its own polished assembly sequence. Secondary and supplementary mappings were flagged and filtered out. We set the minimal chaining score (option –m, approximating the number of matching bases) to 6000, corresponding to about half of the read lengths. Mappings were saved in the BAM format.

As an alternative strategy for sequence read assembly, we also mapped ONT reads using classical single-reference-based read mapping for comparison with the *de-novo* pipeline. In this reference-based mapping pipeline, we used as reference the current *ABO* reference sequence NG_006669.2, which represents an *ABO*A1.01* allele. *Minimap2* (v2.17) was used as described above*.*

3.3 Variant calling and haplotype phasing

We used the tool *medaka_variant* (*Medaka* v1.2.2) for variant calling and phasing of called variants (i.e. haplotype reconstruction). *medaka_variant* first calls SNVs from unphased reads, followed by phasing based on obtained SNV calls, which allows final SNV and indel calling for each haplotype. Output files (before and after phasing) were saved in variant call format (VCF).

We flagged intermediate VCF files (after phasing but before the haplotype-separated variant calling step) if they reported unphased variants above Q9 (n = 17). This either pointed to samples with only one heterozygous SNV (n = 14), in which phasing was irrelevant, or to samples that contained one or more SNVs in the non-overlapping region of both PCR fragments, but not in the overlap (n = 3). The latter were inspected manually with the *Integrative Genomics Viewer* (*IGV,* v2.7.2)[7]. Genetic variation in the overlapping region not surpassing the insensitive pre-phasing SNV calling threshold was detected in all three samples and unphased positions in corresponding intermediate VCF files were manually edited before running the haplotype-separate variant calling step of *medaka_variant*.

For both SNV and indel calling, we set the threshold for low quality calls to ≥ Q20, i.e. only calls above this threshold were finally considered as true variants. This threshold was above the default settings (Q8 for SNVs and Q9 for indels). As models underlying *Medaka* have only been trained with up to 60x sequencing data (as typical in WGS settings), we first ensured that *Medaka* could also handle much higher coverage. We tested and compared variant calls based on read depths up to 5000x reads per PCR amplicon with those based on reads downsampled to 50x per amplicon. As expected, calls were very stable when leveraging read depths ≥ 50x. We observed, however, a slight trend to higher call quality values for a given variant the higher the variant was covered. Since we had in this study sequence coverage of up to 2000x per sample, a more stringent quality score threshold than the default settings seemed justified. In fact, we observed that most variants falling into the Q10 − Q20 quality range were indels in repetitive sequence motifs, which still represent the biggest challenge with nanopore sequencing[8]. Hence, we decided to rather tolerate a slight reduction in sensitivity (i.e. may not reliably call every indel present) than lowering accuracy by calling inexistent indels.

Finally, we used *BCFtools* (v1.11)[9] to generate haplotype FASTA sequences for all study samples from the generated VCF files containing the phased variants, masking out the low quality calls.

3.4 Sanger sequencing validation

To achieve best accuracy of generated FASTA sequences, we validated some sites in repetitive regions by Sanger sequencing. Specific PCR primers were designed and PCRs carried out with a KAPA HiFi HotStart ReadyMix (Roche). Primers and PCR reaction protocols are available upon request. Amplicon purification and Sanger sequencing were outsourced to an external company (Microsynth AG, Balgach, Switzerland). In total, we investigated 12 repetitive regions but managed to obtain clean Sanger sequences for only three, all homopolymeric stretches of 8-10 A/T. Compared to the Sanger sequences, ONT haplotypes consistently harbored one further A/T in these regions. These sequences were manually corrected according to the Sanger data. All other investigated regions that were too large for obtaining high-quality Sanger sequences were processed as stated in SI Section 4.4.

Section 4: Illumina and PacBio HiFi sequencing

For quality validation of obtained ONT sequences, a subset of 12 samples (n = 2 for each *ABO* group; Table S5), which were *ABO* homozygous at pre-typed SNVs, was additionally sequenced using both short-read Illumina sequencing on a MiSeq instrument and long-read PacBio HiFi sequencing on a Sequel II system.

Table S5. List of samples additionally sequenced using an Illumina/PacBio hybrid approach.

