

Comprehensive red blood cell and platelet antigen prediction from whole genome sequencing: proof of principle

William J. Lane,^{1,2} Connie M. Westhoff,³ Jon Michael Uy,¹ Maria Aguad,¹
Robin Smeland-Wagman,¹ Richard M. Kaufman,¹ Heidi L. Rehm,^{1,2,4,5} Robert C. Green,^{2,5,6} and
Leslie E. Silberstein⁷ for the MedSeq Project*

BACKGROUND: There are 346 serologically defined red blood cell (RBC) antigens and 33 serologically defined platelet (PLT) antigens, most of which have known genetic changes in 45 RBC or six PLT genes that correlate with antigen expression. Polymorphic sites associated with antigen expression in the primary literature and reference databases are annotated according to nucleotide positions in cDNA. This makes antigen prediction from next-generation sequencing data challenging, since it uses genomic coordinates.

STUDY DESIGN AND METHODS: The conventional cDNA reference sequences for all known RBC and PLT genes that correlate with antigen expression were aligned to the human reference genome. The alignments allowed conversion of conventional cDNA nucleotide positions to the corresponding genomic coordinates. RBC and PLT antigen prediction was then performed using the human reference genome and whole genome sequencing (WGS) data with serologic confirmation.

RESULTS: Some major differences and alignment issues were found when attempting to convert the conventional cDNA to human reference genome sequences for the following genes: *ABO*, *A4GALT*, *RHD*, *RHCE*, *FUT3*, *ACKR1* (previously *DARC*), *ACHE*, *FUT2*, *CR1*, *GCNT2*, and *RHAG*. However, it was possible to create usable alignments, which facilitated the prediction of all RBC and PLT antigens with a known molecular basis from WGS data. Traditional serologic typing for 18 RBC antigens were in agreement with the WGS-based antigen predictions, providing proof of principle for this approach.

CONCLUSION: Detailed mapping of conventional cDNA annotated RBC and PLT alleles can enable accurate prediction of RBC and PLT antigens from whole genomic sequencing data.

Prediction of red blood cell (RBC) and platelet (PLT) antigens using DNA assays has the potential to augment or replace traditional serologic antigen typing in many situations. DNA-based typing methods are more easily automated, amenable to multiplexing, and do not require expensive and sometimes difficult to obtain serologic immunoglobulin

ABBREVIATIONS: CDS = coding DNA sequence; NGS = next-generation sequencing; SNP(s) = single-nucleotide polymorphism(s); WGS = whole genome sequencing.

From the ¹Department of Pathology, the ⁶Division of Genetics, Department of Medicine, and the ⁷Division of Transfusion Medicine, Department of Pathology, Brigham and Women's Hospital; and ²Harvard Medical School, Boston, Massachusetts; ³New York Blood Center, New York, New York; and the ⁴Laboratory for Molecular Medicine and the ⁵Partners Healthcare Personalized Medicine, Boston, Massachusetts.

Address reprint requests to: William J. Lane, MD, PhD, Pathology Department, Brigham and Women's Hospital and Harvard Medical School, Amory Lab Building 3rd Floor, Room 3-117, 75 Francis Street, Boston, MA 02115; e-mail: wlane@partners.org.

The MedSeq Project is carried out as a collaborative study supported by the National Human Genome Research Institute HG006500. Additional funding was provided by HD077671.

*Additional members of the MedSeq Project are listed in the acknowledgments.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

Received for publication July 27, 2015; revision received September 15, 2015; and accepted October 14, 2015.

doi:10.1111/trf.13416

© 2015 The Authors Transfusion published by Wiley Periodicals, Inc. on behalf of AABB

TRANSFUSION 2015;00:00-00

reagents. As such, DNA-based approaches could allow for more extensive characterization of patient and donor phenotypes and enable enhanced blood product selection and identification of donors with rare phenotypes.

There are 346 serologically distinct RBC blood group antigen phenotypes recognized by the International Society of Blood Transfusion (ISBT).¹ For most RBC antigens there is a known correlation between the antigen phenotype and one or more molecular changes defined by more than 1100 alleles across 45 genes.²⁻⁹ There are 33 serologically distinct human PLT antigen (HPA) phenotypes recognized by the Platelet Nomenclature Committee.¹⁰ For all 33 PLT antigens, the molecular basis is known and can be characterized by 33 alleles within six genes.¹⁰⁻¹² Resources that catalog RBC antigen allele variants include the ISBT website,² the Blood Group Antigen FactsBook,³ the BGMUT website,¹³ and the *RHD* RhesusBase.¹⁴ Alleles encoding PLT antigens are available from the Immuno Polymorphism Database-HPA website.¹⁰⁻¹² These resources provide a means to validate and design single-nucleotide polymorphism (SNP) assays to predict phenotypes. However, current SNP-based molecular typing assays have limitations^{15,16} including: 1) need for specialized testing instruments, reagents, and workflows; 2) do not include all of the known blood group genes; 3) target selective gene regions without evaluating all potentially contributory genetic changes; and 4) more complex antigens require the integration of multiple assays.¹⁶

The RH (e.g., D, C/c, E/e) and MNS (e.g., M/N, S/s) blood group system antigens are challenging to predict given the large number of complex alleles, genetic variation, and gene rearrangements between *RHD/RHCE* and *GYPB/GYPB/GYPE*. Most of the other RBC protein antigens (e.g., K/k, Fy^{a/b}) are the result of single well-characterized inherited missense variants.^{3,4} However, additional molecular changes can cause antigen expression to be weak or silenced (null) due to alternative splicing, premature stop codons, hybrid genes, promoter silencing, and at the protein level, altered membrane insertion or changes to interacting proteins or modifying genes. High-resolution predictive accuracy would require large regions of sequence coverage to identify all potentially relevant changes. Although commercial SNP assays evaluate common polymorphisms to predict protein-based antigens,^{15,17} they do not include all clinically significant changes.

The RBC carbohydrate antigens (e.g., ABO, Le^{a/b}, P1, P^k) are synthesized by enzymes.³ DNA-based determination of carbohydrate antigen expression is not widespread because accurate prediction requires gene sequencing to properly predict the enzymatic and sugar specificity across several genes (e.g., ABO antigen prediction requires evaluation of *ABO* along with *FUT1*^{3,19} and *FUT2*^{21,22}). In addition, alleles associated with carbohydrate antigens are complex, often contain multiple nucleotide changes, and are numerous (e.g., >300 *ABO* alleles reported¹³), with

many different null alleles. The clinical significance of missing one inactivating mutation for ABO is an unacceptable risk for transfusion and, therefore, the limited sequence coverage of SNP targeted typing is currently inadequate.

PLT antigens are mainly associated with single missense variants.²³ Although molecular assays exist to predict PLT antigens,²⁴ matching for patients, with the possible exception of HPA-1a, is underutilized in clinical practice at the present time given the cost and lack of antigen typed donors.

