Automated typing of red blood cell and platelet antigens: a whole-genome sequencing study

William J Lane, Connie M Westhoff*, Nicholas S Gleadall*, Maria Aguad, Robin Smeland-Wagman, Sunitha Vege, Daimon P Simmons, Helen H Mah, Matthew S Lebo, Klaudia Walter, Nicole Soranzo, Emanuele Di Angelantonio, John Danesh, David J Roberts, Nick A Watkins, Willem H Ouwehand, Adam S Butterworth, Richard M Kaufman, Heidi L Rehm, Leslie E Silberstein*, Robert C Green*, on behalf of the MedSeq Project†*

Summary

Background There are more than 300 known red blood cell (RBC) antigens and 33 platelet antigens that differ between individuals. Sensitisation to antigens is a serious complication that can occur in prenatal medicine and after blood transfusion, particularly for patients who require multiple transfusions. Although pre-transfusion compatibility testing largely relies on serological methods, reagents are not available for many antigens. Methods based on single-nucleotide polymorphism (SNP) arrays have been used, but typing for ABO and Rh—the most important blood groups—cannot be done with SNP typing alone. We aimed to develop a novel method based on whole-genome sequencing to identify RBC and platelet antigens.

Methods This whole-genome sequencing study is a subanalysis of data from patients in the whole-genome sequencing arm of the MedSeq Project randomised controlled trial (NCT01736566) with no measured patient outcomes. We created a database of molecular changes in RBC and platelet antigens and developed an automated antigen-typing algorithm based on whole-genome sequencing (bloodTyper). This algorithm was iteratively improved to address cis–trans haplotype ambiguities and homologous gene alignments. Whole-genome sequencing data from 110 MedSeq participants (30×depth) were used to initially validate bloodTyper through comparison with conventional serology and SNP methods for typing of 38 RBC antigens in 12 blood-group systems and 22 human platelet antigens. bloodTyper was further validated with whole-genome sequencing data from 200 INTERVAL trial participants (15×depth) with serological comparisons.

Findings We iteratively improved bloodTyper by comparing its typing results with conventional serological and SNP typing in three rounds of testing. The initial whole-genome sequencing typing algorithm was 99·5% concordant across the first 20 MedSeq genomes. Addressing discordances led to development of an improved algorithm that was 99·8% concordant for the remaining 90 MedSeq genomes. Additional modifications led to the final algorithm, which was 99·2% concordant across 200 INTERVAL genomes (or 99·9% after adjustment for the lower depth of coverage).

Interpretation By enabling more precise antigen-matching of patients with blood donors, antigen typing based on whole-genome sequencing provides a novel approach to improve transfusion outcomes with the potential to transform the practice of transfusion medicine.

Funding National Human Genome Research Institute, Doris Duke Charitable Foundation, National Health Service Blood and Transplant, National Institute for Health Research, and Wellcome Trust.

Introduction

Exposure to non-self red blood cell (RBC) and platelet antigens during transfusion or pregnancy can lead to the development of alloantibodies that can cause mortality and morbidity. Although transfusion-related deaths are rare, about 15% of deaths associated with blood transfusions reported each year are the result of haemolytic transfusion reactions due to blood-group antibodies.[1] Additionally, sensitisation to foreign RBC antigens results in a lifetime risk of delayed or acute haemolytic transfusion reactions, fetal anaemia, and complications in pregnancy.[2] For patients who require chronic transfusion, this sensitisation increases the cost and turnaround time of each subsequent transfusion. Similarly, sensitisation to foreign platelet antigens can be life-threatening because of ineffective platelet transfusion, and can also result in thrombocytopenia of the fetus and newborn, with a risk of intracranial haemorrhage.[2]

Antigen typing and matching of recipients and blood donors for more than the traditional ABO and RhD blood-group antigens (termed extended antigen matching), which can avoid primary sensitisation and improve transfusion safety,[1] is not currently standard of practice. Extended antigen typing with antibody-based serological methods is labour intensive and costly, and reagent antibodies are not available for many clinically important blood-group antigens. DNA array methods that sample single nucleotide polymorphisms (SNPs) have been used for extended blood-group typing and overcome some limitations of serological typing methods.[3] However, SNP

Lancet Haematol 2018

Published Online
May 17, 2018
http://dx.doi.org/10.1016/S2352-3026(18)30053-X

See Online/Comment
http://dx.doi.org/10.1016/S2352-3026(18)30064-4

*Contributed equally

†All other members of the MedSeq Project are listed in appendix p 3

Department of Pathology (W J Lane MD, M Aguad MS, R Smeland-Wagman BS, D P Simmons MD, H H Mah MS, M S Lebo PhD, R M Kaufman MD), **Division of Transfusion Medicine** (Prof L E Silberstein MD), **and Division of Genetics, Department of Medicine** (Prof R C Green MD), **Brigham and Women's Hospital, Boston, MA, USA; Harvard Medical School, Boston, MA, USA** (W J Lane, M S Lebo, H L Rehm PhD, Prof L E Silberstein, Prof R C Green); **New York Blood Center, New York, NY, USA** (C M Westhoff PhD, S Vege MS); **Department of Haematology** (N S Gleadall BSc, Prof N Soranzo PhD, Prof W H Ouwehand FMedSci), **Medical Research Council and British Heart Foundation Cardiovascular Epidemiology Unit** (E Di Angelantonio MD, Prof J Danesh FMedSci, A S Butterworth PhD) **and National Institute for Health Research Blood and Transplant Research Unit in Donor Health and Genomics** (E Di Angelantonio, Prof J Danesh, Prof D J Roberts DPhil, A S Butterworth), **Department of Public Health and Primary Care, and British Heart Foundation Cambridge Centre of Excellence, Department of Medicine** (Prof J Danesh, Prof W H Ouwehand), **University of Cambridge, Cambridge, UK; National Health Service (NHS)**

Blood and Transplant, Cambridge, UK (N S Gleadall, N A Watkins DPhil, Prof W H Ouwehand); **Laboratory for Molecular Medicine, Boston, MA, USA** (M S Lebo, H L Rehm); **Partners Personalized Medicine, Boston, MA, USA** (M S Lebo, H L Rehm, Prof R C Green); **Wellcome Trust Sanger Institute, Hinxton, UK** (K Walter PhD, Prof N Soranzo, Prof J Danesh); **Cambridge Substantive Site, Health Data Research UK, Wellcome Genome Campus, Hinxton, UK** (Prof N Soranzo, E Di Angelantonio, Prof J Danesh, A S Butterworth); **NHS Blood and Transplant–Oxford Centre, Oxford, UK** (Prof D J Roberts, Prof W H Ouwehand); **Biomedical Research Centre Haematology Theme and Radcliffe Department of Medicine, University of Oxford, Oxford, UK** (Prof D J Roberts); **Department of Medicine, Massachusetts General Hospital, Boston, MA, USA** (H L Rehm); **and Broad Institute of Massachusetts Institute of Technology and Harvard, Boston, MA, USA** (H L Rehm, Prof R C Green)

Correspondence to: Dr William J Lane, Department of Pathology, Brigham and Women's Hospital, Boston, MA 02115, USA
wlane@bwh.harvard.edu

See Online for appendix

Research in context

Evidence before this study

We searched PubMed between January, 2000, and December, 2012, using the search terms "whole genome sequencing", "next generation sequencing", "blood groups", "RBC antigens", and "platelet antigens" to identify studies that used next-generation sequencing to type for red blood cell (RBC) or platelet antigens. We identified one review paper suggesting that RBC antigen typing with next-generation sequencing has future potential as a blood-typing method, and one research paper piloting the use of next-generation sequencing in RhD typing. Use of next-generation sequencing for antigen typing in clinical practice will require development of algorithms that are capable of accurately translating next-generation sequencing data into antigen phenotypes.

