Published in final edited form as:

Transfusion. 2018 November ; 58(11): 2693–2704. doi:10.1111/trf.14953.

Genomic Coordinates and Continental Distribution of 120 Blood Group Variants Reported by the 1000 Genomes Project

Celina Montemayor-Garcia¹, Panagiota Karagianni², David A. Stiles¹, Erika M. Reese¹, Danielle A. Smellie¹, Debrean A. Loy¹, Kimberly Y. Levy¹, Magdalene Nwokocha¹, Marina U. Bueno¹, Jeffery L. Miller³, Harvey G. Klein¹

¹Department of Transfusion Medicine, NIH Clinical Center, National Institutes of Health, Bethesda, MD, USA

²B.S.R.C. Alexander Fleming, Vari, Greece

³Genetics of Development and Disease Branch, NIDDK, National Institutes of Health, Bethesda, MD, USA

Abstract

Background—The 1000 Genomes Project provides a database of genomic variants from whole genome sequencing of 2,504 individuals across 5 continental superpopulations. This database can enrich our background knowledge of worldwide blood group variant geographic distribution, and identify novel variants of potential clinical significance.

Corresponding author: Celina Montemayor MD PhD; Department of Transfusion Medicine, Clinical Center, National Institutes of Health, Bldg10, Room 1C711, Bethesda, MD 20892. celina.montemayorgarcia@nih.gov.

Authorship Contributions

C.M.G. conceived the project and wrote the manuscript; C.M.G., P.K., H.G.K. and J.L.M. designed the project; C.M.G., P.K., and D.A.S. queried all blood group alleles with the 1000 Genomes database, extracted coordinates and frequencies, identified novel variants and evaluated for possible clinical relevance, and built the tables; E.M.R., D.S., D.A.L., K.Y.L., M.N., and M.U.B. verified coordinate and LiftOver accuracy and extracted additional database values; all authors reviewed and edited the manuscript.

Web Resources

Wetterstrand K. DNA Sequencing Costs: Data from the NIH GRI Genome Sequencing Program (GSP): <http://www.genome.gov/sequencingcostsdata> (accessed May 30, 2018)

The International Society of Blood Transfusion (ISBT): <http://www.isbtweb.org> (accessed May 30, 2018)

The Blood Group Antigen Gene Mutation Database (dbRBC): <https://www.ncbi.nlm.nih.gov/projects/gv/mhc/xslcgi.cgi?cmd=bgmuthome> (now archived, accessed September 29, 2017)

UCSC Genome Browser: <https://genome.ucsc.edu> (accessed May 30, 2018)

USCS Variant Annotation Integrator: https://genome.ucsc.edu/cgi-bin/hgVai?hgid=584663737_VIwxalgvz0fyL96HqMxz6IJ152R (accessed May 30, 2018)

SIFT: <http://sift.jcvi.org> (accessed May 30, 2018)

PolyPhen-2 prediction of functional effects of human nsSNPs: <http://genetics.bwh.harvard.edu/pph2> (accessed May 30, 2018)

Mutation Taster: <http://www.mutationtaster.org> (accessed May 30, 2018)

Mutation Assessor functional impact of protein mutations: <http://mutationassessor.org/r3> (accessed May 30, 2018)

Index of 1000 Genomes exome pull down targets: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/exome_pull_down_targets/

(accessed May 30, 2018)

gnomAD browser beta | genome Aggregation dataset: <http://gnomad.broadinstitute.org/> (for reader's reference)

TOPMed Freeze5 on GRCh38: <https://bravo.sph.umich.edu/freeze5/hg38/> (for reader's reference)

Conflict of Interest: The authors declare that they have no conflicts of interest relevant to the manuscript submitted to TRANSFUSION.

Disclosure of Conflicts of Interest

The authors declare having no competing financial interest relevant to this article.

Statement of Disclaimer: The views expressed do not necessarily represent the view of the National Institutes of Health, the Department of Health and Human Services, or the U.S. Federal Government.

Study Design and Methods—The 1000 Genomes database was analyzed to 1) expand knowledge about continental distributions of known blood group variants; 2) identify novel variants with antigenic potential and their geographic association; and 3) establish a baseline scaffold of chromosomal coordinates to translate Next Generation Sequencing (NGS) output files into a predicted RBC phenotype.