Next-generation sequencing (NGS) would overcome many of the limitations associated with SNP-based assays. NGS-based molecular prediction has been successfully applied to human leukocyte antigens²⁵⁻³⁰ and human neutrophil antigens.³¹ However, there are no published reports of NGS-based PLT antigen prediction and only three reports of targeted NGS-based RBC antigen prediction: 1) *RHD* in 26 samples with weak D antigens,³² 2) K/k allelic polymorphism (c.578) using cell-free fetal DNA in three pregnant females,³³ and 3) 18 genes that control 15 blood group systems in four individuals.³⁴ Recently, an algorithm was published³⁵ that used the BGMUT database¹³ to predict RBC antigens for ABO and D typed individuals from the personal genome.^{36,37}

With the emergence of genomic approaches and personalized medicine, NGS-based whole genome sequencing (WGS) data could be used to evaluate genes encoding RBC and PLT antigens to predict the presence of antigens with a known molecular basis. There are no reports describing comprehensive WGS-based RBC or PLT antigen prediction. One of the challenges for this approach is that the allele reference sources list the nucleotide changes according to coding DNA sequence (CDS) positions based on cDNA sequences. It is not readily possible to predict RBC and PLT antigens from WGS data, since the data use genomic coordinates linked to the human reference genome. In this article we describe an approach for the prediction of RBC and PLT antigens from WGS data and demonstrate the feasibility of the approach.

MATERIALS AND METHODS

Conversion of conventional cDNA positions to genomic coordinates

Conventional cDNA reference sequence CDS positions were converted to genomic coordinates: 1) reference cDNA and protein sequences were downloaded from GenBank; 2) human reference genome UCSC genomic transcripts, corresponding to the splicing pattern of the conventional cDNA sequence, were downloaded in a format identifying the exons and introns and the genomic start and end positions (exons, uppercase; introns, lowercase); 3) the cDNA reference sequence and the human reference genome sequences were aligned using Clustal

Omega v1.1.1;³⁸ 4) the start and termination codon genomic positions were manually determined in the Integrated Genomic Viewer Version 2.3.26;³⁹ and 5) the CDS start position and alignments were then used as a reference to convert between cDNA, gene, and genomic coordinate positions.

Predicting antigens from the human reference genome

RBC and PLT antigens encoded by each cDNA reference sequence are well established.^{2,3,10} The conventional cDNA reference and human reference genome alignments were used to determine the CDS and amino acid positions that differed. The known alleles^{2-4,10,12,23} were used to manually determine if any difference altered the presence or absence of a RBC and PLT antigen, which allowed for the prediction of the RBC and PLT antigens encoded by the human reference genome.

WGS-based antigen prediction

With approval from the Partners HealthCare Human Research Committee, a sample for RBC phenotyping and genomic DNA isolation was collected from a patient participating in the MedSeq Project.⁴⁰ Whole genomic sequencing was performed by the CLIA-certified, CAP-accredited Illumina Clinical Services Laboratory (San Diego, CA) using paired-end 100-bp reads on the Illumina HiSeq platform and sequenced to at least 30× mean coverage.⁴¹ The genomic data from the MedSeq project has been submitted to the dbGaP website. The genome used in this article is from dbGaP subject ID 1270611. Sequence read data were aligned to the human reference sequence (GRCh37/hg19) using Burrows-Wheeler Aligner 0.6.1-r104.⁴² Variant calls for 45 RBC and six HPA genes (300 bases upstream of start codon, exons, and 10 bases into each intron) were made using the Genomic Analysis Tool Kit Version 2.3-9-gdcgccbb and saved as a variant calling format (.vcf) file showing differences between the WGS data and the reference genome.⁴³ Sequencing coverage was extracted from the alignment file using BEDTools v2.17.0.⁴⁴ The Integrative Genomics Viewer³⁹ was used to verify coverage and sequence identity for positions in the .vcf file.

The genomic coordinates from the .vcf file were converted into CDS positions relative to the conventional cDNA sequences. Each variant was then compared to published allele tables.^{2-4,10,12,23} For alleles with nucleotide changes in the 3', 5', and intronic regions, the genomic coordinates were manually determined and evaluated in the NGS alignment file for the presence or absence of the antigen from published allele tables.^{2-4,10,12,23} To predict antigens with positions not in the .vcf file, the sequence coverage was analyzed for adequacy and if adequate the human reference genome prediction was used.

RBC serology

RBC serologic antigen typing by tube method was performed according to standard blood banking practices in the Brigham and Women's Hospital Blood Bank. Commercially available serologic typing reagents were used to type for the ABO, D, c, C, e, E, K, k, Fy^a, Fy^b, Jk^a, Jk^b, M, N, S, and s antigens.

RESULTS

Conversion of cDNA positions to genomic coordinates

RBC and PLT antigen polymorphisms have historically been defined using CDS positions referenced to published cDNA sequences with the A of the start codon (ATG) as Position 1. To predict antigens from NGS data, a manual workflow was created to map the CDS nucleotide changes to the respective genomic coordinates using alignments between the cDNA reference sequence and the GRCh37/hg19 human reference genome sequence for the cDNA, CDS, and protein sequences. Figure 1 and Table 1 illustrate the process for the Duffy system. The Fy^a/Fy^b antigens, c.125G/A, map to chr1:159,175,354G/A, and genomic coordinates were also determined for other reported FY alleles.

Differences between conventional cDNA and human reference genome

From the conversion and alignment process, differences were observed between conventional cDNA and the human reference genome. Minor differences included: 1) silent variants that do not encode amino acid changes, 2) different antigen allele, and 3) potentially nonrelevant missense changes. Table 2 summarizes the blood group, gene, nucleotide, and amino acid differences; location of change; or predicted impact on antigen expression. Major differences that would challenge interpretation were encountered in the following genes: *ABO*, *A4GALT*, *RHD*, *RHCE*, *FUT3*, *ACKRI* (previously *DARC*), *ACHE*, *FUT2*, *CRI*, *GCNT2*, and *RHAG*, summarized in Table 2.

ABO

The *ABO* gene determines the transferase enzymes responsible for the carbohydrate antigens, A and B. Any mutation that results in absence of transferase activity results in Group O. The conventional cDNA reference is an A allele. By analyzing the *ABO* gene region in the human reference genome it was found to represent sequence regions from two separate human reference genome sequencing contigs representing different haplotype alleles: 1) The AL158826.23 contig, which contains Exons 1 to 5 corresponds to the *ABO.O.01.02* allele, and 2) the AL732364.9 contig, which contains Exons 6 to 7 matches the *ABO*O.01.01* allele.³ Therefore, the reference sequence contains a deletion characteristic of *O*^I alleles