Added value of this study

We curated a database of molecular changes in RBC and platelet antigens, from which we developed an automated antigen-typing software based on whole-genome sequencing (bloodTyper). The performance of bloodTyper was evaluated and sequentially improved by use of whole-genome sequencing data from 110 individuals from the whole-genome sequencing group of the MedSeq Project randomised controlled trial (30×depth), and then validated in 200 genomes from the INTERVAL study (15×depth). The final algorithm was 99·8% concordant with conventional antigen typing by serology (17 antigens in six blood-group systems) and DNA-based single nucleotide polymorphism assays (35 antigens in 11 blood group systems). In the validation analysis, bloodTyper was 99·2% concordant with conventional antigen typing by serology (21 antigens in seven blood-group systems).

Implications of all the available evidence

Typing of RBC and platelet antigens with whole-genome sequencing has the potential to be used in routine clinical practice to predict extended blood-group antigen profiles; however, further investigation is needed. The results of this study suggest that bloodTyper could be used as a comprehensive and accurate approach to improve transfusion typing, and therefore safety.

approaches do not target all blood groups, detect all inactive (null) alleles and complex gene rearrangements, or reliably ascertain alleles encoding ABO and Rh, the major blood groups.[4]

Next-generation sequencing, particularly whole-genome sequencing, might overcome these limitations through providing accurate high-resolution typing in pre-transfusion antibody screening, enabling routine prophylactic extended blood-group matching whenever possible. However, in the absence of computerised algorithms capable of robust interpretation of RBC and platelet antigens directly from next-generation sequencing data, the translation to antigen phenotypes is time intensive and requires considerable expertise.[5–12]

In a proof-of-principle analysis,[7] we showed that a subject-matter expert could analyse whole-genome sequencing data to comprehensively assess RBC and platelet antigens. Subsequently, we hypothesised that it would be possible to create automated antigen-typing software based on whole-genome sequencing that would be concordant with conventional typing assays based on serology and SNPs. In this study, we aimed to develop and validate such software.

Methods

Study design and participants

This whole-genome sequencing study is a subanalysis of the MedSeq Project randomised controlled trial.[13–18] Participants were ten primary care physicians and 100 of their healthy patients, and ten cardiologists and 100 of their patients who had hypertrophic cardiomyopathy or dilated cardiomyopathy, from a single centre in Boston, MA, USA. In the trial, participants were randomly assigned to undergo either standardised family history assessment plus whole-genome sequencing (n=100) or standardised family history assessment alone (n=100; control group). To expand the ethnic diversity of the evaluated genetic changes, in 2016–17, an additional ten African–American individuals were recruited to the whole-genome sequencing group as part of an extension phase.

Participants were eligible for enrolment to the MedSeq study if they were aged 18–90 years and did not have cardiac disease (other than hypertrophic or dilated cardiomyopathy in those patients enrolled by participating cardiologists), diabetes, progressive debilitating illness, or untreated clinical anxiety or depression (as indicated by a Hospital Anxiety and Depression Scale score >11 at baseline), and were not pregnant. The MedSeq Project whole-genome sequencing data are available through the database of Genotypes and Phenotypes (accession number phs000958; appendix p 4). Participants were informed of the risks of whole-genome sequencing and possible results before providing written informed consent.

In this substudy we assessed 110 participants enrolled to the whole-genome sequencing group of the MedSeq Project between Dec 19, 2012, and Jan 26, 2017, without any intended comparison with the control group. Instead, we compared whole-genome sequencing with conventional serological and SNP methods for typing of RBC and platelet antigens within participants. The self-identified ethnicities of these participants were European ancestry (n=89), African ancestry (n=13), Asian (n=4), and Hispanic (n=4; appendix p 4). RBC and platelet extended antigen profiles were summarised for each participant and provided to their physician.[14]

We also used 220 genomes of European ancestry with matching RBC serological phenotyping from the INTERVAL study.[19] The initial 20 genomes were used for technical troubleshooting, and the remaining 200 were used as an external dataset for validation studies (appendix p 5).

Database of antigen alleles

A comprehensive curated database of RBC and platelet antigen alleles was created by subject-matter experts (WJL with input from CMW and SV) using published sources.[20–25] During the curation process, errors and omissions in the published antigen alleles were detected by manually comparing the published sources with automated cross-correlation checks (eg, agreement between nucleotide change and position and aminoacid identity and position). Our previously published, semi-automated, nucleotide-position-conversion process[7] was fully automated and used to convert all relevant nucleotide positions in complementary (c)DNA (eg, 578T in *KEL* cDNA; GenBank sequence M64934) to genomic DNA coordinates (eg, chr7:142 655 008T in human reference genome GRCh37/hg19). We will update the database as new blood-group systems and alleles are officially assigned by the international antigen workgroups.

For more on the **antigen allele database** see http://bloodantigens.com

Serological antigen typing, SNP typing, and *RHD* zygosity testing

Figure 1 shows a flowchart for this study. Blood samples in edetic acid were collected from the MedSeq Project participants between Dec 19, 2012, and Jan 26, 2017, and conventional RBC serological antigen typing was done according to standard in-vitro blood-typing methods.[2] Serological typing reagents from Bio-Rad (Hercules, CA, USA) were used to identify ABO, M, N, S, s, D, C, c, E, e, K, k, Fy^a, Fy^b, Jk^a, and Jk^b antigens. Additionally, we identified Fy^b using serological typing reagents from Ortho Clinical Diagnostics (Raritan, NJ, USA) and Jk^a and Jk^b using serological typing reagents from Immucor (Norcross, GA, USA).

DNA was isolated from the white blood cells with standard methods, and the PreciseType BeadChip HEA (human erythrocyte antigen) array (Immucor) was used to detect SNPs in 35 RBC antigens: M, N, S, s, U, C, c, E, e, V, VS, Lu^a, Lu^b, K, k, Kp^a, Kp^b, Js^a, Js^b, Fy^a, Fy^b, Jk^a, Jk^b, Di^a, Di^b, Sc1, Sc2, Do^a, Do^b, Hy, Jo^a, Co^a, Co^b, LW^a, and LW^b. The Immucor BioArray HPA (human platelet antigen) BeadChip array (Immucor, Norcross, GA, USA) was used to detect 22 human platelet antigens (HPAs): HPA-1a, HPA-1b, HPA-2a, HPA-2b, HPA-3a, HPA-3b, HPA-4a, HPA-4b, HPA-5a, HPA-5b, HPA-6aw, HPA-6bw, HPA-7aw, HPA-7bw, HPA-8aw, HPA-8bw, HPA-9aw, HPA-9bw, HPA-11aw, HPA-11bw, HPA-15a, and HPA-15b.

RBC serological typing data for participants in the INTERVAL study[19] were extracted from the UK National Health Service Blood and Transplant's PULSE blood bank control and management system. Specifically, we extracted data on ABO, M, N, S, s, D, C, c, E, e, Lu^a, Lu^b, K, k, Kp^a, Kp^b, Fy^a, Fy^b, Jk^a, and Jk^b antigens.

We did conventional *RHD* zygosity testing on a subset of MedSeq Project blood samples using the hybrid box assay, according to previously published methods.[26] Briefly, allele-specific PCR was done with primers designed to amplify a 1507-bp product within the hybrid box sequence (appendix p 5).[26] PCR products were visualised via agarose-gel electrophoresis with ethidium bromide staining. Participants were defined as homozygous for *RHD* when serological RhD was positive and no hybrid box was present, as hemizygous for *RHD* when serological RhD was positive and the hybrid box was present, and *RHD* negative or null when serological RhD was negative and the hybrid box was present.