Sample ID	Assigned *ABO* group for the two haplotypes[a]	ABO serology	*ABO* genotype by MALDI-TOF MS[b]	Coverage[c]
s41	*ABO*A1 | ABO*A1*	A_1	A1 | A1	2991
s46	*ABO*A1 | ABO*A1*	A_1	A1 | A1	3109
s104	*ABO*A2 | ABO*A2*	A_2	A2 | A2	3038
s89	*ABO*A2 | ABO*A2*	A_2	A2 | A2	3222
s31	*ABO*B | ABO*B*	B	B | B	3166
s43	*ABO*B | ABO*B*	B	B | B	3038
s107	*ABO*O.01.01 | ABO*O.01.01*	O	O1.01 | O1.01	3063
s49	*ABO*O.01.01 | ABO*O.01.01*	O	O1.01 | O1.01	3126
s51	*ABO*O.01.02 | ABO*O.01.02*	O	O1.02 | O1.02	3195
s74	*ABO*O.01.02 | ABO*O.01.02*	O	O1.02 | O1.02	3005
s111	*ABO*O.02 | ABO*O.02*	O	O2 | O2	2983
s68	*ABO*O.02 | ABO*O.02*	O	O2 | O2	2820

[a] *ABO* group assignment of both haplotypes (maternal and paternal sequences) for the purpose of this study based on serological and genetical (i.e. MALDI-TOF MS genotype data) prevalues; [b] Genotype of both haplotypes obtained by MALDI-TOF MS genotyping (see SI Section 1.2); [c] Mean coverage based on combined Illumina/PacBio sequencing data.

4.1 Illumina library preparation and sequencing

All steps of the Illumina DNA library preparation starting from the enzymatic fragmentation, followed by the end repair, adapter ligation, and paired-end index PCR were done with the NEBNext®Ultra™DNA Library Prep Kit for Illumina (New England Biolabs) according to the manufacturer instructions. Prior to the sequencing run, we pooled the barcoded amplicons, separated them on a 1.6% TAE agarose gel and isolated the final library fraction ranging between 600-800 bp with the GeneJet Gel Extraction Kit (ThermoFisher Scientific). For library quantification, we used the Kapa Library Quantification Kit (Roche) according to the manufacturer's supplied protocol. Before loading the sequencing cartridges, the libraries were diluted to 15 pmol and denatured according to standard Illumina loading procedures. Sequencing was performed on a MiSeq instrument (Illumina) running 500 cycles of v2 chemistry.

4.2 PacBio HiFi library preparation and sequencing

SMRTbell® Express Template Prep Kit 2.0 libraries (PacBio) were prepared from the amplicons as Multiplexed Microbial Libraries. The Sequel® II Binding Kit 2.0 (PacBio) was used to bind prepared DNA template libraries to the Sequel® II Polymerase 2.0. Long-read sequencing was performed on a PacBio Sequel II system.

4.3 Bioinformatic analyses of Illumina and PacBio data

The demultiplexed Illumina reads were mapped with the *Burrows-Wheeler Aligner* (*bwa;* v.0.7.17-r1188)[10] against the *ABO* reference sequence NG_006669.2. PacBio sequences were processed with the *PacBio Secondary Analysis Tools* on Bioconda (https://github.com/PacificBiosciences/pbbioconda). Circular consensus sequences with at least 5 passes were generated from the demultiplexed raw reads. These consensus sequences were mapped against the *ABO* reference sequence using *pbmm2* (v.1.3.0). A combined variant calling was done with the *Genome Analysis Toolkit* (*GATK*; v.4.1.4.1)[11], following the best practice guidelines for germline short variant discovery[12].

4.4 Quality validation of ONT data with Illumina/PacBio data

For performance comparison of the two alternative ONT assembly pipelines (see SI Section 3), we aligned for each sample the two ONT haplotype sequences and its corresponding unphased Illumina/PacBio sequence. This was done separately for sequences from both ONT assembly pipelines. Heterozygous positions were manually checked and proved concordant. Remaining deviations were all limited to highly repetitive sequence motifs in intronic regions, mainly longer homopolymers. Both assembly pipelines performed similarly well in comparison with the Illumina/PacBio data. Because of the general advantages of a per sample *de-novo* assembly over single-reference based read mapping[3,4], we used the sequences generated by the ONT *de-novo* assembly pipeline for all analyses and submission to the GenBank sequence database. Observed deviations in the highly repetitive sequence motifs were manually corrected with the corresponding Illumina/PacBio sequence based on a multiple sequence alignment.

Section 5: Genetic diversity analyses

To investigate genetic diversity patterns within and between the six *ABO* groups, we calculated several diversity statistics based on the analysis sequence alignment using *DNAsp.v6*[13] (Table 3 and S6). For each *ABO* group, we computed (i) the number of segregating sites (S), corresponding to the number of SNVs[14], (ii) the number of insertion/deletion (indel) events, (ii) the number of unique haplotypes (h)[14], (iv) haplotype diversity (Hd)[14], (v) the average number of nucleotide differences between sequences (k)[15], and (vi) nucleotide diversity (π), which is the average number of nucleotide differences per site between two sequences[14]. To study genetic diversity between *ABO* groups, we calculated the average number of nucleotide differences between groups and the number of fixed nucleotide differences (i.e. sites for which one group has one allele and the other group the other allele).