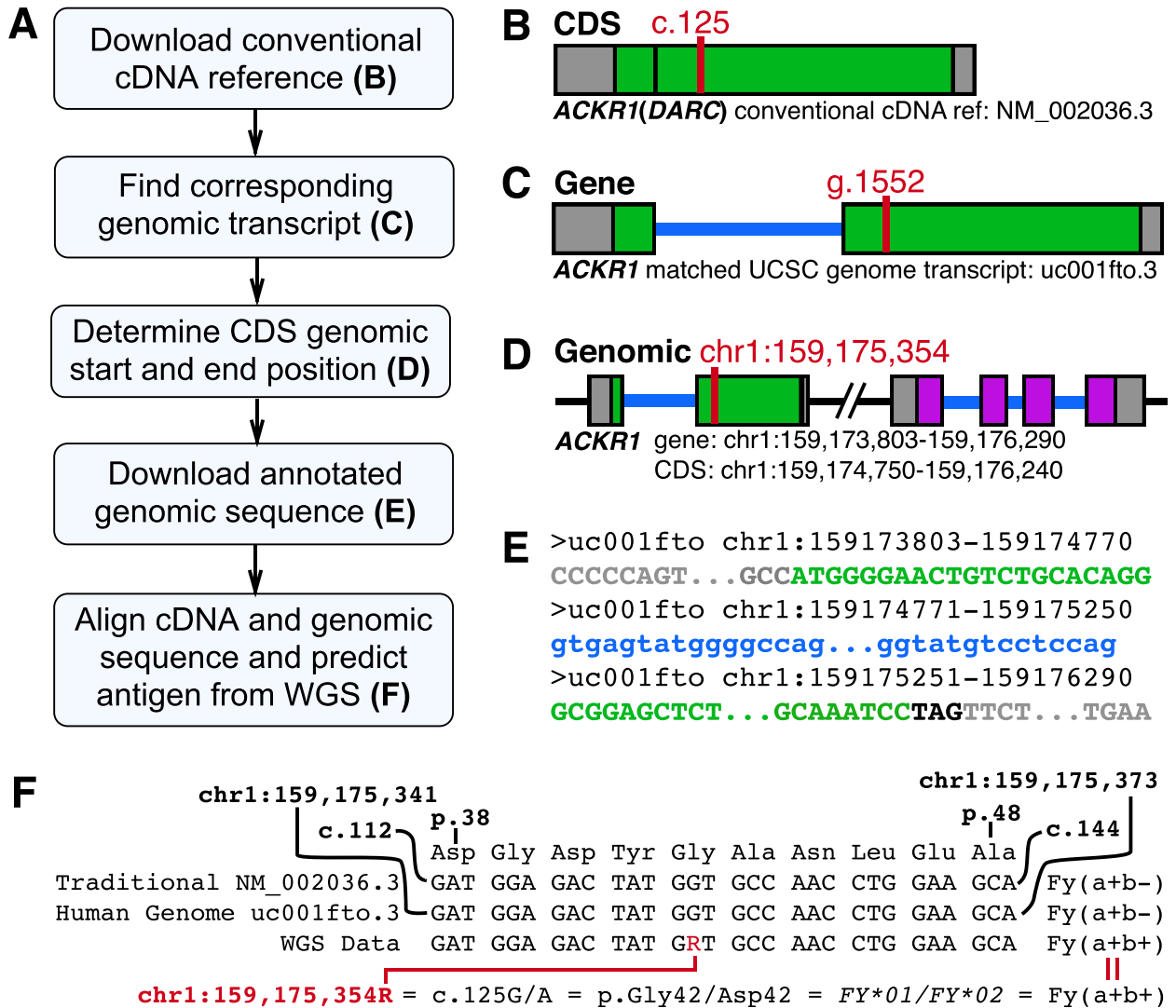


Fig. 1. Approach for mapping conventional cDNA reference sequence positions to genomic coordinates. (A) Process developed to convert conventional CDS positions to genomic coordinates with FY as an example. (B) CDS positions referenced to cDNA sequence. (C and D) Genome transcript and genomic coordinates according to the human reference genome. (E) UCSC genomic sequence in which each exon and intron is annotated as separate sequence entry preceded by the genomic coordinates. The sequence regions are colored: 3' and 5' (gray), CDS (green uppercase), and intron (blue lowercase). (F) FY gene conversion between genomic coordinates and cDNA reference sequence.

(c.261delG) and when analyzing NGS data *A* and *B* allele sequences would appear to have an insertion (chr1:136132908_136132909insG) and *O*^l alleles would not have the characteristic deletion.

A4GALT (P1PK)

The *A4GALT* gene encodes a lactosylceramide 4- α -galactosyltransferase enzyme responsible for the carbohydrate P1PK system antigens: P1 and P^k. All of the known P1PK null alleles are referenced using the splice form I reference sequence (GU902278-GU902281), but all of the nucleotides associated with P1+ and P1- expression are located

in a skipped Exon 2a, which is only found in the alternative spliced form IV of *A4GALT* (AJ245581).⁴⁵ Therefore, mapping required use of both conventional reference sequences to obtain the corresponding genomic coordinates.

RH

Relative to the conventional reference sequence for *RHD* (L08429) the human reference genome is c.1136C>T (chr1:25,643,553C>T) p.Thr379Met, which corresponds to the family of DAU alleles,⁴⁶ specifically *RHD***DAU0*, which is primarily found in African Americans.⁴⁷ The conventional reference sequence (DQ322275) for *RHCE***ce*

TABLE 1. FY alleles, cDNA and genomic coordinates, bases(s) found, WGS coverage, and result

Allele	CDS	Gene	Genome	Base(s) found	Coverage	Result
<i>FY*01N.01</i>	c.-67	g.881	chr1:159,174,683	T	28×	Absent
<i>FY*01/02</i>	c.125	g.1552	chr1:159,175,354	G/A	15/15×	Fy(a+b+)
<i>FY*02M.02</i>	c.145	g.1572	chr1:159,175,374	G	24×	Absent
<i>FY*02M.01/02</i>	c.265	g.1692	chr1:159,175,494	C	27×	Absent
<i>FY*01N.02</i>	c.281_295	g.1708_1722	chr1:159,175, 510_chr1:159,175,524	CTGGCT GGCCTGTCC	31-36×	Absent
<i>FY*01N.04</i>	c.287	g.1714	chr1:159,175,516	G	34×	Absent
<i>FY*02M.01/02</i>	c.298	g.1725	chr1:159,175,527	G	32×	Absent
<i>FY*01N.05</i>	c.327	g.1754	chr1:159,175,556	C	37×	Absent
<i>FY*02N.02</i>	c.407	g.1834	chr1:159,175,636	G	27×	Absent
<i>FY*01N.03</i>	c.408	g.1835	chr1:159,175,637	G	28×	Absent

TABLE 2. Differences encountered when aligning the conventional cDNA with the human reference genome (shown as conventional reference > reference genome)

Symbol	Gene	CDS nucleotide (genomic coordinate) [amino acid]	Differences
ABO	<i>ABO</i>	c.261delG (no genomic coordinate)	Genome: inactive enzyme Exons 1-5 correspond to <i>ABO.O.01.02</i> Exons 6 and 7 correspond to <i>ABO*O.01.01</i>
MNS	<i>GYPA</i>	c.38(chr4:145,041,741)C>A [p.Ala13Glu]; c.59(chr4:145,041,720)C>T [p.Ser20Lue], c.71(chr4:145,041,708)G>A, c.72(chr4:145,041,707)T>G [p.Gly24Glu]; c.93(chr4:145,041,686)C>T [p.Thr31Thr]	Probable nonrelevant missense change in cleaved N-term; Antigenic difference: M+N->M-N+; Silent change
MNS	<i>GYPB</i>	c.251(chr4:144,918,712)C>G [p.Thr84Ser]	Presumed non-relevant missense change c.251G is part of the S-s-U+w [GYPB.NY] allele which has additional changes c.208G>T and c.230C>T not present in the reference genome
P1PK	<i>A4GALT</i>		Two different cDNA reference sequences
RH	<i>RHD</i>	c.1136(chr1:25,643,553)C>T [p.Thr379Met]	Genome: common African black allele
RH	<i>RHCE</i>	c.48(chr1:25,747,230)G>C [p.Trp16Cys]	Genome: common African black allele
LU	<i>BCAM (LU)</i>	c.1615(chr19:45,322,744)G>A [p.Ala539Thr]	Antigenic difference: Au(a-b+) > Au(a+b-)
LE	<i>FUT3</i>	c.202(chr19:5,844,649)T>C [p.Trp68Arg], c.314(chr19:5,844,537)C>T [p.Thr105tMet]	Genome: inactive enzyme associated with a Le(a-b-) phenotype
FY	<i>ACKR1 (DARC)</i>		Conventional: reference is alternative splice form with different numbering than the alleles
DI	<i>SLC4A1</i>	c.357(chr17:42,337,900)T>C [p.Val119Val]	Silent change
YT	<i>ACHE</i>		Conventional reference is alternative splice form with different numbering and splice form was not deposited into GenBank
DO	<i>ART4</i>	c.378(chr12:14,993,854)C>T [p.Tyr126Tyr], c.624(chr12:14,993,608)T>C [p.Leu208Leu]; c.793(chr12:14,993,439)A>G [p.Asn265Asp]	Silent changes; Antigenic difference: Do(a+b-) > Do(a-b+)
H	<i>FUT2</i>		Conventional reference is numbered to alternative splice form rather than the deposited long isoform
KN	<i>CR1</i>	c.4828(chr1:207,782,916)T>A [p.Ser1610Thr]	Genome: rare SI3- phenotype
IN	<i>CD44</i>	c.326(chr11:35,201,913)A>C [p.Tyr109Ser]	Found in association with rare In(a+b-) phenotype but with additional changes c.137G>C p.Arg46Pro and c.716G>A p.Gly239Glu. It is unclear if c.326A>C alone can lead to antigenic change.
OK	<i>BSG</i>	c.537(chr19:581,407)T>C [p.Asp179Asp]	Silent change
RAPH	<i>CD151</i>	c.579(chr11:837,582)A>G [p.Gly193Gly]	Silent change
I	<i>GCNT2</i>	c.816(chr6:10,587,038)G>C [p.Glu272Asp]	Genome: uncommon allele with unclear phenotype
GIL	<i>AQP3</i>	c.61(chr9:33,447,468)T>C [p.Leu21Leu], c.105(chr9:33,447,424)C>G [p.Leu35Leu], c.390(chr9:33,442,952)T>C [p.Phe130Phe], c.543(chr9:33,442,466)T>C [p.Pro181Pro]	Silent changes
RHAG	<i>RHAG</i>	c.724(chr6:49,582,483)G>A [p.Asn242Asp]	Conventional reference sequence sequencing error