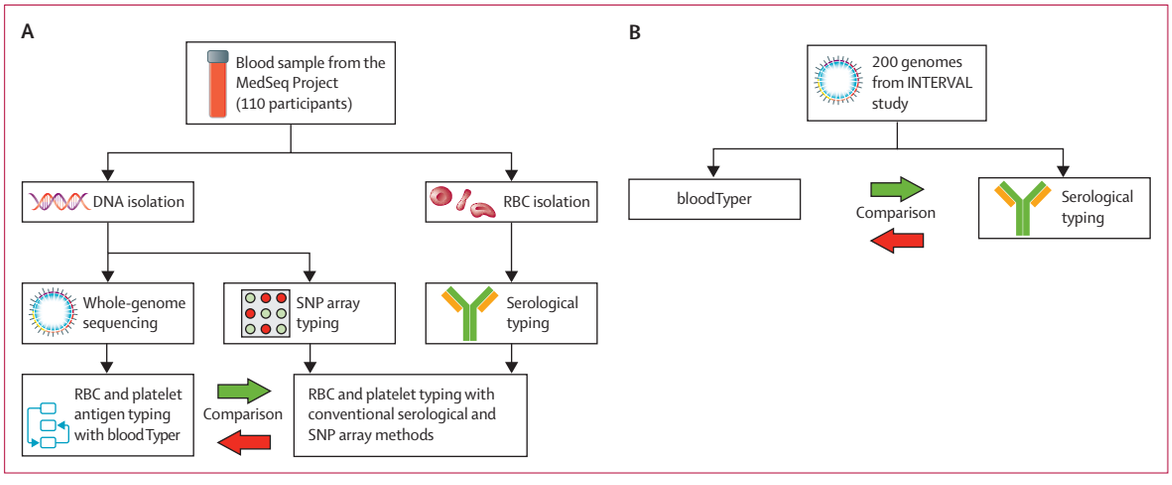

Figure 1: **Flow diagram**
(A) Process for development of the typing algorithm based on whole-genome sequencing. (B) Process for validation of the final bloodTyper algorithm in INTERVAL genomes.[19] RBC=red blood cell. WGS=whole-genome sequencing. SNP=single nucleotide polymorphism.

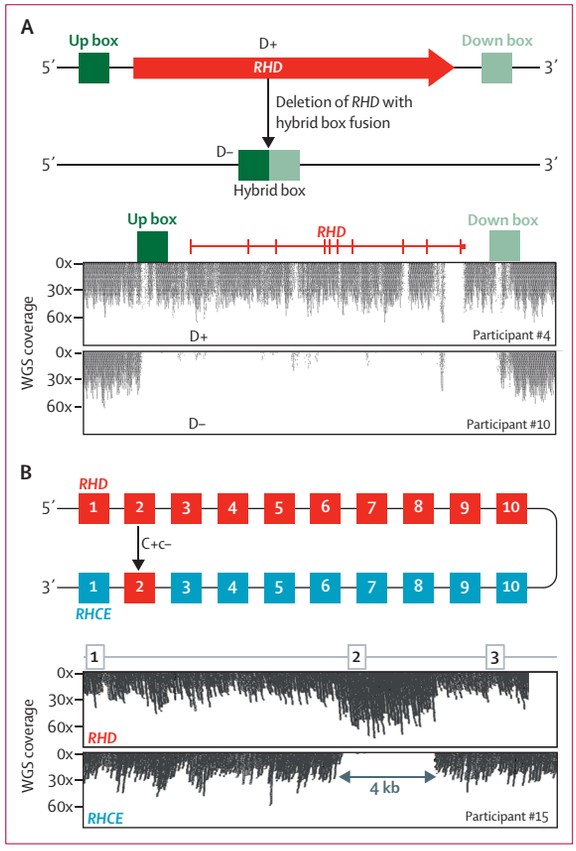

Figure 2: **Rh typing algorithm considerations**
(A) Absence of the D antigen is most commonly caused by deletion of the *RHD* gene, which results in fusion of the upstream and downstream Rhesus boxes into a hybrid box. In Rh D-negative individuals, this fusion leads to loss of WGS sequence reads over the *RHD* gene region. (B) Presence of the C antigen results in misalignment of *RHCE* exon 2 (loss of sequence reads) to *RHD* exon 2 (gain of sequence reads). WGS=whole-genome sequencing.

Whole-genome sequencing

The whole-genome sequencing workflow in the MedSeq Project randomised controlled trial, including methods for genomic DNA isolation and quality control, has been previously described.[13,14] Briefly, whole-genome sequencing was done on blood samples with the Illumina HiSeq 2000 platform (Illumina, San Diego, CA, USA). Genomes were sequenced using 100-bp paired-end reads to a depth of coverage of 30×.[27] Sequence read data were aligned with the human reference sequence (GRCh37/hg19), and the alignments were processed to remove duplicates, recalibrate, and realign around indels.

In addition to the MedSeq Project whole-genome sequencing data, we used whole-genome sequencing data (unpublished) from the 220 INTERVAL participants who had the largest number of serologically typed antigens. We used data from 20 of these individuals for initial technical troubleshooting and from 200 individuals for algorithm validation. INTERVAL blood samples were sequenced to a coverage of 15× depth at the Wellcome Sanger Institute (Hinxton, UK) on the Illumina HiSeq X platform (Illumina, San Diego, CA, USA). The raw sequencing reads were converted directly into sequence alignment files (BAM format) with Illumina2BAM version 1.03 and, after several quality control steps to remove duplicates, converted into compressed sequence alignment files (CRAM format). To standardise the INTERVAL genomes with those from the MedSeq Project, sequence reads were extracted from the INTERVAL CRAMs and aligned as GRCh37/hg19 BAMs.

The same general workflow was used to predict RBC and platelet antigens from both the MedSeq and INTERVAL BAM files. Variant calls for 45 RBC and six platelet genes (listed in appendix p 6) were made for each exon, 300 bp upstream of the start codon, and 10 bp into each intron with the Genome Analysis Tool Kit version 2.3-9-gdcdccbb. These variants were saved as a variant calling format file that showed differences between the whole-genome sequencing data and the reference genome.[28] Sequencing coverage was extracted from the alignment file with BEDTools version 2.17.0.20.[29] The Integrative Genomics Viewer was used as needed to verify coverage and sequence identity.[30]

Copy number analysis

The Rh blood-group system comprises the homologous genes *RHD* and *RHCE*, which can be problematic for whole-genome sequencing alignment algorithms. To detect misalignment, we ascertained the copy number for each exon and intron within *RHD* and *RHCE* using a depth-of-coverage approach in which the copy number for a particular region was equal to the average coverage of a region divided by the average background coverage, multiplied by two. The average coverage across the *RHCE* gene was used as the background coverage because two copies of *RHCE* are usually present.

We used the copy number calculations of the introns and exons to detect structural changes associated with the presence or absence of D and C antigens. Absence of the D antigen is most commonly caused by deletion of the *RHD* gene, which is indicated by an absence of whole-genome sequence reads across the *RHD* gene region (figure 2A). The C antigen occurs when exon 2 of *RHCE* is replaced by exon 2 of *RHD*, causing exon 2 of the *RHCE* C antigen to misalign to *RHD* exon 2, indicated by an increase in aligned *RHD* exon 2 sequences and an absence of *RHCE* exon 2 sequences (figure 2B). D antigen (*RHD*) copy number (zygosity) was calculated as the average coverage for *RHD* divided by the average coverage for *RHCE*, multiplied by two. Samples were homozygous for *RHD* if the copy number was 1·6–2·5, hemizygous if the copy number was 0·6–1·5, and null or negative if the copy number was 0·0–0·5.

The copy number of C antigen was calculated on the basis of misalignment of *RHCE* exon 2 (ie, loss of sequence reads aligned to *RHCE*), and was equal to the

average coverage across C antigen divided by the average coverage across *RHCE*, multiplied by two. We evaluated two different regions of C antigen: *RHCE* exon 2 only (0·2 kb), and *RHCE* exon 2 and parts of the surrounding introns (4 kb). C antigen genotype was assigned with copy number ranges: less than 0·5 (C antigen positive, c antigen negative), 0·5–1·4 (C antigen positive, c antigen positive), and 1·5 or higher (C antigen negative, c antigen positive).