Genetic diversity was much higher between *ABO* groups than within groups (Tables 3, 4, and S6). Within *ABO* groups (Tables 3 and S6), nucleotide diversity (π) was particularly low for *ABO*A1* (0.00002), *B* (0.00004) and *A2* (0.00006). The group of *ABO*O.01.01* showed comparatively highest within-group diversity (π = 0.00043). This appeared to be linked to deep within-group substructure into two phylogenetic clades (see Figures 3 and 4), which is inflating diversity measurements.

Table S6. Detailed statistics on genetic diversity among the 154 *ABO* haplotype sequences. This is an extended version of the Table 3 with more detailed statistics. For comparison, statistics are also provided for *ABO*O.01* without separating the two subgroups *ABO*O.01.01* and *ABO*O.01.02*.

	All	*ABO*A1*	*ABO*A2*	*ABO*B*	*ABO*O.01*	*ABO*O.01.01*	*ABO*O.01.02*	*ABO*O.02*
No. of sequences (N)	154	39	21	20	58	27	31	16
No. of segregating (i.e. polymorphic) sites (S)[a]	230	7	7	6	110	23	14	18
No. of indel events; [No. of indel sites in bp][b]	16 [204]	0	0	0	2 [16]	0	0	1 [9]
No. of unique haplotypes (h)	47	5	5	6	25	14	11	7
Haplotype diversity (Hd)	0.919	0.197	0.652	0.516	0.900	0.895	0.725	0.775
Average no. of nucleotide differences (k)[c]	66.439	0.408	1.095	0.768	44.747	8.473	1.918	1.842
Nucleotide diversity (π)[d]	0.00339	0.00002	0.00006	0.00004	0.00228	0.00043	0.00010	0.00009

[a]Corresponding to the number of SNVs; [b]number of insertion/deletion (indel) events and total number of indel sites in basepairs; [c]average number of nucleotide differences between two sequences[15]; [d]average number of nucleotide differences per site between two sequences[14].

Section 6: Recombination analyses

We used the recombination detection program *RDP4*[16] to detect potential recombination events between *ABO* allele groups based on the complete analysis sequence alignment containing all 154 *ABO* haplotypes. We first run the *PHI*-test implemented in *RDP4*, which tests for the presence of overall recombination signals in the dataset. As this statistical test found strong evidence for recombination events ($p < 10^{-5}$), we further investigated where on the gene (i.e. breakpoints) and between which haplotypes recombination events might have occurred. We applied seven different methods implemented in *RDP4*, i.e. RDP, bootscan, maxchi, chimaera, 3seq, geneconv, and siscan. Default settings were used except for the option specifying 'linear' sequences instead of 'circular'. Only events depicted by at least five different methods were considered reliable (Figure S2).

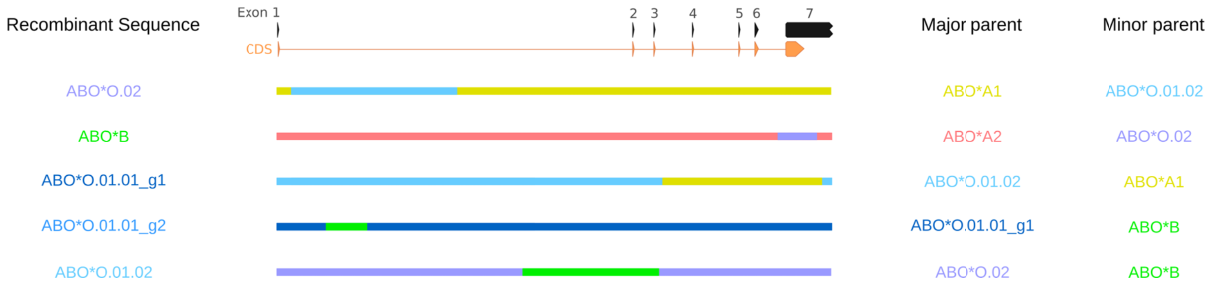

Figure S2. *ABO* recombination events identified by the recombination detection program *RDP4*.