(*RHCE*01*) encodes a c+e+ phenotype.⁴⁶⁻⁴⁸ The human reference genome *RHCE* sequence is c.48G>C, chr1:25,747,230G>C (p.Trp16Cys),⁴⁹ which corresponds to *RHCE*ce(48C)* (*RHCE*01.01*), again, an allele more often found in African Americans.

ACKRI previously *DARC* (*FY*)

The gene that encodes the Duffy antigens, Fy^a and Fy^b, has a minor 338 amino acid product (Variant 1, U01839) and a major 336-amino-acid product (Variant 2, NM_002036.3).^{3,50,51} The nucleotide position responsible for the Fy^a/Fy^b phenotype is c.131G/A (p.Gly44Asp) in Variant 1 and c.125G/A (p.Gly42Asp) in Variant 2.⁵² Some allele sources list the reference sequence as U01839,^{2,3} but many of the null allele nucleotide positions did not correlate with the U01839 sequence. The original report⁵² used Variant 1 (which they did not deposit at the time of publication, but corresponds to NM_002036.3). The two sequences differ in length by six nucleotides, and both sequences have a GAC (Gly) codon six nucleotides upstream of the actual Fy^a GAC (Gly) codon, making the disparity in reference sequence difficult to detect.

FUT3

Relative to the conventional reference sequence (X53578), the human reference genome sequence is a reported inactive form of the enzyme with nucleotide changes c.202T>C (chr19:5,844,649T>C) p.Trp68Arg and c.314C>T (chr19:5,844,537C>T) p.Thr105tMet corresponding to a Le(a-b-) phenotype.⁵³

ACHE (*YT*)

ACHE has several alternative splicing variants including: 1) the conventional cDNA reference sequence (Variant 1, M55040, 614 amino acids) and 2) a cDNA sequence that is primarily expressed in erythroid tissue (Variant 2, NM_015831.2, 617 amino acids).^{54,55} Variant 1 and Variant 2 only differ in the C-terminal region and the nucleotide numbering of the only known antigens (Yt^a and Yt^b) are not affected by this difference. Published allele source lists the conventional cDNA reference sequence (Variant 1), but shows the amino acid sequence for Variant 2.³

FUT2

The *FUT2* gene product has two isoforms: a 332-amino-acid short isoform and a 334-amino-acid long isoform (extra 11 amino acid N-term). The original *FUT2* paper found both isoforms, but although secretor mutations were referenced to the short isoform, only the long isoform was submitted (U17894).⁵⁶ Subsequent alleles have continued to be referenced to the short isoform, but incorrectly list the long isoform as reference.^{2,3} We took the long isoform (UCSC transcript: uc002pke.4) and removed the first 33 nucleotide (11 amino acids) so that the allele positions would correlate with those published.

CR1 (*KN*)

Relative to the conventional reference sequence (Y00816), the human reference genome is c.4828T>A (chr1:207,782,916T>A) p.Ser1610Thr, which corresponds to the allele encoding lack of the high frequency Knops antigen (Sl3) and a Sl3- (Sl:1,-2,-3) phenotype.⁵⁷ The Exome Variant Server⁵⁸ was used to determine the allele frequency for c.4828T>A, which is 2.5% (207/8041) European Americans and 0.4% (16/3818) African Americans.

GCNT2 (*I*)

Relative to the conventional reference sequence for *GCNT2* (AF458026), the human reference genome is c.816G>C (chr6:10,587,038G>C) p.Glu272Asp, which according to one source² is the null allele *GCNT2*N.03* that encodes for an I- (i adult) phenotype associated with cataracts. However, although c.816G>C was found in an individual with an I- (i adult) phenotype, it was present with another change c.1006G>A, Gly336Arg (*GCNT2*N.04*).⁵⁹ BGMUT indicates c.816G>C has been found in both adult I+ and I- (i adult) individuals. The Exome Variant Server⁵⁸ was used to determine the allele frequency for c.816G>C (dbSNP rs539351) as 0.1% (11/8589) European Americans and 0.05% (2/4404) African Americans.

RHAG

Relative to the conventional reference sequence for *RHAG* (X64594) the human reference genome is c.724G>A (chr6:49,582,483G>A) p.Asp242Asn. However, aside from the original *RHAG* report (X64594),⁶⁰ all subsequent sequences are c.724A p.Asn242 (AF031549, AF179682, AF179684, AF179685, AF187847, AF178841), and dbSNP indicates that c.742A p.Asn242 (rs1058063) has an allele frequency of 100% and is found in 590 of 590 tested chromosomes from a mix of Europeans, Asians, and Africans. Therefore, the c.724G in X64594 was likely a sequencing error with c.724A being the correct nucleotide.