Results—Forty-two genes were investigated. A total of 604 known variants were mapped to the GRCh37 assembly; 120 of these were reported by 1000 Genomes in at least one superpopulation. All queried variants, including the *ACKR1* promoter silencing mutation, are located within exon pull-down boundaries. The analysis yielded 41 novel population distributions for 34 known variants, as well as 12 novel blood group variants that warrant further validation and study. Four prediction algorithms collectively flagged 79 of 109 (72%) known antigenic or enzymatically-detrimental blood group variants, while 4 of 12 variants that do not result in an altered RBC phenotype were flagged as deleterious.

Conclusion—NGS has known potential for high-throughput and extended RBC phenotype prediction; a database of GRCh37 and GRCh38 chromosomal coordinates for 120 worldwide blood group variants is provided as a basis for this clinical application.

Keywords

1000 Genomes Project; Next Generation Sequencing; Blood Group Genomics

Introduction

Fifteen years after the completion of the Human Genome Project, sequencing technology has evolved to allow inexpensive and rapid whole genome sequencing of thousands of individuals. In 2015, these efforts culminated in the release of the 1000 Genomes Project, an international collaboration that sequenced the genome of 2,504 individuals from 26 populations.^{1,2} Subjects were grouped into 5 superpopulations corresponding to the major continental groups sampled: Africa, Europe, South Asia, East Asia, and the Americas.¹ This project provided a worldwide snapshot of human genomic variation, validating a substantial fraction of the previously known genomic polymorphisms and revealing millions of novel variants.^{1,2}

Given the continuous decline in sequencing costs and the increasing throughput of this technology, the major challenge in the last few years has been to interpret this vast amount of data and translate it into medical practice.³ A number of successful clinical applications have been reported,⁴ and further efforts are underway in the areas of oncology, hemostasis, obstetrics, and pharmacology to identify the clinical benefits of genomic medicine.^{3,5} Transfusion Medicine is an additional discipline with a strong genetics foundation that is exploring this approach.^{6–17}

The genetic basis of blood groups has been largely elucidated and applied by commercially available red blood cell (RBC) genotyping platforms. The International Society for Blood Transfusion (ISBT) recognizes 36 blood group systems, encompassing 41 genes and over 1,100 haplotype alleles. Single nucleotide variants (SNVs) in coding sequences account for

the majority of worldwide RBC antigen variations, but nucleotide changes in splice sites and transcription factor binding sites, as well as insertions, deletions, and gene rearrangements, are also documented. This catalogue of blood group genomic variations was one of the first tools in population genetics,¹⁸ but most of the background knowledge of RBC phenotype distribution derives from studies on limited ethnic groups from a single geographic location.

Accurate blood donor and patient RBC antigen typing has been historically accomplished by serology, representing the earliest example of personalized medicine.¹⁹ However, RBC antigen prediction through genotyping has proven advantageous for a growing number of clinical scenarios, including recently transfused patients, patients with antibodies to high or low-incidence antigens for whom serology has limited value (especially “historical” antibodies no longer detectable), and patients receiving monoclonal therapies such as anti-CD38 and anti-CD47 that obscure traditional typings.^{20–22} Most commercially available RBC genotyping platforms employ allele-specific probe or primer sets, addressing a limited number of genes and associated variants. Typically these tests query the most common SNVs for a given blood group gene, failing to interrogate gene segments where rare, but clinically significant, antigenic or null mutations are determined. These assays are thus often blind to novel variants that may be encountered in clinical practice, and, in the absence of serologic data, may pose a risk of misclassifying and inappropriately transfusing patients carrying rare alleles.¹⁴

By sequencing entire genomes or exomes, Next Generation Sequencing (NGS) may permit a more comprehensive and accurate RBC phenotype prediction, an application that has been demonstrated by a growing number of publications.^{6,7,9–11,13,15,16} We queried the 1000 Genomes NGS data with a focus on 42 blood group-related genes and 3 main objectives: 1) expand knowledge about continental distributions of known blood group variants; 2) identify novel blood group variants with antigenic potential and their geographic association; and 3) establish a baseline scaffold of chromosomal coordinates in the 2 most recent human reference genome builds to translate NGS output files into a predicted RBC phenotype.