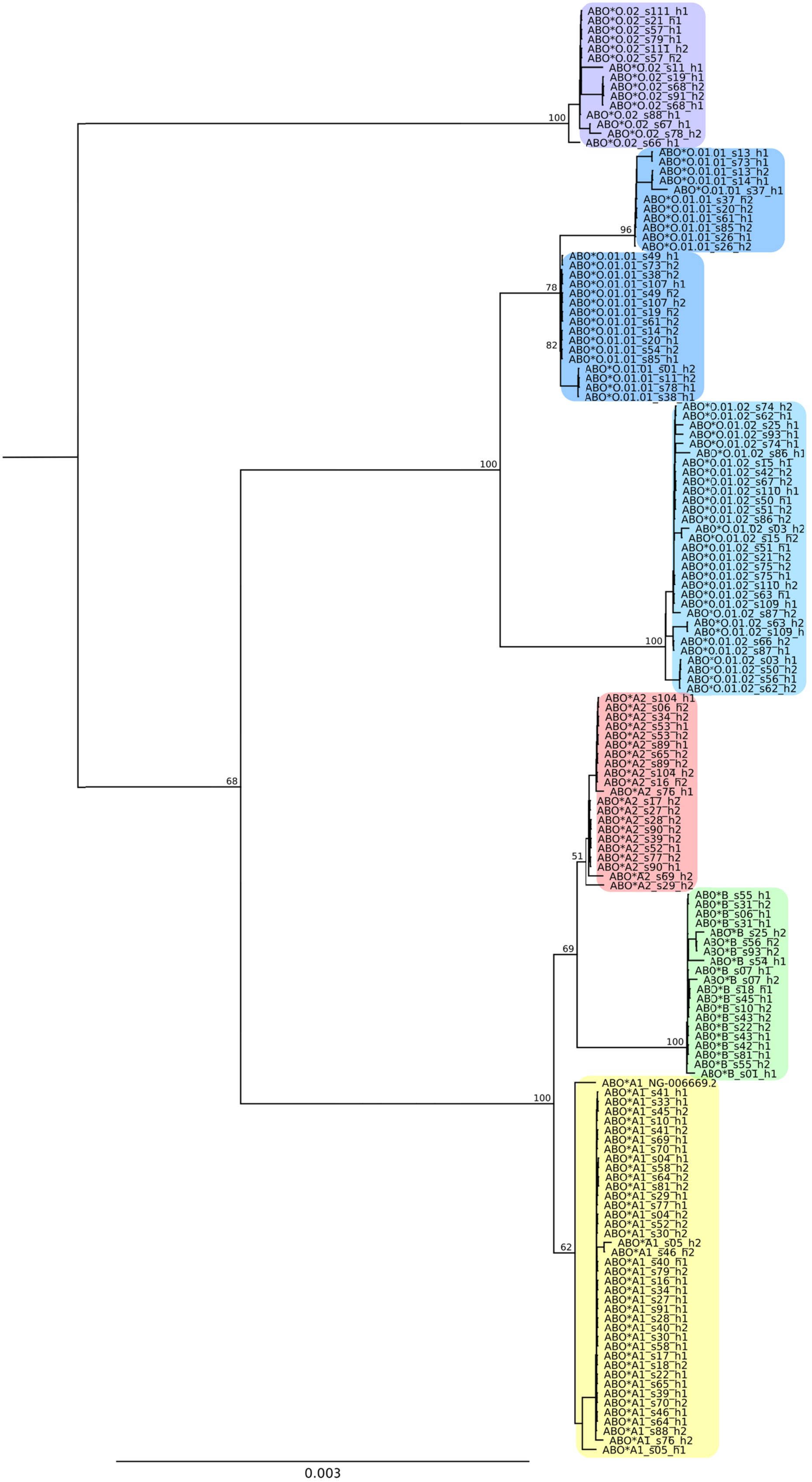

Figure S3. Maximum-likelihood phylogenetic tree based on *ABO* alignment with identified recombinant regions removed. All *ABO* groups are monophyletic. Bootstrap support is provided for main branching points. The tree was rooted with central chimpanzee sequence (not shown).

Section 7: Validation of putative *ABO*A1*-diagnostic variants in a multi-ethnic cohort

7.1 Blood allele group determination from whole-genome sequencing data

We aimed to study diagnostic accuracy of our discovered putatively *ABO*A1*-specific variants (Figure 1) in a larger and ethnically more diverse cohort. For this, we downloaded WGS data from 4,872 individuals participating in the Multi-Ethnic Study of Atherosclerosis (MESA)[17-19] from the database of Genotypes and Phenotypes (dbGaP, phs001416.v2.p1). Demographic information was available for 4,845 subjects (phs000209.v13.p3). First, we extracted the allele-defining variants for *ABO*A2*, *B*, *O.01*, and *O.02* (exact variants see SI Section 1.2) as well as the four *ABO*A1* candidate variants from the WGS data. The SNV rs115478735 of the compound dinucleotide variant rs1554760445 served as its proxy since the latter was not called in the dataset. Although the *ABO*A2*, *B*, *O.01*, and *O.02* defining variants were generally mutually exclusive, we statistically phased the variants with *Beagle* (v.5.2)[20] to allow better interpretation of the presence of more than two alleles per sample.