Comprehensive whole genome antigen prediction

WGS data from a 47-year-old female of European ethnicity in generally good health were first analyzed to determine the sequencing coverage of the genes encoding RBC and PLT antigens. For genes encoding the RBC antigens there was an average coverage of 34× over 1,091,334 bp (Fig. 2, Fig. S1 [available as supporting information in the online version of this paper], Tables 1 and 3). For genes encoding PLT antigens, there was an average coverage of 38× over 323,222 bp (Fig. 2, Fig. 1S, Table 4). There were some regions with missing sequence coverage and/or poor sequencing quality in the following RBC genes: *RHD* (Exon 8), *C4B*, *C4A*, and *CR1* (Fig. 2, Fig. 1S). However, all of the RBC and PLT genes had adequate sequencing coverage (Fig. 1S) and quality to allow for prediction of phenotypes

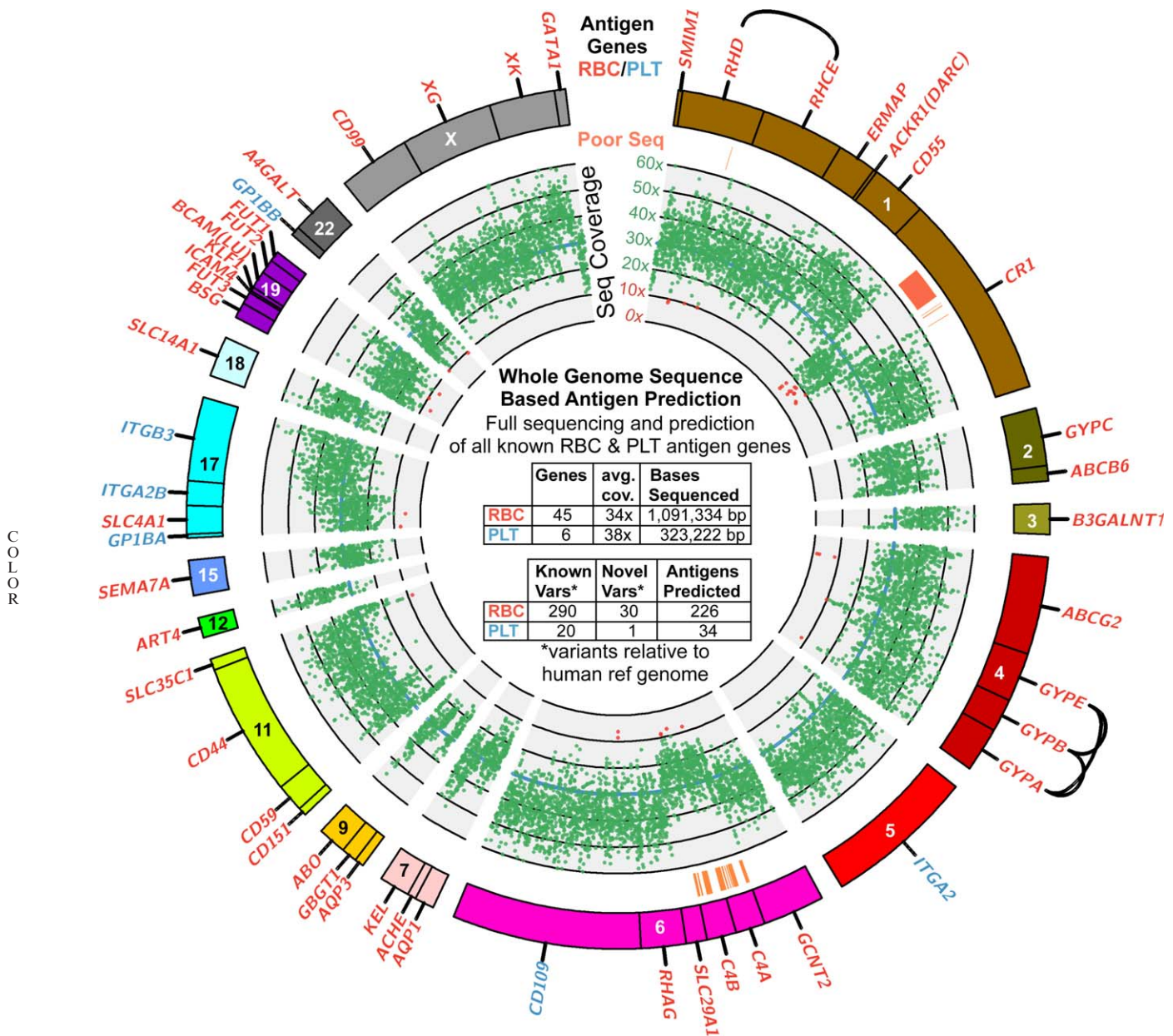


Fig. 2. WGS-based RBC and PLT gene sequencing. Circos plot⁶¹ of the WGS data that has been filtered to only show the RBC and PLT genes with a circular plot of the sequence coverage (100-bp bins).

from the known allele nucleotide positions. The low coverage for *RHD* Exon 8 is likely due to the human reference genome Exon 8 containing a mismatched *RHD*DAU0* allele change. In addition, without the presence of the *RHD*DAU0* allele, Exon 8 is identical in both *RHD* and *RHCE*. Therefore, *RHD* Exon 8 sequences either misaligned to *RHD* and/or did not align at all.

Variant calling on WGS data determined the nucleotide positions that differed in relation to the human reference genome. The sequence alignments between the human reference genome CDS and the cDNA reference sequence were then used as a guide to convert genomic

coordinates from the variant calling process into the conventional CDS positions. By combining the human reference genome antigen predictions with manual identification of the CDS converted variants using published allele tables,^{2-4,10,12,23} the WGS data were used to comprehensively predict all RBC and PLT antigens (Tables 3 and 4). As part of the process, nucleotide changes were found that are not known to encode antigenic epitopes; while most were silent changes that did not alter the amino acid sequence, there were a few missense changes that do alter the amino acid sequence (Table 5). A RBC sample was tested for RBC antigens using available commercial