Whole-genome sequencing typing algorithm

We designed and implemented the whole-genome sequencing typing algorithm (bloodTyper) using custom-made typing software that was iteratively improved during the study (figure 3). Whole-genome sequencing data from the first 20 MedSeq participants were typed with an initial algorithm, and the typing results were compared with those of conventional serological and SNP typing methods for 38 RBC and 22 platelet antigens (encoded by 17 RBC and six platelet genes, respectively) to guide improvement of the algorithm.

The improved algorithm was then used on blood samples from the remaining 90 MedSeq participants, and the typing results were compared with serological (MA and RS-W) and SNP (SV) typing results. Discordances between the results of the different typing methods were investigated, and an updated final algorithm was created that used a combination of gene sequence, sequence coverage, copy number analysis, and misalignment detection to select the correct antigen alleles. These alleles were then integrated to determine the antigen phenotype. The final algorithm was then validated on 200 genomes from the INTERVAL study,[19] in typing of 21 RBC antigens encoded by 14 genes.

Although we used the whole-genome sequencing typing algorithm to evaluate all antigens with known DNA sequences, it was only possible to assess the performance of our algorithm for typing of antigens that have commonly available serological reagents and those covered by the SNP arrays.

Statistical analysis

The MedSeq Project was designed as a pilot randomised controlled trial, and exploratory statistics were used to compare outcomes between the groups that did and did not undergo whole-genome sequencing.[13–18] In this substudy we used data from participants assigned to the whole-genome sequencing group, without any intended comparison with the control group, to compare whole-genome sequencing versus conventional serological and SNP methods for typing of RBC and platelet antigens in the same individual. We used Excel version 15.33 to calculate performance statistics for this comparison, including sensitivity, specificity, positive predictive value, negative predictive value, and accuracy. The MedSeq Project is registered with ClinicalTrials.gov, number NCT01736566.

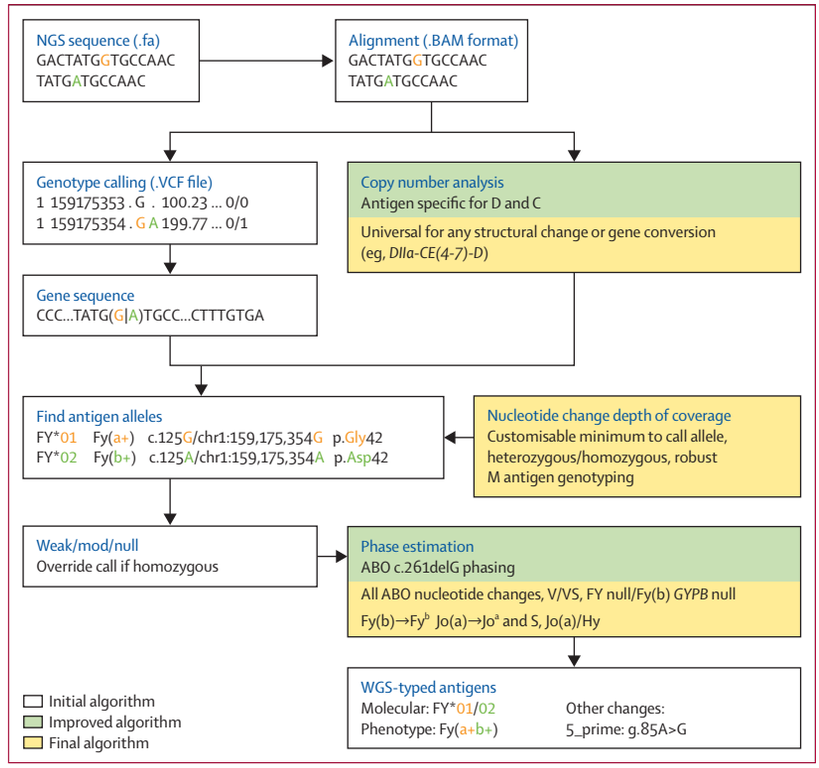

Figure 3: **bloodTyper algorithm**
Orange text indicates Fy^a antigen changes. Green text indicates Fy^b antigen changes. NGS=next-generation sequencing. WGS=whole-genome sequencing.

Role of the funding source

The sponsors of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

Similar to our previous findings from whole-genome-based typing of one individual,[7] we found a few regions of low coverage (*RHD* [exon 8], *C4B*, *C4A*, and *CR1*). However, all MedSeq Project genomes had adequate sequencing coverage to allow antigen typing from the relevant nucleotide positions at a calling cutoff for each nucleotide of 4× coverage, except for antigens involving Rh gene conversions (eg, C antigen) and the M antigen. Rh gene conversions caused misalignment of whole-genome sequences, which we were able to identify using copy number analysis. The M antigen—which is defined by three nucleotide changes in exon 2 of *GYPA*—required a calling cutoff of 2× coverage at each nucleotide. The lower M antigen coverage could be due to inefficient whole-genome sequence alignment because the human reference genome encodes the N antigen, which is defined by three different nucleotides at the same positions as the M antigen.

Complete typing results on all 110 MedSeq participants are shown in the appendix (pp 8–23). We used the first