For interpretation purposes, we additionally extracted the variant differentiating *ABO*A1.02* from *A1.01* (c.467C>T, rs1053878), as well as further variants with minor allele counts ≥ 1, defining rare allele subgroups for *ABO*A2* (c.1054C>T, rs56390333; c.1009A>G, rs566015043), and *ABO*O* (c.628GTGGAC, rs782433608; c.542G>A, rs55727303; c.496del, rs563704490); and a splice donor variant (NG_006669.2:g.20397del, rs782023144). Finally, we also extracted the more frequent c.646T>A (rs8176740) and c.829G>A (rs8176748) variants, which act in combination with other variants and define an *ABO*O.09* allele.

7.2 Linkage disequilibrium between *ABO*A1*-candidate variants

We detected a substantially lower minor allele frequency (MAF) for the *ABO*A1*-candidate variant rs115478735 (MAF = 0.120) compared to the other three candidate variants rs532436 (MAF = 0.164), rs507666 (MAF = 0.164), and rs2519093 (MAF = 0.163). This corresponds to 21.7% of subjects carrying the minor allele of rs115478735 (952 hetero-, 107 homozygotes), while 29.7% (1293/153), 29.6% (1292/152), and 29.5% (1286/152) have minor alleles of rs532436, rs507666 and rs2519093, respectively. The three more frequent *ABO*A1*-candidate variants showed pairwise very high linkage disequilibrium (LD, r^2 = 0.97 to 1.00) and partial disagreement in only 25 samples. The less frequent variant rs115478735, as proxy for the compound *ABO*A1*-candidate variant rs1554760445, showed materially lower LD with the other three *ABO*A1*-candidates (r2 = 0.68 to 0.69).

7.3 Estimation of *ABO*A1* specificity and sensitivity

Table S7 shows predicted allele groups from the phased data and congruence with the observed presence of *ABO*A1* candidate variants. Specificity was calculated by the number of *ABO*A2*, *B*, *O.01*, or *O.02* alleles not containing *ABO*A1* candidate variants divided by the total number of *ABO*A2*, *B*, *O.01*, or *O.02* alleles (n = 8,155). Sensitivity was calculated as the ratio of the number of predicted *ABO*A1* alleles containing the *ABO*A1* candidate variants and the total number of predicted *ABO*A1* alleles (n = 1,589).

As for the three candidate SNVs in high LD, their presence on haplotypes with allele-defining variants for *ABO*A2*, *B*, *O.01,* or *O.02* (between n = 33 for rs2519093 and n = 48 for rs532436) suggested specificities between 99.4% and 99.6%. Due to lying in high LD, combining the SNVs did not increase specificity, which was only as good as the value for the best scoring variant rs2519093. In more detail, all the 33 predicted non-*ABO*A1* alleles with presence of *ABO*A1*-candidate SNV rs2519093 also contained the candidate SNVs rs507666 and rs532436.

Sensitivities for these three *ABO*A1*-candidate variants were between 97.5% for rs507666 and 98.0% for rs2519093. For the latter, there were 32 predicted *ABO*A1*-alleles that did not contain the variant. Sensitivity marginally improved to 98.4% when combining the three *ABO*A1* candidate SNVs (i.e. 25 alleles predicted as *A1* did not contain any of the three SNVs). The main reason for the slightly better sensitivity performance of rs2519093 compared to the other two candidate SNVs was related to uncertainties in phasing. For instance, the heterozygous minor allele of the candidate variant in a few predicted *A1*/*A2* samples was phased to the *A1* allele in the case of rs2519093, but to the *A2* allele in the case of the other two variants, hence lowering both, their sensitivity and specificity. As statistical phasing within a dataset like MESA without the availability of large population-based haplotype sequence collections is error-prone, we caution against inferring rs2519093 being superior to the other markers. In agreement, a look-up of publically available pairwise LD values between rs2519093 and either rs532436 or rs507666 in the multi-ethnical 1000 Genomes study (phase 3) revealed also slightly higher values (r2 = 0.99).

The less frequent candidate variant rs115478735 co-occurred with other allele-defining variants on only 23 haplotypes (specificity = 99.7%), but showed low sensitivity (71.9%) according to its absence on alleles not containing *ABO*A2*, *B*, *O.01,* or *O.02* defining variants. As rs115478735 was found to be mutually exclusive from the *ABO*A1.02* defining variant rs1053878 in predicted *ABO*A1* alleles, we hypothesized that it could specifically tag the *ABO*A1.01* allele subgroup, which is common in Europeans. Indeed, *A1*-sensitivity increased to 97.1% when only including samples of European ancestry (n = 1,755), which was in the range of overall sensitivities for the other three candidate variants. However, sensitivity of rs115478735 for tagging *ABO*A1.01* remained somewhat lower at 90.0%. We found that 77% of those putative *ABO*A1* alleles that were only supported by the three more frequent *ABO*A1* candidate variants (but not by the rarer rs115478735) could be assigned to *ABO*A1.02* by containing rs1053878. The remainder, i.e. those approx. 100 haplotypes that lower *ABO*A1.01* specificity, were alleles almost exclusively assigned to samples of African ethnicity. We concluded that our compound dinucleotide candidate variant rs1554760445 remained promising as *ABO*A1.01* tagging variant, but only in ethnicities of non-African descent.