TABLE 3. Comprehensive RBC antigen prediction from a patient's whole genome*

System	Gene	Average coverage	Phenotypes
001 ABO	<i>FUT1, ABO</i>	27×, 31×	A1
002 MNS	<i>GYP A</i>	41×	M+N+, Vr-, Mt(a-), Ri(a-), Ny(a-), Or-, ERIK-, Os(a-), ENEP+, ENEH+, ENAV+, ENEV+, MN TD-
003 P1PK	<i>GYP B</i>	45×	S+s+, U+, En(a+), He-, Mi(a-), Mur-, Mv-, s(D-), Mit-, Dantu-
004 RH	<i>A4GALT</i>	29×	P1+/P1-, pk+, NOR-
	<i>RHD</i>	34×	D+, Tar-
	<i>RHCE</i>	33×	C-c+E+e-, C ^W -, C ^X -, E ^W -, V-, VS-, Rh26+LOCR-, Be(a-), DAK-, Go(a-), Rh32-, Crawford-CELO+, JAL-CEST+, STEM-, JAHK-
005 LU	<i>BCAM(LU), KLF1, GATA1</i>	24×, 21×, 25×	Lu(a-b+), LURC+, Lu4+, Lu5+, Lu6+, Lu7+, Lu8+, Lu13+, Lu16+, Lu17+, Au(a+b-), Lu20+, Lu21+
006 KEL	<i>KEL</i>	33×	K-k+, Kp(a-b+c-), Js(a-b+), Ul(a-), K11+, K12+, K13+, K14+, K18+, K19+, K22+, K23-, VLAN-VONG-, TOU+, RAZ+, KALT+, KTIM+, KYO-, KUCI+, KANT+, KASH+, KELP+, KETI+, KHUL+
007 LE	<i>FUT2,3</i>	28×, 28×, 27×, 23×	Le(a+b-)
008 FY	<i>ACKR1(DARC)</i>	28×	Fy(a+b+)
009 JK	<i>SLC14A1</i>	37×	Jk(a+b+)
010 DI	<i>SLC4A1</i>	26×	Di(a-b+), Wr(a-b+), Wd(a-), Rb(a-), WARR-, ELO-, Bp(a-), Mo(a-), Hg(a-), Vg(a-), Sw(a-), BOW-, NFLD-, Jn(a-), KREP-, Tr(a-), Fr(a-), SW1-, Wu-DISK+
011 YT	<i>ACHE</i>	23×	Yt(a+b-)
013 SC	<i>ERMAP</i>	37×	Sc1+Sc2-, Rd-, STAR+, SCER+, SCAN+
014 DO	<i>ART4</i>	40×	Do(a+b+), Jo(a+), DOYA+, Hy+, DOMR+, DOLG+
015 CO	<i>AQP1</i>	27×	Co(a+b-), Co4+
016 LW	<i>ICAM4</i>	23×	LW(a+b-)
017 CH/RG	<i>C4B</i>	25×	Ch1+, Ch2+, Ch3+, Ch4+, Ch5+, Ch6+, Rg1-, Rg2-
017 CH/RG	<i>C4A</i>	25×	Ch1-, Ch2-, Ch3-, Ch4-, Ch5-, Ch6-, Rg1+, Rg2+
018 H	<i>FUT1,2,SLC35C1</i>	27×, 28×, 29×	H+
019 XK	<i>XK</i>	39×	Kx+
020 GE	<i>CYP C</i>	32×	Ge2+, Ge3+, Ge4+, Es(a+), Wb-, An(a-), Dh(a-), GEIS-, GELP+, GEAT+, GETI+
021 CROM	<i>CD55</i>	43×	Cr(a+), Tc(a+b-c-), Dr(a+), Es(a+), WES(a-b+), UMC+, GUTI+, SERF+, ZENA+, CROV+, CRAM+, CROZ+
022 KN	<i>CR1</i>	35×	Kn(a+b-), McC(a+b-), Sla+Vil-, Yk(a+), Si3+, KCAM+/KCAM-
023 IN	<i>CD44</i>	39×	In(a-b+), INFI+, INJA+
024 OK	<i>BSG</i>	22×	Ok(a+), OKGV+, OKVM+
025 RAPH	<i>CD151</i>	22×	MER2+
026 JMH	<i>SEMA7A</i>	27×	JMHK+, JMHL+, JMHG+, JMHM+, JMHQ+
027 I	<i>GCNT2</i>	37×	I+
028 GLOB	<i>B3GALNT1</i>	40×	P+
029 GIL	<i>AQP3</i>	27×	GIL+
030 RHAG	<i>RHAG</i>	39×	Duclos+, Ol(a-), DSLK+, RHAG4-
031 FORS	<i>GBGT1</i>	27×	FORS+
032 JR	<i>ABCG2</i>	37×	Jra+
033 LAN	<i>ABCB6</i>	29×	Lan+
034 VEL	<i>SMIM1</i>	22×	Vel+
035 CD59	<i>CD59</i>	36×	CD59.1+
036 AT	<i>SLC29A1</i>	28×	At(a+)

* Serologic RBC confirmation = A+, B-, D+, C-, c+, E+, e-, K-, k+, Fy(a+b+), Jk(a+b+), M+, N+, S+, s+.
FUT1 = active; *FUT2* = inactive; *FUT3* = active; *SLC35C1* = active; *ABO* = active A1; *B3GALNT1* = active; *GCNT2* = active; *GBGT1* = inactive; *KLF1* = active; *GATA1* = active.

serologic typing reagents and all of the antigen predictions were correct for the serologically tested RBC antigens (ABO, D, c, C, e, E, K, k, Fy^a, Fy^b, Jk^a, Jk^b, M, N, S, and s).

DISCUSSION

Advantages of antigen prediction by WGS

In this analysis we showed that it is possible to perform comprehensive RBC and PLT antigen prediction using WGS

data. WGS-based antigen prediction has advantages over current methods such as DNA CHIP, polymerase chain reaction, and Sanger sequencing. Although the current commercial DNA chip-based assays enable antigen prediction, they are limited in the number of SNPs analyzed, which impacts unambiguous allele resolution. Assays for the RH blood group system are not capable of detecting all known variant RH alleles and additional assays need to be performed to determine *RHD* zygosity. Sanger sequencing could be used to determine all of the known alleles, but the

TABLE 4. Comprehensive PLT antigen prediction from a patient's whole genome

Gene	Average coverage	Predicted HPA phenotypes
<i>ITGB3</i>	34×	1a+, 1b+, 4a+, 4b-, 6bw-, 7bw-, 8bw-, 10bw-, 11bw-, 14bw-, 16bw-, 17bw-, 19w-, 21w-, 23bw-, 26bw-
<i>GP1BA</i>	25×	2a+, 2b-
<i>ITGA2B</i>	27×	3a+, 3b-, 9bw-, 20w-, 22bw-, 24bw-, 27bw-, 28bw-
<i>ITGA2</i>	40×	5a+, 5b-, 13bw-, 18w-, 25bw-
<i>GP1BB</i>	21×	12bw-
<i>CD109</i>	39×	15a-, 15b+

method is labor-intensive and requires the development and validation of many individual assays. In contrast, NGS-based sequencing can evaluate whole gene sequences and detect gene rearrangements, and copy number analysis could determine zygosity. Laboratories could use whole genome or exome approaches or develop targeted NGS-based panels that allow for more affordable sequencing of specific genomic regions by pooling patient specimens using molecular barcodes. In addition, the current generation of benchtop NGS instruments have a 24- to 48-hour turnaround time.

Considerations for antigen prediction with WGS

Traditional serologic antigen testing for the most commonly tested antigens (ABO, D, c, C, e, E, K, k, Fy^a, Fy^b, Jk^a, Jk^b, M, N, S, and s), performed independently and without knowledge of the WGS predictions, agreed with the WGS-based antigen predictions. Although the prediction algorithms successfully predicted the ABO, C/c, M/N, and S/s antigens in this first genome analysis, it is anticipated that these antigens might be more challenging to reliably predict in patients with more extensive genomic variation. In general, robust and reliable automated algorithms for predicting ABO and other carbohydrate antigens require the integration of analyses across several genes. Furthermore, the known alleles for the carbohydrate antigens and the duplicated gene families *GYP A/GYP B* and *RHD/RHCE* often rely on multiple distant variant positions and haplotype ambiguities can occur due to the short read length of most current WGS platforms. Resolution of these ambiguities will ultimately require sequencing technologies that allow for longer read lengths, but in the meantime allele population prevalence could be used to select the most likely haplotype.

The correct alignment of NGS sequence reads is anticipated to be more difficult in the duplicated gene families *GYP A/GYP B* and *RHD/RHCE*. For example, the C antigen results from gene transfer of Exon 2 from *RHD* into *RHCE*, thus the NGS reads for a C+ *RHCE* Exon 2 might misalign to *RHD* Exon 2 without the appropriate algorithm. Similar issues with alignment are likely to

TABLE 5. Changes not known to encode new or altered antigenic epitopes

Gene	CDS nucleotide (genomic coordinate) and [amino acid]
<i>GYP A</i>	hom c.38(chr4:145,041,741)A>C [missense p.Glu13Ala] Note: aa position 13 is within the N-term of protein which is cleaved from the native protein.
<i>A4GALT</i>	het c.109(chr22:43,089,849)A>G [missense p.Met37Val]
<i>CR1</i>	het c.3623(chr1:207,753,621)A>G [missense p.His1208Arg]; het c.5480(chr1:207,790,088)C>G [missense p.Pro1827Arg]; hom c.5905(chr1:207,795,320)A>G [missense p.Thr1969Ala]
<i>CD109</i>	hom c.3722(chr6:74,521,947)C>T [missense p.Thr1241Met]

hom = homozygous, het = heterozygous

occur with other gene rearrangements. However, it might be possible to use the sequence read depth along each gene to look for misaligned sequences to infer the correct antigen or find a rearrangement. NGS rearrangement detection algorithms⁶¹ could also be used to look for the rearrangement breakpoint. Prediction algorithms capable of detecting *RHD/RHCE* rearrangements would be of great value in detecting these potential clinically significant changes in sickle cell patients and pregnant women with weak D or partial D phenotypes. Performing NGS-based RBC predictions on a diverse population of serologically and conventionally molecularly typed individuals will aid development of interpretation algorithms.