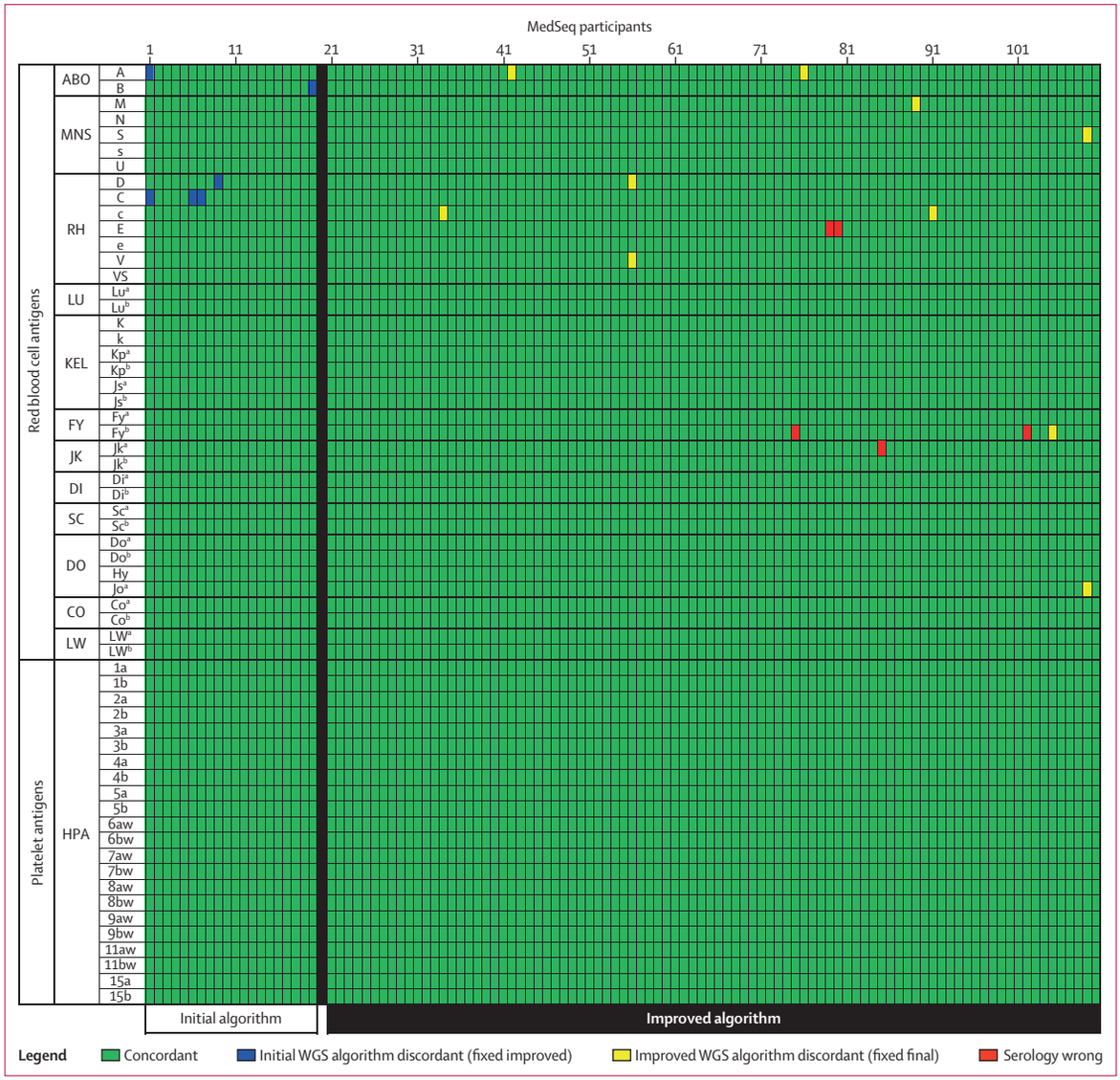

Figure 4: **Concordance of WGS antigen typing with serological and SNP typing methods in 110 MedSeq participants**
WGS=whole-genome sequencing. SNP=single nucleotide polymorphism. HPA=human platelet antigen.

20 MedSeq genomes to find errors in the curated database and to design an algorithm to translate whole-genome sequencing data into the corresponding RBC and platelet antigens. Compared with conventional serology and SNP typing methods, the initial algorithm was discordant for ABO antigens in two individuals, C antigen in three individuals, and D antigen in one individual (figure 4, table; appendix p 24). The algorithm made 1194 correct calls out of 1200 individual antigen typings, giving a concordance of 99·5%. The performance statistics for the initial algorithm according to the antigen are shown in the appendix (pp 25, 26).

To improve the algorithm, we analysed the cause of each discordant call and modified the algorithm accordingly (table, figure 3). The major changes responsible for ABO transferase specificity and activity are in exons 6 and 7 of the *ABO* gene.[20] Although changes in exon 7 span only 277 bp, the changes in exons 6 and 7 are separated by a 1316-bp intron.[20] Therefore, because the average size of whole-genome sequencing fragments is 300 bp, we could not use the sequence reads to directly determine the haplotype phase of exons 6 and 7 (ie, the cis or trans relation between the nucleotide changes in exon 6 and those in exon 7; figure 5A). However, the haplotype phase of the nucleotide changes could be imputed or inferred on the basis of known allele frequencies (figure 5B). To improve ABO typing, we added better integration of heterozygous and homozygous blood type O c.261delG changes to the improved algorithm on the basis of a decision tree

	Antigen	Serology	SNP array	WGS algorithm			Discordance	Algorithm modification
				Initial	Improved	Final		
1	ABO	B	..	AB	B	B	Initial algorithm did not integrate or phase heterozygous nucleotide changes in A, B, and O alleles	Population allele haplotype frequencies were added to estimate haplotype phase of O-type c.261delG as in cis to A nucleotide changes
1, 6	C	Positive	Positive	Negative	Positive	Positive	Misalignment of *RHCE* exon 2 sequences onto *RHD* exon 2	Copy number analysis of *RHCE* exon 2 sequence included to detect C antigen
7	C	Negative	Negative	Positive	Positive	Negative	C antigen misalignment not detected because its coverage was slightly less than the cutoff set after WGS of participant 6	Copy number analysis of *RHCE* exon 2 and surrounding introns to detect C antigen
9	D	Positive	..	No call	Positive	Positive	The copy number cutoff to detect the presence of *RHD* in the initial algorithm was too high to call this hemizygous participant	Modified copy number analysis (coverage cutoff over *RHD* lowered to 0·5×) to detect D antigen in participants who are hemizygous for *RHD*
20	ABO	O	..	B	O	O	Algorithm did not correctly integrate or phase homozygous c.261delG change in O allele with heterozygous nucleotide changes in A and B alleles	Modified to call O when homozygous for c.261delG, regardless of other nucleotide changes
34, 91	c	Negative*	Negative	..	Positive	Negative	One errant sequence causing incorrect C antigen c.307C genotype	Modified minimum nucleotide coverage cutoff for calling an allele
42, 76	ABO	O	..	..	A	O	Inability to integrate or phase two different O alleles (O*01.01 and O*02.01)	Modified to estimate the phase of heterozygous O alleles with different causative nucleotide changes as in trans
56	D	Positive	..	..	No call	Positive (DIIIa-CE(4-7)-D†)	Copy number analysis indicated that the participant was heterozygous for wildtype *RHD*, with *RHD* with exons 4–7 misalignment and c.186T, c.410T, and c.455C changes; algorithm could not integrate or phase results	Modified the *RHD* copy number analysis to still call D antigen if *RHD* misalignment was present with one copy of wildtype *RHD*
56	V	..	Negative	..	Positive	Negative	Inability to integrate or phase heterozygous V-antigen nucleotide changes c.733C/G and c.1006G/T	Population allele haplotype frequencies included to phase c.733G and c.1006T changes in cis to correctly call V negative (c.1006T destroys expression of V)
75	Fy^b	Positive	Negative	..	Negative	Negative	Repeat serological testing on a follow-up sample agreed with WGS algorithm	Initial serological typing error
79	E	Positive	Negative	..	Negative	Negative	SNP array testing agreed with WGS; serological typing of this participant was done at the same time as that of participant 80; results for E antigen were probably inverted between samples	Initial serological typing error
80	E	Negative	Positive	..	Positive	Positive	SNP array testing agreed with WGS; serological typing of this participant was done at the same time as that of participant 79; results for E antigen were probably inverted between samples	Initial serological typing error
85	Jk^a	Negative	Positive	..	Positive	Positive	Repeat serological testing on a follow-up sample agreed with WGS algorithm	Initial serological typing error
89	M	Positive	Positive	..	Negative	Positive	*GYPA* exon 2 sequences misaligned to *GYPE* exon 2; M antigen changes did not reach call level in some samples)	Calling cutoff lowered to 2× coverage
102	Fy^b	Negative	Positive	..	Positive	Positive	Repeat serological testing on a frozen aliquot agreed with WGS algorithm	Initial serological typing error
105	Fy^b	Negative	Negative	..	Positive	Negative	Inability to integrate or phase heterozygous c.–67c null with Fy^b nucleotide change	Population allele haplotype frequencies added to phase c.–67c in cis with Fy^b nucleotide change and in trans with Fy^a nucleotide change
109	S	Negative	Negative	..	Positive	Negative	Inability to integrate or phase heterozygous null GYPB*03N.04 allele nucleotide changes with heterozygous S and s nucleotide changes	Population allele haplotype frequencies added to phase GYPB*03N.04 in cis with S antigen nucleotide changes and in trans with s antigen nucleotide changes
109	Jo^a	..	Negative	..	Positive and negative	Negative	Inability to integrate or phase heterozygous Jo^a c.350C/T nucleotide change and heterozygous Hy c.323G/T nucleotide change	Modified algorithm based on population allele haplotype frequencies to phase Hy and Jo^a nucleotide changes

SNP=single nucleotide polymorphism. WGS=whole-genome sequencing. *Serological typing was not done for participant 34. †Participant is heterozygous (D positive, *DIIIa-CE(4-7)-D* positive; confirmed with allele-specific PCR).