7.4 Conservative accuracy estimates

Importantly, specificity and sensitivity values computed in this study are overall very conservative as estimated solely on the allele level. They would be materially higher (99.89% and above) if the deduced or serologically-typed ABO phenotype was considered, as normally done in blood group determination from genotype or sequencing data[21,22]. Reason is that *ABO*O* defining variants allow to assign an O allele group independent of co-occurring *ABO*A* or *B* defining variants as they impair allelic function. Presence of such co-occurring variants would hence not compromise phenotype specificity. For instance, we only find three alleles with the *ABO*A1* candidate variant rs2519093 coinciding with allele-defining variants for either *ABO*A2* or *ABO*B*, which suggests a phenotype specificity as high as 99.96%. The stringency of our allele-based method becomes also obvious when computing estimates for the diagnostic *ABO*B* SNV (rs8176747). Although regarded as entirely specific and sensitive for *ABO*B* within the community, this SNV shares a relatively high amount (n = 98) of alleles with the *ABO*O.01* variant in the MESA data (Table S7), which renders specificity to below 99% applying analogous calculation as for the *ABO*A1* candidates.

7.5 Rare *ABO*A* allele subgroups

Finally, we aimed to test the potential of our candidate variants to discriminate *ABO*A* allele subgroups (e.g. *A3*, *Ax*, A_{weak}, *Am*, *Ael*) from *A1*, but underlying variants were not present in sufficient number in MESA participants. We additionally considered low frequent *ABO*A2* and *O* alleles reported in ISBT tables (see SI Section 7.1)[23]. While rare *ABO*A2*-alleles were only found on *ABO*O.01* background, we indeed detected two presumable null alleles (*ABO*O.09* due to the presence of rs8176740 and rs8176748) that were originally predicted as *ABO*A1* alleles. Since they did not contain any of the *ABO*A1* candidate variants, their presence would slightly improve sensitivity (to 98.1% for the best scoring SNV rs2519093 and to 98.6% when combining the three candidate SNVs in LD). Nevertheless, other rare alleles may have been missed as we neither included variant combinations contributing to weak alleles nor could we include data on structural variants that had been reported in other multi-ethnic WGS studies[24]. The presence of hybrid alleles as well as phasing and sequencing errors are further issues that could have led to an underestimation of the diagnostic accuracy of the *ABO*A1* candidate variants.

Table S7. Validation of *ABO*A1* candidate variants using genotype prediction from MESA sequencing data. *ABO* genotypes were predicted by the presence of *ABO* allele-defining variants. Phasing information showing on which allele the *ABO*A1* candidate variant was present (1) or absent (0) is provided in brackets next to the sample counts. For homozygous samples, phasing information is not provided because irrelevant (displayed by "/"). **Part (I)** contains all samples predicted to have no *ABO*A1* alleles due to the presence of causative *ABO*A2*, *B*, *O.01* or *O.02* variants on both alleles. Bold numbers represent discrepant samples for which at least one allele additionally harbored an *ABO*A1* candidate variant. Total number of alleles (2*3,435) as well as discrepant alleles contribute to the specificity calculation for each *ABO*A1* candidate variant. **Part (II)** contains all samples predicted to have one *ABO*A1* allele due to the absence of causative *ABO*A2*, *B*, *O.01* or *O.02* variants on one allele. Bold numbers represent discrepant samples for which the allele with the causative variant or neither of the two alleles harbored an *ABO*A1* candidate variant. Total number of alleles (2*1,285) is assigned at equal parts to specificity and sensitivity. Alleles from discrepant cases contribute to specificity (if *ABO*A1* candidate variant is present on first allele, i.e. 1|0 or 1|1) and/or to sensitivity (if *ABO*A1* candidate variant is absent on second allele, i.e. 1|0 or 0|0). **Part (III)** contains all samples predicted to have two *ABO*A1* alleles due to the absence of causative *ABO*A2*, *B*, *O.01* or *O.02* variants on both alleles. Bold numbers represent discrepant samples for which at least one allele did not contain an *ABO*A1* candidate variant. Total number of alleles (2*152) as well as discrepant alleles contribute to the sensitivity calculation for each *ABO*A1* candidate variant.