Clinical benefits of antigen prediction with WGS

Oncology patients often receive RBC and PLT transfusions. For a minor added cost RBC and PLT antigen prediction could be added to NGS assays already being performed for oncologic diagnosis and drug selection. It might also be possible to replace the current SNP-based antigen typing assays with targeted NGS-based RBC and PLT predictions to aid in difficult serologic work-ups and PLT refractory evaluations and help prevent alloantibody formation in chronically transfused patients. As clinical WGS becomes more commonplace for general disease screening and risk assessment, these existing WGS data could be used for large population level antigen prediction. This would allow for easy identification of donors; assist with compatibility testing of alloimmunized recipients; and prevent alloantibody formation using extended prophylactic matching and the identification of individuals at increased risk for posttransfusion purpura, hemolytic disease of the newborn or fetus, and neonatal alloimmune thrombocytopenia.

Future directions

In this article, we have shown proof of principle that it is possible to comprehensively predict RBC and PLT antigens from WGS data. WGS-based antigen predictions may someday enable accurate determination of blood group antigens, including ABO and RH, at a level of fidelity that cannot be achieved with current DNA chip analysis. To fully realize this potential, we are currently developing and validating prediction algorithms capable of automatically detecting and integrating across the known antigen alleles, which will allow for quick and easy antigen prediction from both WGS and targeted NGS. We are also extending our analysis algorithms for use with the newest human reference genome (GRCh38).

ACKNOWLEDGMENTS

The authors thank the staff and participants of the MedSeq Project for their important contributions. Additional funding was provided by the Brigham and Women's Hospital Pathology Department Stanley L. Robbins M.D. Memorial Research Fund Award.

Members of the MedSeq Project

Members of the MedSeq Project are as follows: David W. Bates, MD, Alexis D. Carere, MA, MS, Allison Cirino, MS, Kurt D. Christensen, MPH, PhD, Robert C. Green, MD, MPH, Carolyn Y. Ho, MD, Lily Hoffman-Andrews, Joel B. Krier, MD, William J. Lane, MD, PhD, Denise M. Perry, MS, Lisa Lehmann, MD, PhD, MSc, Calum A. MacRae, MD, PhD, Cynthia C. Morton, PhD, Christine E. Seidman, MD, Shamil Sunyaev, PhD, Jason L. Vassy, MD, MPH, SM, Rebecca Walsh, Brigham and Women's Hospital and Harvard Medical School; Sandy Aronson, ALM, MA, Ozge Ceyhan-Birsoy, PhD, Siva Gowrisankar, PhD, Matthew S. Lebo, PhD, Ignat Leschiner, PhD, Kalotina Machini, PhD, MS, Heather M. McLaughlin, PhD, Danielle R. Azzariti, MS, Heidi L. Rehm, PhD, Partners Center for Personalized Genetic Medicine; Jennifer Blumenthal-Barby, PhD, Lindsay Zausmer Feuerman, MPH, Leila Jamal, ScM, Kaitlyn Lee, Amy L. McGuire, JD, PhD, Jill Oliver Robinson, MA, Melody J. Slashinski, MPH, PhD, Julia Wycliff, Baylor College of Medicine, Center for Medical Ethics and Health Policy; Philip Lupo, PhD, MPH, Baylor College of Medicine, Department of Pediatrics; Stewart C. Alexander, PhD, Shubhangi Arora, Kelly Davis, Christine Kirby, MS, Peter A. Ubel, MD, Duke University; Peter Kraft, PhD, Harvard School of Public Health; J. Scott Roberts, PhD, University of Michigan; Judy E. Garber, MD, MPH, Dana-Farber Cancer Institute; Dmitry Dukhovny, MD, MPH, Oregon Health & Science University; Tina Hambuch, PhD, Illumina, Inc.; Michael F. Murray, MD, Geisinger Health System; and Isaac Kohane, MD, PhD, Sek Won Kong, MD, Boston Children's Hospital.

CONFLICT OF INTEREST


Dr. Green's research is supported by grants from the National Institutes of Health and Illumina, Inc. Dr. Green has received

compensation for advisory services or speaking from Invitae, Prudential, Arivale, Illumina, AIA, Helix and Roche. The other authors have disclosed no conflicts of interest.

REFERENCES

1. Storry JR, Castilho L, Daniels G, et al. International Society of Blood Transfusion Working Party on red cell immunogenetics and blood group terminology: Cancun report (2012). *Vox Sang* 2014;107:90-6.
2. International Society of Blood Transfusion (ISBT). Red cell immunogenetics and blood group terminology [Internet]. Amsterdam: ISBT Central Office; 2015 [cited 2015 Jul 1]. Available from: <http://www.isbtweb.org/working-parties/red-cell-immunogenetics-and-blood-group-terminology/blood-group-terminology/blood-group-allele-terminology/>.
3. Reid ME, Lomas-Francis C, Olsson ML. The blood group antigen factsbook. 3rd ed. San Diego: Academic Press; 2013.
4. Daniels G. Human blood groups. 3rd ed. Oxford: Wiley-Blackwell; 2013.
5. Ballif BA, Helias V, Peyrard T, et al. Disruption of SMIM1 causes the Vel- blood type. *EMBO Mol Med* 2013;5:751-61.
6. Storry JR, Jöud M, Christophersen MK, et al. Homozygosity for a null allele of SMIM1 defines the Vel-negative blood group phenotype. *Nat Genet* 2013;45:537-41.
7. Cvejic A, Haer-Wigman L, Stephens JC, et al. SMIM1 underlies the Vel blood group and influences red blood cell traits. *Nat Genet* 2013;45:542-5.
8. Daniels G, Ballif BA, Helias V, et al. Lack of the nucleoside transporter ENT1 results in the Augustine-null blood type and ectopic mineralization. *Blood* 2015;125:3651-4.
9. Anliker M, von Zabern I, Höchsmann B, et al. A new blood group antigen is defined by anti-CD59, detected in a CD59-deficient patient. *Transfusion* 2014;54:1817-22.
10. Immuno polymorphism database: IPD-HPA [Internet]. Cambridgeshire: EMBL-EBI; 2015 [cited 2015 May 15]. Available from: <http://www.ebi.ac.uk/ipd/hpa/>.
11. Robinson J, Halliwell JA, McWilliam H, et al. IPD—the Immuno Polymorphism Database. *Nucleic Acids Res* 2013;41:D1234-40.
12. Metcalfe P, Watkins NA, Ouwehand WH, et al. Nomenclature of human platelet antigens. *Vox Sang* 2003;85:240-5.
13. 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:D1023-9.
14. Wagner FF. RhesusBase [Internet]. Springer: DRK Blutspendedienst NSTOB; 2015 [cited 2015 Jun 22]. Available from: <http://www.rhesusbase.info/>.
15. 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.
16. Liu Z, Liu M, Mercado T, et al. Extended blood group molecular typing and next-generation sequencing. *Transfus Med Rev* 2014;28:177-86.