Table: **MedSeq Project discordances and WGS algorithm fixes**

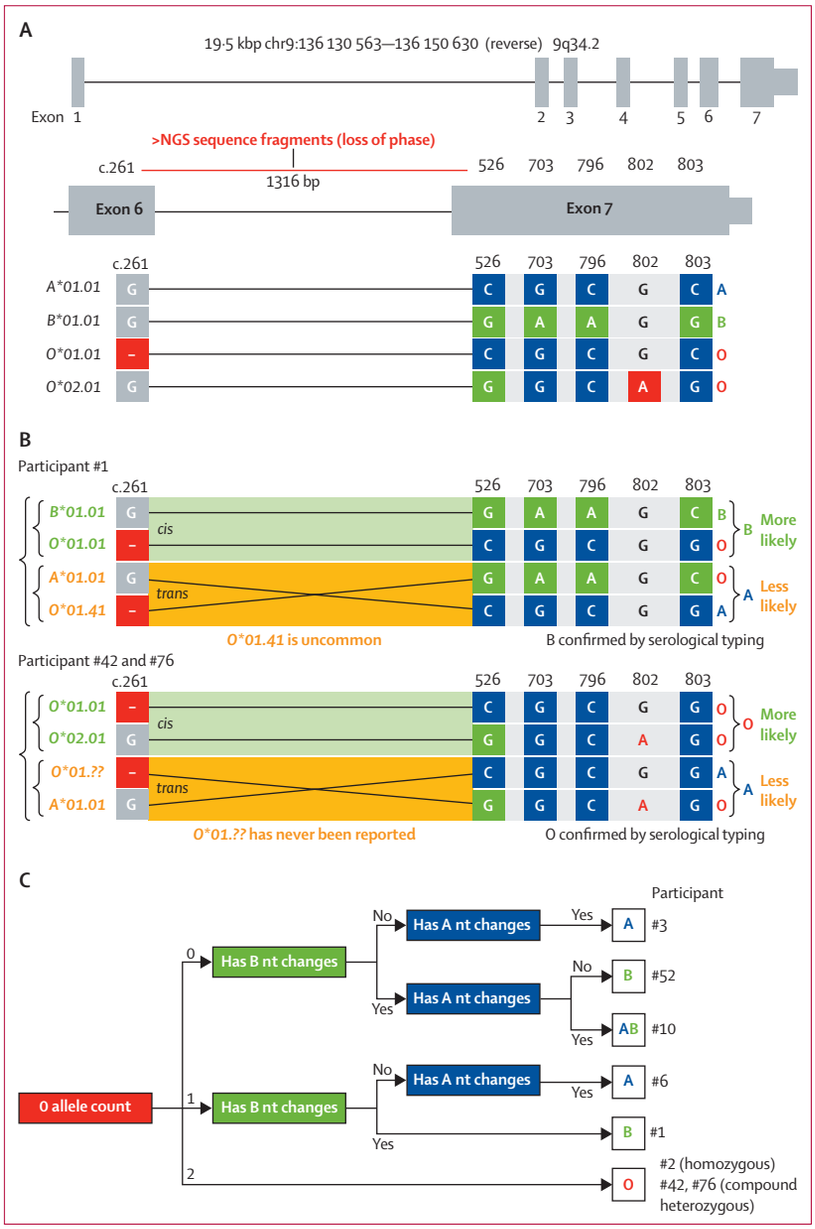

Figure 5: **ABO typing algorithm considerations**
(A) ABO exons 6 and 7 contain the nucleotide positions largely responsible for the activity and specificity of the transferase. (B) Allele haplotypes can be inferred from known population frequencies to impute the phase between exon 6 O*01.01 allele deletion (c.261) and exon 7 changes characteristic of A versus B transferase enzymes (c.526, c.703, c.796, c.803) and the O.02.01 allele nucleotide change (c.802). (C) Decision tree for imputing the ABO phenotype based on known haplotype frequencies. The decision tree first evaluates the number of distinct O nucleotide alleles (eg, c.261delG or c.802A), followed by an evaluation of c.526, c.703, c.796, and c.803 for the presence of B and then A allele nucleotide changes. Representative participants are listed for each decision output. chr=chromosome. NGS=next-generation sequencing. nt=nucleotide.

(figure 5C), in which c.261delG is phased cis to blood type A nucleotide changes in exon 7. This approach should be highly accurate across most ethnicities.[20]

We added copy number analysis to the initial algorithm to improve typing for the D and C antigens, as well as various allele combinations (appendix pp 29, 30). When testing the two different approaches to C antigen copy number analysis across the first 20 MedSeq participants, we found that when only exon 2 was considered the copy number was discordant with serology for one participant. By contrast, when considering exon 2 and parts of the surrounding introns, the copy number was concordant with serology for all 20 participants. Therefore, we incorporated assessment of exon 2 and surrounding introns into the improved algorithm. The performance of these two different approaches in all 110 participants is shown in the appendix (pp 29, 30).

After using data from the first 20 MedSeq genomes to improve the algorithm, we typed the remaining 90 MedSeq genomes for 38 RBC and 22 platelet antigens using the improved algorithm (figure 3), and compared the results with those of serological and SNP typing methods. The improved algorithm was 99·8% concordant (5390 correct calls out of 5400 individual antigen typings; appendix pp 27, 28), with ten discordant typings for RBC antigens and none for platelet antigens (figure 4, table). Discordances included six cis–trans haplotype ambiguities and four misalignments in homologous genes (figure 4). Five additional discordant results were due to incorrect serological RBC typing, which were confirmed by comparison with SNP testing and on repeat serological testing. The improved algorithm with inclusion of the *RHD* copy number analysis correctly predicted the presence or absence of the D antigen in all 110 MedSeq participants (appendix pp 29, 30), including *RHD* zygosity in 40 participants with conventional hybrid box PCR zygosity testing (appendix pp 29, 30).

To further improve the final algorithm, we added typing using the second most common blood type O nucleotide change c.802A, phased in cis to blood type A nucleotide changes in exon 7, but trans to blood type O c.261delG nucleotide changes if present in a compound heterozygous type O individual (figure 5B and C). Additionally, we programmed the final algorithm to detect any structural change or gene conversion between *RHD* and *RHCE*. For example, *RHD*DIIIa-CE(4-7)-D* is a hybrid *RHD* gene in which exons 4–7 of *RHD* are replaced by exons 4–7 of *RHCE* (appendix pp 29, 30). This hybrid gene, which is not uncommon in people of African ancestry,[20] encodes a D-negative phenotype, as well as a clinically important partial C phenotype. In one MedSeq participant, the *RHD*DIIIa-CE(4-7)-D* hybrid was trans to the wildtype *RHD*01*. In the whole-genome sequencing alignment, exons 4–7 (and the intervening introns) of *RHD* misaligned to *RHCE* in that individual (appendix pp 29, 30). Therefore, we programmed the final algorithm to analyse the *RHD–RHCE* misalignment with copy number analysis to detect the presence of both the *RHD*DIIIa-CE(4-7)-D* hybrid and the C antigen gene conversion.

We validated the final algorithm in 200 genomes from the INTERVAL study with an average coverage of 15× depth.[19] The final algorithm was 99·2% concordant

with serological methods in typing of 21 RBC antigens encoded by 14 genes (3486 correct calls out of 3515 individual antigen typings; appendix pp 31–42, 45). Analysis of the discordances showed that most were attributable to technical limitations caused by the lower average depth of coverage for INTERVAL genomes than for MedSeq genomes (15× *vs* 30×)—for example, the correct antigen nucleotides were detected but were present below the 4× nucleotide cutoff (appendix pp 43, 44). In particular, the low sequencing coverage of the INTERVAL genomes caused difficulties when typing for the M antigen, with some genomes having M antigen alignments as low as 1× and even 0×. We addressed this problem after our initial technical troubleshooting round using the first 20 INTERVAL samples by setting the cutoff that defined M antigen positivity to 1× and including loss of *GYPA* exon 2 (M antigen nucleotide location) as a backup method for confirming the presence or absence of M antigen.