<table>
<tr><th>Genotype[a]</th><th>N</th><th>rs532436</th><th>rs507666</th><th>rs2519093</th><th>rs115478735[b]</th></tr>
<tr><td colspan="6">(I) Samples predicted to lack ABO*A1 candidate variants</td></tr>
<tr><td>A2 | A2</td><td>8</td><td>8 (0|0)</td><td>8 (0|0)</td><td>8 (0|0)</td><td>8 (0|0)</td></tr>
<tr><td>B | B</td><td>53</td><td>53 (0|0)</td><td>53 (0|0)</td><td>53 (0|0)</td><td>53 (0|0)</td></tr>
<tr><td>O.01 | O.01</td><td>2171</td><td>2154 (0|0)
17 (0/1)</td><td>2154 (0|0)
17 (0/1)</td><td>2155 (0|0)
16 (0/1)</td><td>2167 (0|0)
4 (0/1)</td></tr>
<tr><td>O.02 | O.02</td><td>1</td><td>1 (0|0)</td><td>1 (0|0)</td><td>1 (0|0)</td><td>1 (0|0)</td></tr>
<tr><td>A2 | B</td><td>58</td><td>58 (0|0)</td><td>58 (0|0)</td><td>58 (0|0)</td><td>58 (0|0)</td></tr>
<tr><td>A2 | B O.01</td><td>5</td><td>5 (0|0)</td><td>5 (0|0)</td><td>5 (0|0)</td><td>5 (0|0)</td></tr>
<tr><td>A2 | O.01</td><td>345</td><td>344 (0|0)
1 (0|1)</td><td>344 (0|0)
1 (0|1)</td><td>344 (0|0)
1 (0|1)</td><td>344 (0|0)
1 (0|1)</td></tr>
<tr><td>A2 | O.02</td><td>5</td><td>3 (0|0)
2 (0|1)</td><td>3 (0|0)
2 (0|1)</td><td>3 (0|0)
2 (0|1)</td><td>3 (0|0)
2 (0|1)</td></tr>
<tr><td>A2 O.01 | B</td><td>1</td><td>1 (0|0)</td><td>1 (0|0)</td><td>1 (0|0)</td><td>1 (0|0)</td></tr>
<tr><td>A2 O.01 | O.01</td><td>1</td><td>1 (0|0)</td><td>1 (0|0)</td><td>1 (0|0)</td><td>1 (0|0)</td></tr>
<tr><td>B | B O.01</td><td>6</td><td>6 (0|0)</td><td>6 (0|0)</td><td>6 (0|0)</td><td>6 (0|0)</td></tr>
<tr><td>B | O.01</td><td>634</td><td>632 (0|0)
2 (0|1)</td><td>632 (0|0)
2 (0|1)</td><td>632 (0|0)
2 (0|1)</td><td>633 (0|0)
1 (0|1)</td></tr>
<tr><td>B | O.02</td><td>13</td><td>13 (0|0)</td><td>13 (0|0)</td><td>13 (0|0)</td><td>13 (0|0)</td></tr>
<tr><td>B O.01 | B O.01</td><td>3</td><td>3 (0|0)</td><td>3 (0|0)</td><td>3 (0|0)</td><td>3 (0|0)</td></tr>
<tr><td>B O.01 | O.01</td><td>62</td><td>61 (0|0)
1 (0|1)</td><td>61 (0|0)
1 (0|1)</td><td>62 (0|0)</td><td>61 (0|0)
1 (0|1)</td></tr>
<tr><td>O.01 | O.02</td><td>69</td><td>65 (0|0)
1 (0|1)
3 (1|0)</td><td>65 (0|0)
1 (0|1)
3 (1|0)</td><td>66 (0|0)
1 (0|1)
2 (1|0)</td><td>66 (0|0)
1 (0|1)
2 (1|0)</td></tr>
<tr><td>∑ (discrepancies)</td><td>3,435</td><td>**27 (0.79%)**</td><td>**27 (0.79%)**</td><td>**24 (0.70%)**</td><td>**12 (0.35%)**</td></tr>
<tr><td colspan="6">(II) Samples predicted to harbor ABO*A1 candidate variants on one allele</td></tr>
<tr><td>A2 | A1</td><td>84</td><td>75 (0|1)
7 (1|0)
1 (0|0)</td><td>75 (0|1)
7 (1|0)
1 (0|0)</td><td>81 (0|1)
1 (1|0)
1 (0|0)</td><td>60 (0|1)
5 (1|0)
18 (0|0)</td></tr>
</table>