Materials and Methods

Genes of interest

A comprehensive list of 48 blood group-related genes was created, based on the ISBT and the archived dbRBC²³ databases. Table 1 includes the complete list of the interrogated genes with the corresponding ISBT blood group number, name, and symbol, if available.

Data analysis

The 1000 Genomes sequencing data were accessed through the UCSC Genome Browser²⁴ GRCh37 (hg19) assembly, focusing on variant call information and superpopulation frequencies. We queried the genomic variants associated with 692 blood group alleles, as documented by ISBT, 2 blood group literature references, and the archived dbRBC.^{23,25,26} For this study, only single nucleotide and short (<50bp) insertion/deletion (indel) variants that are individually associated with a blood group allele were evaluated. Chromosomal GRCh37 coordinates, superpopulation frequencies, sequencing depth and associated dbSNP/

dbVar ID were extracted through the UCSC Genome Browser.²⁴ Coordinates were converted to the GRCh38 assembly using the UCSC LiftOver tool and confirmed by manual review. The process of data acquisition and analysis is summarized in Figure 1. All base calls are reported in the plus strand, as would be needed to interpret a typical NGS .vcf output file. SIFT,²⁷ PolyPhen-2,²⁸ Mutation Taster,²⁹ and Mutation Assessor³⁰ prediction algorithms were accessed through UCSC's Variant Annotation Integrator tool and through their individual websites.

Results

Based upon the 1000 Genomes Project sequencing strategy, short NGS reads of rearranged exons might be misaligned and require phenotype-driven optimization of filter thresholds;^{10,16} therefore, five blood group related genes with known rearrangements (*RHD*, *RHCE*, *GYP A*, *GYP B*, and *GYP E*) were not included in this study (Figure 1). *ABO* was also excluded from this analysis, since accurate prediction of this phenotype often requires precise phasing algorithms.^{10,13}

With the remaining 42 genes, we first surveyed the total number and density of associated missense and nonsense variants worldwide, since these account for the majority of known blood group antigenic changes. The number of nonsynonymous variants ranged from 2 (*SMIM1* and *XK*) to 115 (*CR1*), with an average of 28.5 per gene. Although *SMIM1* and *CR1* are the smallest and largest coding sequences in our database, respectively, there was not a strong correlation between coding sequence length and the number of missense polymorphisms ($r^2 = 0.51$), which ranged from 1 variant per 18.76 coding base pairs (bp) for *FUT2*, to 1 variant per 667.5 coding bp in *XK* (Table 1).

Continental Frequencies of Known Blood Group Variants

The 1000 Genomes Database was queried for antigenic, null and weak alleles documented for each of the 42 blood group genes in our final list, including coding and non-coding regions. We did not address the 88 alleles that are defined by more than 1 variant, since this analysis has been conducted previously by Möller et al,¹⁷ and is dependent on the precision of phasing algorithms. We queried a total of 604 variants, but only 120 were reported as present in at least one superpopulation by 1000 Genomes (Figure 1 and Table S1). The average low-coverage whole genome sequencing read depth for the reported SNVs was 18,211, which corresponds to the aggregate read depth of all 2504 individuals and thus is equivalent to the reported mean low-coverage depth for the entire project (7.4x).¹ Additionally, all queried variants (including the Duffy promoter silencing mutation) were subject to targeted exome sequencing by the 1000 Genomes Project, as determined by the 1000 Genomes UCSC track annotation, as well as by intersection with the 1000 Genomes exome pull-down target BED coordinates published online (listed under Web Resources). The reported exome sequencing mean depth for the entire 1000 Genomes Project is 65.7x.¹ Selected blood groups are discussed below and listed in Table 2. Table S1 contains the complete dataset, which includes ISBT system and name, gene, UCSC transcript ID, nucleotide and aminoacid change, GRCh38 coordinates, read depth, and dbSNP/dbVar ID. Thousand Genomes did not identify single-nucleotide or short indel variants in *FUT6*, *XG*,

CD99, C4A, C4B, XK, BSG, SEMA7A, AQP3, SMIM1, CD59, C1GALT1, C1GALT1C1, or GATA1 that met our criteria.