When adjusted for 15× depth coverage, the typing algorithm based on whole-genome sequencing was 99·9% concordant with serological typing (3486 correct calls out of 3490 individual antigen typings; appendix p 46). The final algorithm was 100% concordant with serology in typing for ABO and D antigen, and 99·5% accurate at typing the C antigen. The one C antigen discordance was probably due to the copy number analysis misinterpreting a 1× loss of exon 2 coverage over *RHCE* as C-antigen positive, when this particular loss of coverage was in the context of a larger 1× loss over exons 2–6, likely indicating the presence of a heterozygous *RHCE-D(2-6)-CE* gene conversion, known to be C-negative (INTERVAL genome EGAN00001288526). The performance statistics of the initial, improved, and final bloodTyper algorithms can be found in the appendix (p 47).

Discussion

In this study, we developed and iteratively improved an algorithm based on whole-genome sequencing for the typing of RBC and platelet antigens. During the development stage, we compared our initial algorithm with conventional serological and SNP typing methods in the first 20 participants from the whole-genome sequencing group of the MedSeq randomised controlled trial. An improved version of the algorithm was then compared with conventional serological and SNP typing methods for 38 RBC and 22 platelet antigens in the remaining 90 participants of the MedSeq Project, with 99·8% concordance. We then created a final version of the algorithm and further validated it in 200 genomes from the INTERVAL study. We found that the final algorithm was 99·9% concordant with serological methods in typing of 21 RBC antigens. The final bloodTyper algorithm is available online.

Blood transfusion is commonly used in clinical medicine, with 112·5 million units of blood collected worldwide each year.[31] Pre-transfusion testing involves matching the patient and donor for ABO and RhD blood types on the basis of principles that have not changed much in more than 60 years. Matching of donors and recipients for other common RBC antigens (eg, C, E, and K) is practised in several high-income countries, but not routinely in the USA, with the exception of some centres that treat patients with sickle-cell disease or thalassaemia. Exposure to non-self RBC antigens via transfusion leads to production of antigen-specific antibodies in roughly 3% of white recipients and 30–50% of individuals of African ancestry who receive long-term blood transfusion therapy.[1] Once patients are sensitised to non-self antigens, they are at increased risk of development of additional antibodies to RBCs,[32] and all future donor units must be tested and found to be negative for those antigens to avoid transfusion reactions. After sensitisation, the risk of haemolytic reactions increases over time as the reactivity of antibodies decreases to below the level of detection.[33] Between five and 16 deaths from haemolytic transfusion reactions associated with antibodies to non-ABO blood group antigens are reported to the US Food and Drug Administration each year, almost all due to the inability to detect pre-existing antibodies or the need for emergency transfusion in patients who have previously been sensitised.

The risk of alloantibody complications after blood transfusion is 3–30% (ie, 3·4 million complications based on 112·5 million units of blood per year worldwide[31]). This risk has been accepted by the medical community in the absence of efficient strategies for reducing the risk of alloantibody complications after transfusion. Antibody-based serological typing methods are labour intensive and are not easily scaled-up, and serological reagents are not available to type for all clinically important antigens. Existing DNA-based SNP typing methods are limited by the number of polymorphisms that can be targeted, because they do not interrogate structural changes (such as gene conversion events), and because they are not comprehensive enough to definitively ascertain ABO and RhD antigens.

Our automated analytical software algorithm could be transformative in the implementation of population-level RBC and platelet antigen typing. However, further characterisation of antigenic changes not validated in this study is required. The ability to test large populations of donors and recipients for clinically important antigens that do not have serological reagents could greatly reduce transfusion-related morbidity and mortality. As whole-genome sequencing becomes more common in clinical practice, secondary analysis of existing data could allow inexpensive, comprehensive blood-group typing to become part of donor and patient medical records.

Several studies[5,6,11,12,34] have investigated RBC antigen prediction based on next-generation sequencing. However, these analyses were often restricted to a few targeted SNPs, or required interpretation by an expert (in some instances,

For the **online algorithm** see https://bloodantigens.com/bloodTyper

one who already knew the results obtained through conventional antigen-typing), so these methods might not offer a scalable solution for widespread clinical implementation.

Our study is not the first to create an algorithm for RBC and platelet antigen typing based on whole-genome sequencing. Giollo and colleagues[35] designed an algorithm to predict ABO and D RBC antigens using Hidden Markov Models, RBC antigen allele data from the now-retired Blood Group Antigen Gene Mutation Database,[21] and individuals sequenced in the Personal Genome Project.[36,37] When compared with serological results, concordance was 94% for ABO (n=71) and 94% for D-antigen (n=69), but their Hidden Markov Model approach makes additional improvements difficult because the typing method is abstracted in the predictive model without clear means to address specific discordances. By contrast, our algorithmic rules-based approach allows for iterative improvements on the basis of molecular analysis of discordances. Our improved algorithm, tested in a masked setting, had a concordance of 98% for ABO (n=90) and 99% for D-antigen (n=90) with serology. We were then able to update the algorithm further, such that on subsequent masked testing, using another dataset, the final algorithm had a concordance of 100% with serology for ABO (n=200) and D-antigen (n=200). Our efforts also included masked concordance testing of our whole-genome sequencing typing algorithm for an additional 35 RBC antigens and 22 platelet antigens.

The MedSeq Project cohort represents a cross-section of the population treated at a major academic medical centre that serves a large urban area in North America. One individual was negative for Lu^b and another for Jo^a, which is rarely observed in clinical practice. We typed for and identified several RBC antigen changes that are commonly found in individuals of African ancestry, including positivity for V, VS, and Js^a antigens; negativity for Fy^a and Fy^b antigens; and the presence of *RHD*DIIIa-CE(4-7)-D*. Knowledge of antigenic changes could allow for better matching of recipients and donors.

Antigen profiles are useful in that they indicate which antigens are absent in a patient, thus offering the patient's physician insight into risk of alloantibody sensitisation to aid pre-transfusion antibody identification and prenatal antibody screening. This information could be integrated into clinical decision support and used if and when a patient needs a blood transfusion. Knowledge of antigen profiles could also be used to recruit individuals with uncommon or rare antigen combinations as blood donors. Antigen profiling could be particularly important for individuals who do not have common platelet antigens to prevent antibody sensitisation associated with fetal and neonatal alloimmune thrombocytopenia, post-transfusion purpura, or idiopathic platelet transfusion refractoriness.

Our study has some limitations. We could not test every known RBC or platelet antigen because some antigens are very rare, or only common in specific ethnic groups. Similarly, it was not possible to test all known hybrid and structural changes in Rh and MNS genes, so the copy number analyses of the algorithm will probably require optimisation in the future. We did not test ABO subtypes and hybrid ABO genes, which will necessitate updates to the phase-estimation decision tree and some long-range experimental phasing. Full validation of bloodTyper for all known antigenic changes will require the testing of additional samples with these untested phenotypes.

In summary, we have built a comprehensive database of antigen allele genotypes and an automated algorithm for typing of RBC and platelet antigens based on whole-genome sequencing. Further investigation is needed, but this algorithm might facilitate routine genetic prediction of all key blood-group antigens with a similar level of fidelity to that of current serological or SNP array approaches, potentially transforming the way in which safe blood products are provided to patients.

Contributors

WJL, MA, RS-W, RMK, HLR, LES, and RCG designed the study. MA, RS-W, SV, and HHM contributed to collection of MedSeq Project data. WJL, CMW, MA, RS-W, SV, and DPS analysed MedSeq Project data. NSG, KW, NS, EDA, JD, DJR, NAW, WHO, and ASB contributed to the INTERVAL study design, data collection, and data analysis. WJL did the blood type analysis, wrote the first draft, and designed the tables and figures. All authors contributed to reviewing and editing the final version.