		1 (1|1)	**1 (1|1)**	**1 (1|1)**	**1 (1|1)**
A2 O.01 | A1	1	1 (0|1)	1 (0|1)	1 (0|1)	**1 (0|0)**
B | A1	168	163 (0|1) **4 (0|0)** **1 (1|1)**	162 (0|1) **5 (0|0)** **1 (1|1)**	162 (0|1) **5 (0|0)** **1 (1|1)**	97 (0|1) **71 (0|0)**
B O.01 | A1	19	18 (0|1) **1 (0|0)**	18 (0|1) **1 (0|0)**	18 (0|1) **1 (0|0)**	7 (0|1) **12 (0|0)**
O.01 | A1	990	966 (0|1) **6 (1|0)** **12 (0|0)** **6 (1|1)**	967 (0|1) **5 (1|0)** **13 (0|0)** **5 (1|1)**	968 (0|1) **1 (1|0)** **16 (0|0)** **5 (1|1)**	725 (0|1) **3 (1|0)** **260 (0|0)** **2 (1|1)**
O.02 | A1	23	23 (0|1)	23 (0|1)	23 (0|1)	22 (0|1) **1 (0|0)**
∑ (discrepancies)	1,285	**39 (3.04%)**	**39 (3.04%)**	**32 (2.49%)**	**374 (29.11%)**
(III) Samples predicted to harbor *ABO*A1* candidate variants on both alleles					
A1 | A1	152	145 (1|1) **7 (0/1)**	145 (1|1) **7 (0/1)**	145 (1|1) **7 (0/1)**	104 (1|1) **21 (0/1)** **27 (0|0)**
∑ (discrepancies)	152	**7 (4.61%)**	**7 (4.61%)**	**7 (4.61%)**	**48 (31.58%)**
Accuracy measures					
Specificity	8,155 alleles	**99.41%**	**99.44%**	**99.60%**	**99.72%**
Sensitivity	1,589 alleles	**97.61%**	**97.55%**	**97.99%**	**71.93%**

[a] *ABO* allele-defining variants: *ABO*A2* (c.1061delC, rs56392308), *ABO*B* (c.803G>C, rs8176747), *ABO*O.01* (c.261delG, rs8176719), *ABO*O.02* (c.802G>A, rs41302905); [b] The SNV rs115478735 was used as proxy for the dinucleotide variant rs1554760445.

References

1. Gassner C, Degenhardt F, Meyer S, et al. Low-frequency blood group antigens in Switzerland. *Transfusion Medicine and Hemotherapy.* 2018;45(4):239-250.

2. 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. *Transfusion Medicine Reviews.* 2013;27(1):2-9.

3. Barbitoff YA, Bezdvornykh IV, Polev DE, et al. Catching hidden variation: systematic correction of reference minor allele annotation in clinical variant calling. *Genetics in Medicine.* 2018;20(3):360-364.

4. Ballouz S, Dobin A, Gillis JA. Is it time to change the reference genome? *Genome biology.* 2019;20(1):1-9.

5. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research.* 2017;27(5):722-736.

6. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34(18):3094-3100.

7. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics.* 2013;14(2):178-192.

8. Shafin K, Pesout T, Chang P-C, et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nature methods.* 2021;18(11):1322-1332.

9. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *Gigascience.* 2021;10(2):giab008.

10. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics.* 2009;25(14):1754-1760.

11. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research.* 2010;20(9):1297-1303.

12. Van der Auwera GA, O'Connor BD. *Genomics in the cloud: using Docker, GATK, and WDL in Terra*: O'Reilly Media; 2020.

13. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics.* 2009;25(11):1451-1452.

14. Nei M. *Molecular evolutionary genetics*: Columbia university press; 1987.

15. Tajima F. Evolutionary relationship of DNA sequences in finite populations. *Genetics.* 1983;105(2):437-460.

16. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus evolution.* 2015;1(1).

17. Bild DE, Bluemke DA, Burke GL, et al. Multi-ethnic study of atherosclerosis: objectives and design. *American journal of epidemiology.* 2002;156(9):871-881.

18. Burke G, Lima J, Wong ND, Narula J. The Multiethnic Study of Atherosclerosis. *Global Heart.* 2016;11(3):267-268.

19. Olson JL, Bild DE, Kronmal RA, Burke GL. Legacy of MESA. *Global Heart.* 2016;11(3):269-274.

20. Browning BL, Tian X, Zhou Y, Browning SR. Fast two-stage phasing of large-scale sequence data. *The American Journal of Human Genetics.* 2021;108(10):1880-1890.

21. 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.

22. Lane WJ, Westhoff CM, Gleadall NS, et al. Automated typing of red blood cell and platelet antigens: a whole-genome sequencing study. *The Lancet Haematology.* 2018;5(6):e241-e251.

23. The International Society of Blood Transfusion. ABO (ISBT 001) Blood Group Allele Table. https://www.isbtweb.org/fileadmin/user_upload/Working_parties/WP_on_Red_Cell_Immunogenetics_and/001_ABO_Alleles_v1.2.pdf.

24. 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 advances.* 2016;1(3):240-249.