Lutheran

The *BCAM* c.230G>A (Lu^a) variant was identified within the expected continental distribution.^{25,26} The c.326G>A SNV responsible for loss of the Lu5 antigen was found 3 times in Africa. The Lu9 antigen (LU:–6) SNV was documented in Europe, and for the first time in South Asia. The substitution (c.233C>T) that results in the LURC- phenotype (LU:–22), described in only one proband thus far,³¹ was detected twice in East Asia. The continental distribution of genomic variants associated with prediction of Lu8/Lu14, Lu16, and Lu18/Lu19 was as previously reported.^{25,26}

Kell

The K polymorphism (c.578C>T) was found most frequently in Europe, but was present in every ethnic group except for East Asia (likely due to insufficient sampling since the K+k- phenotype has been documented in Japan³²). This antigen is reported most frequently in the Arabian and Sinai peninsulas,³³ populations not sampled by 1000 Genomes. The rare c.388C>T polymorphism associated with KEL:–18 phenotype, documented previously in Caucasians,^{25,26} was identified only in East and South Asia. The genomic change associated with the loss of KEL26 (c.1217G>A), demonstrated previously in one Hispanic and one Native American family,³⁴ was identified only in South Asia. *KEL* c.913G>A (KEL:–30), which has been documented in a single white American proband,³⁵ was detected 3 times within the African population. Continental frequencies and distribution of the SNVs for Kp^a/Kp^b, Js^a/Js^b, U1^a, KEL17, KEL19, KEL31, KEL36, and KEL37 were as expected.^{25,26} One known Kell_{null} variant, the alternative splicing polymorphism c.924+1 g>t associated with *KEL* *02.N13, was identified in heterozygous state in a South Asian individual. *KEL* c.1719 C>T, the SNV that defines *KEL* *02M.05, was documented in the Americas.

Lewis

The c.59T>G *FUT3* polymorphism, which is associated with Le(a-b-) red cells and may affect anchoring of the enzyme with the Golgi membrane,³⁶ was identified in every superpopulation. The c.202T>C and c.1067T>A *FUT3* variants, which lead to a decreased enzymatic activity in transfection experiments,^{36,37} were also detected worldwide. Le⁴⁷⁸ and Le³⁰⁴, both reported to result in an active enzyme in cell culture models, were identified with the expected continental distribution.^{38,39}

Duffy

The SNV that defines the Fy^a/Fy^b polymorphism was detected in every superpopulation, with a frequency corresponding to previous studies.^{25,26} In accordance with previous reports, the silencing –67T>C SNV was detected with highest frequency in Africa,⁴⁰ and was also present in Europe and America. Notably, this silencing variant, which is located in the upstream promoter of the erythroid *ACKR1* transcript, aligns with the 5'UTR of the *ACKR1* transcript expressed in non-erythroid cells. Since exome-sequencing target boundaries were designed according to the RefSeq (non-erythroid) gene, this important

promoter variant is included within the exome pull down target boundaries of this dataset. The c.395G>A variant, described in a single African American,⁴¹ was identified in South Asia.⁴² The FyX associated SNV, c.265C>T, was present in every population except for Africa.

Kidd

The Jk^b-associated SNV (c.838A>G) frequency was close to 50% in Europe, East Asia, and the Americas, but lower in South Asia (37%) and lowest in Africa (22%). Five polymorphisms associated with null phenotype (JK:-3) were identified and their distribution corresponds to previous reports.^{25,26} Finally, four variants associated with weak Kidd alleles were identified; c.28G>A and c.226G>A were both identified in South Asia, Europe and Americas, in addition to their known presence in individuals of African descent.⁴³

Scianna

ERMAP c.169G>A, associated with SC2, was found in Europeans, and for the first time, in South Asians. The SNV linked with SC:-5 (c.139G>A) reported previously in a single proband of Irish and English descent,⁴⁴ was detected in Europe and America. The variant predictive of SC:-7 (c.103G>A), formerly reported in a single proband of German-English/ Native American descent,⁴⁵ was present in Europe, America, and for the first time documented in South Asia.