Declaration of interests

WJL reports non-financial support from Illumina, outside the submitted work. MSL reports grants from the National Human Genome Research Institute (NHGRI) of the National Institutes of Health, during the conduct of the study. JD reports grants from UK Medical Research Council (MRC), British Heart Foundation, UK National Institute of Health Research (NIHR), and European Commission, during the conduct of the study. He also reports personal fees and non-financial support from Merck Sharp & Dohme (MSD) and Novartis, and grants from British Heart Foundation, European Research Council, MSD, NIHR, NHS Blood and Transplant, Novartis, Pfizer, UK MRC, Wellcome Trust, and AstraZeneca, outside the submitted work. ASB reports grants from MSD, Biogen, Pfizer, AstraZeneca, and Novartis, and personal fees from Novartis, outside the submitted work. HLR reports personal fees from Brigham and Women's Hospital, outside the submitted work. RCG reports personal fees from AIA, Americord, Veritas, and Helix, outside the submitted work. All other authors declare no competing interests.

Acknowledgments

The MedSeq Project was supported by the NHGRI (U01-HG006500). WJL is supported by the Brigham and Women's Hospital Pathology Department Stanley L Robbins MD Memorial Research Fund Award. DPS is supported by the National Heart, Lung, and Blood Institute (T32-HL007627), and CMW by the Doris Duke Charitable Foundation (2011097 and 2015133). RCG is supported by the National Institutes of Health (U19-HD077671, U01-HG008685, R03-HG008809, UG3-OD023156, U01-AG24904, R01-CA154517, P60-AR047782, R01-AG047866), as well as funding from the Broad Institute and Department of Defense. The authors thank the staff and participants of the MedSeq Project. Recruitment into the INTERVAL study was supported by NHS Blood and Transplant, National Institute for Health Research, British Heart Foundation, and the UK Medical Research Council. Sequencing in INTERVAL was supported by the Wellcome Sanger Institute; data analysis was partly supported by the Cambridge Substantive Site of Health Data Research UK.

References

1 Hendrickson JE, Tormey CA, Shaz BH. Red blood cell alloimmunization mitigation strategies. *Transfus Med Rev* 2014; **28:** 137–44.

2 AABB. Technical manual, 19th edn. Bethesda, MD: American Association of Blood Banks, 2017.

3 Paccapelo C, Truglio F, Antonietta Villa M, Revelli N, Marconi M. HEA BeadChip technology in immunohematology. *Immunohematology* 2015; **31:** 81–90.

4 Hashmi G, Shariff T, Zhang Y, et al. Determination of 24 minor red blood cell antigens for more than 2000 blood donors by high-throughput DNA analysis. *Transfusion* 2007; **47:** 736–47.

5 Stabentheiner S, Danzer M, Niklas N, et al. Overcoming methodical limits of standard *RHD* genotyping by next-generation sequencing. *Vox Sanguinis* 2011; **100:** 381–88.

6 Fichou Y, Audrezet MP, Gueguen P, Le Marechal C, Ferec C. Next-generation sequencing is a credible strategy for blood group genotyping. *Br J Haematol* 2014; **167:** 554–62.

7 Lane WJ, Westhoff CM, Uy JM, et al. Comprehensive red blood cell and platelet antigen prediction from whole genome sequencing: proof of principle. *Transfusion* 2016; **56:** 743–54.

8 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:** 240–49.

9 Baronas J, Westhoff CM, Vege S, et al. *RHD* zygosity determination from whole genome sequencing data. *J Blood Disord Transfus* 2016; **7:** 365.

10 Lang K, Wagner I, Schone B, et al. ABO allele-level frequency estimation based on population-scale genotyping by next generation sequencing. *BMC Genomics* 2016; **17:** 374.

11 Chou ST, Flanagan JM, Vege S, et al. Whole-exome sequencing for RH genotyping and alloimmunization risk in children with sickle cell anemia. *Blood Adv* 2017; **1:** 1414–22.

12 Orzinska A, Guz K, Mikula M, et al. A preliminary evaluation of next-generation sequencing as a screening tool for targeted genotyping of erythrocyte and platelet antigens in blood donors. *Blood Transfus* 2018; **16:** 285–92.

13 Vassy JL, Lautenbach DM, McLaughlin HM, et al. The MedSeq Project: a randomized trial of integrating whole genome sequencing into clinical medicine. *Trials* 2014; **15:** 85.

14 Vassy JL, McLaughlin HM, MacRae CA, et al. A one-page summary report of genome sequencing for the healthy adult. *Public Health Genomics* 2015; **18:** 123–29.

15 McLaughlin HM, Ceyhan-Birsoy O, Christensen KD, et al. A systematic approach to the reporting of medically relevant findings from whole genome sequencing. *BMC Med Genet* 2014; **15:** 134.

16 Vassy JL, Christensen KD, Schonman EF, et al. The impact of whole-genome sequencing on the primary care and outcomes of healthy adult patients: a pilot randomized trial. *Ann Intern Med* 2017; **167:** 159–69.

17 Roberts JS, Robinson JO, Diamond PM, et al. Patient understanding of, satisfaction with, and perceived utility of whole-genome sequencing: findings from the MedSeq Project. *Genet Med* 2018; published online Jan 4. DOI:10.1038/gim.2017.223.

18 Christensen KD, Vassy JL, Phillips KA, et al. Short-term costs of integrating whole-genome sequencing into primary care and cardiology settings: a pilot randomized trial. *Genet Med* 2018; published online March 22. DOI:10.1038/gim.2018.35.

19 Di Angelantonio E, Thompson SG, Kaptoge S, et al. Efficiency and safety of varying the frequency of whole blood donation (INTERVAL): a randomised trial of 45 000 donors. *Lancet* 2017; **390:** 2360–71.

20 Reid ME, Lomas-Francis C, Olsson ML. The blood group antigen FactsBook. 3rd edn. Waltham, MA: Academic Press, 2013.

21 Patnaik SK, Helmberg W, Blumenfeld OO. BGMUT: NCBI dbRBC database of allelic variations of genes encoding antigens of blood group systems. *Nucleic Acids Res* 2012; **40:** D1023–29.

22 International Society of Blood Transfusion. Red cell immunogenetics and blood group terminology. http://www.isbtweb.org/working-parties/red-cell-immunogenetics-and-blood-group-terminology (accessed Nov 1, 2017).

23 Robinson J, Halliwell JA, McWilliam H, Lopez R, Marsh SG. IPD—the Immuno Polymorphism Database. *Nucleic Acids Res* 2013; **41:** D1234–40.

24 Metcalfe P, Watkins NA, Ouwehand WH, et al. Nomenclature of human platelet antigens. *Vox Sanguinis* 2003; **85:** 240–45.

25 Wagner FF, Flegel WA. The Human RhesusBase. http://www.rhesusbase.info (accessed Dec 20, 2016).

26 Chiu RW, Murphy MF, Fidler C, Zee BC, Wainscoat JS, Lo YM. Determination of RhD zygosity: comparison of a double amplification refractory mutation system approach and a multiplex real-time quantitative PCR approach. *Clin Chem* 2001; **47:** 667–72.

27 Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008; **456:** 53–59.

28 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:** 1297–303.

29 Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010; **26:** 841–42.

30 Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013; **14:** 178–92.

31 WHO. Blood safety and availability. 2017. http://www.who.int/mediacentre/factsheets/fs279/en (accessed March 21, 2018).

32 Schonewille H, van de Watering LM, Brand A. Additional red blood cell alloantibodies after blood transfusions in a nonhematologic alloimmunized patient cohort: is it time to take precautionary measures? *Transfusion* 2006; **46:** 630–35.

33 Tormey CA, Stack G. The persistence and evanescence of blood group alloantibodies in men. *Transfusion* 2009; **49:** 505–12.

34 Rieneck K, Bak M, Jonson L, et al. Next-generation sequencing: proof of concept for antenatal prediction of the fetal Kell blood group phenotype from cell-free fetal DNA in maternal plasma. *Transfusion* 2013; **53** (suppl 2)**:** 2892–98.

35 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:** e0124579.

36 Ball MP, Thakuria JV, Zaranek AW, et al. A public resource facilitating clinical use of genomes. *Proc Natl Acad Sci USA* 2012; **109:** 11920–27.

37 Church GM. The Personal Genome Project. *Mol Syst Biol* 2005; **1:** 2005.