17. Chou ST, Westhoff CM. The role of molecular immunohematology in sickle cell disease. *Transfus Apher Sci* 2011;44:73-9.
18. Avent ND, Martinez A, Flegel WA, et al. The Bloodgen Project of the European Union, 2003-2009. *Transfus Med Hemother* 2009;36:162-7.
19. Storry JR, Olsson ML. The ABO blood group system revisited: a review and update. *Immunohematology* 2009;25:48-59.
20. Svensson L, Rydberg L, de Mattos LC, et al. Blood group A(1) and A(2) revisited: an immunochemical analysis. *Vox Sang* 2009;96:56-61.
21. Liu Y, Fujitani N, Koda Y, et al. Presence of H type 3/4 chains of ABO histo-blood group system in serous cells of human submandibular gland and regulation of their expression by the secretor gene (FUT2). *J Histochem Cytochem* 1999;47:889-94.
22. Lofling JC, Hauzenberger E, Holgersson J. Absorption of anti-blood group A antibodies on P-selectin glycoprotein ligand-1/immunoglobulin chimeras carrying blood group A determinants: core saccharide chain specificity of the Se and H gene encoded alpha1,2 fucosyltransferases in different host cells. *Glycobiology* 2002;12:173-82.
23. Robinson J, Mistry K, McWilliam H, et al. IPD—the Immuno Polymorphism Database. *Nucleic Acids Res* 2010;38:D863-9.
24. Arinsburg SA, Shaz BH, Westhoff C, et al. Determination of human platelet antigen typing by molecular methods: importance in diagnosis and early treatment of neonatal alloimmune thrombocytopenia. *Am J Hematol* 2012;87:525-8.
25. Lind C, Ferriola D, Mackiewicz K, et al. Next-generation sequencing: the solution for high-resolution, unambiguous human leukocyte antigen typing. *Hum Immunol* 2010;71:1033-42.
26. Gabriel C, Furst D, Fae I, et al. HLA typing by next-generation sequencing—getting closer to reality. *Tissue Antigens* 2014;83:65-75.
27. Shiina T, Suzuki S, Ozaki Y, et al. Super high resolution for single molecule-sequence-based typing of classical HLA loci at the 8-digit level using next generation sequencers. *Tissue Antigens* 2012;80:305-16.
28. Wang C, Krishnakumar S, Wilhelmy J, et al. High-throughput, high-fidelity HLA genotyping with deep sequencing. *Proc Natl Acad Sci U S A* 2012;109:8676-81.
29. Erlich RL, Jia X, Anderson S, et al. Next-generation sequencing for HLA typing of class I loci. *BMC Genomics* 2011;12:42.
30. Bentley G, Higuchi R, Högland B, et al. High-resolution, high-throughput HLA genotyping by next-generation sequencing. *Tissue Antigens* 2009;74:393-403.
31. Chu HT, Lin H, Tsao TT, et al. Genotyping of human neutrophil antigens (HNA) from whole genome sequencing data. *BMC Med Genomics* 2013;6:31.
32. Stabentheiner S, Danzer M, Niklas N, et al. Overcoming methodical limits of standard RHD genotyping by next-generation sequencing. *Vox Sang* 2011;100:381-8.
33. 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;53:2892-8.
34. Fichou Y, Audrézet MP, Guéguen P, et al. Next-generation sequencing is a credible strategy for blood group genotyping. *Br J Haematol* 2014;167:554-62.
35. Giollo M, Minervini G, Scalzotto M, et al. 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 U S A* 2012;109:11920-7.
37. Church GM. The personal genome project. *Mol Syst Biol* 2005;1:2005 0030.
38. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 2011;7:539.
39. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013;14:178-92.
40. 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.
41. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008;456:53-9.
42. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 2010;26:589-95.
43. 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.
44. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26:841-2.
45. Thuresson B, Westman JS, Olsson ML. Identification of a novel A4GALT exon reveals the genetic basis of the P1/P2 histo-blood groups. *Blood* 2011;117:678-87.
46. Wagner FF, Ladewig B, Angert KS, et al. The DAU allele cluster of the RHD gene. *Blood* 2002;100:306-11.
47. Chou ST, Jackson T, Vege S, et al. High prevalence of red blood cell alloimmunization in sickle cell disease despite transfusion from Rh-matched minority donors. *Blood* 2013;122:1062-71.
48. Westhoff CM, Vege S, Horn T, et al. RHCE*ceMO is frequently in cis to RHD*DAU0 and encodes a hr(S) -, hr(B) -, RH:-61 phenotype in black persons: clinical significance. *Transfusion* 2013;53:2983-9.
49. Westhoff CM, Silberstein LE, Wylie DE, et al. 16Cys encoded by the RHce gene is associated with altered expression of the e antigen and is frequent in the R0 haplotype. *Br J Haematol* 2001;113:666-71.
50. Chaudhuri A, Polyakova J, Zbrzezna V, et al. Cloning of glycoprotein D cDNA, which encodes the major subunit of the Duffy blood group system and the receptor for the Plasmodium vivax malaria parasite. *Proc Natl Acad Sci U S A* 1993;90:10793-7.

51. Iwamoto S, Li J, Omi T, et al. Identification of a novel exon and spliced form of Duffy mRNA that is the predominant transcript in both erythroid and postcapillary venule endothelium. *Blood* 1996;87:378-85.
52. Iwamoto S, Omi T, Kajii E, et al. Genomic organization of the glycoprotein D gene: Duffy blood group Fya/Fyb alloantigen system is associated with a polymorphism at the 44-amino acid residue. *Blood* 1995;85:622-6.
53. Elmgren A. Significance of individual point mutations, T202C and C314T, in the human Lewis (FUT3) gene for expression of Lewis antigens by the human alpha (1,3/1,4)-fucosyltransferase, Fuc-TIII. *J Biol Chem* 1997;272:21994-8.
54. Li Y, Camp S, Rachinsky TL, et al. Gene structure of mammalian acetylcholinesterase. Alternative exons dictate tissue-specific expression. *J Biol Chem* 1991;266:23083-90.
55. Bartels CF, Zelinski T, Lockridge O. Mutation at codon 322 in the human acetylcholinesterase (ACHE) gene accounts for YT blood group polymorphism. *Am J Hum Genet* 1993;52:928-36.
56. Kelly RJ, Rouquier S, Giorgi D, et al. Sequence expression of a candidate for the human Secretor blood group alpha(1,2)-fucosyltransferase gene (FUT2). Homozygosity for an enzyme-inactivating nonsense mutation commonly correlates with the non-secretor phenotype. *J Biol Chem* 1995;270:4640-9.
57. Moulds JM, Zimmerman PA, Doumbo OK, et al. Expansion of the Knops blood group system and subdivision of Sl(a). *Transfusion* 2002;42:251-6.
58. NHLBI GO Exome Sequencing Project (ESP) Exome Variant Server [Internet]. Seattle (WA): University of Washington; 2015 [cited 2015 Jul 15]. Available from: <http://evs.gs.washington.edu/EVS/>.
59. Inaba N, Hiruma T, Togayachi A, et al. A novel I-branching beta-1,6-N-acetylglucosaminyltransferase involved in human blood group I antigen expression. *Blood* 2003;101:2870-6.
60. Ridgwell K, Spurr NK, Laguda B, et al. Isolation of cDNA clones for a 50 kDa glycoprotein of the human erythrocyte membrane associated with Rh (rhesus) blood-group antigen expression. *Biochem J* 1992;287:223-8.
61. Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. *Genome Res* 2009;19:1639-45.
62. Drier Y, Lawrence MS, Carter SL, et al. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res* 2013;23:228-35. 

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Fig. S1. WGS antigen gene coverage.