Dombrock

The c.793A>G variant associated with Do^a was detected in every superpopulation and its frequency is as expected.^{25,26} The rare c.350C>T change responsible for the Jo(a-) phenotype was found in Africa, Americas, and Europe. The c.673T>A SNV, documented in a single Sri Lankan DOLG- proband,⁴⁶ was identified once in South Asia.

Cromer

Six known *CD55* variants were identified. The c.679G>C change, linked to the Cr(a-) phenotype, was found in Africa and with a lower frequency in Europe and the Americas. The c.155G>T substitution (Tc^b) was identified in Africa, in concordance with prior reports.^{25,26} The G>C substitution in the same position (Tc^c) was documented in Europe as expected,^{47,48} but was also found in South Asia. CROM8 was only predicted in Africa by c.245T>G. *CD55* c.748C>T, which has been documented in a single Japanese proband,⁴⁹ was detected in East Asia. The CROM:-17 SNV (c.649T>G), originally identified in a Thai proband,⁵⁰ was found in East Asia as well.

Junior

The *ABCG2* c.34G>A variant is part of 3 null alleles, but it has also been associated with a Jr(a-) phenotype when present in a homozygous state, presumably by affecting the membrane localization of the protein.⁵¹ This variant has been described in Asians and Caucasians,²⁵ but it was detected in every superpopulation by 1000 Genomes. Three other Jr(a-) associated variants were identified, all rare and in a continent-restricted distribution, consistent with previous reports.⁵¹⁻⁵³

Novel Blood Group Variants

In our list of 42 target genes, 1000 Genomes identified 1198 nonsynonymous variants; 116 are individually associated with known blood group alleles and are listed in Table S1. We scrutinized the remaining nonsynonymous variants for possible clinical significance on the basis of overlap with known blood group alleles, or the presence of a nonsense change preceding an antigenic site or a known null-associated variant. We also analyzed small insertions/deletions <50 bp that do not meet the definition of structural variants. Our analysis yielded 12 novel variants that warrant further validation and study, shown in Table 3 and Table S2.

Two missense variants are possibly antigenic. The first, a novel *ERMAP* c.241C>T (p.R81W) variant in Africa and East Asia, targets the same amino acid associated with the known European SCER- allele,⁴⁵ constituting the first possible antigenic Scianna variation described in East Asia. Similarly, the novel *CD55* 679 G>T (p.A227S) change found in Europe targets the amino acid responsible for the Cr(a-) phenotype described in African Americans and in one Hispanic individual.^{54,55}

Table 3 lists a novel East Asian variant in Kell (c.2107G>C in coding strand; p.G703R) that results in the same amino acid substitution as the known *KEL* *02M.04 allele (c.2107G>A), and thus is predicted to affect the expression of this protein. A novel *SLC14A1* 24bp deletion in Africa encompasses the 28G residue that defines *JK**01W.03, and thus possibly alters Kidd protein expression. This novel in-frame deletion is predicted to eliminate Kidd amino acids D3 through V10, located in the intracytoplasmic N-terminal portion. The 59 N-terminal residues of Kidd are required for membrane localization of this protein in oocytes.⁵⁶ Interestingly, the known null *JK**01.N09 allele is also a deletion of 8 aminoacids that overlaps this novel variant only through loss of V10.

The remaining novel variants are nonsense changes; Table S2 indicates the location of the predicted premature stop codon and lists the known antigenic or null blood group alleles that are encoded by variants downstream of it. A total of 12 additional novel premature stop codons were reported by 1000 Genomes in the *KEL*, *FUT6*, *C4B*, *FUT1*, *CD151*, *GBGT1*, *ABCG2*, and *ABCB6* genes, but their location in the coding sequence is distal to all known alleles in the corresponding blood group (data not shown).

Of the nonsynonymous variants identified by 1000 Genomes in our 42 blood group-related genes, only a small portion correspond to known alleles or are classified above as possible novel null or antigenic changes. The challenge is to determine whether any of the remaining nonsynonymous changes is clinically relevant. To determine if available prediction algorithms would be suitable for this purpose, we tested SIFT,²⁷ PolyPhen-2,²⁸ Mutation Taster²⁹, and Mutation Assessor³⁰ with the control set of known blood group variants listed in Table S1. Collectively, at least one of these 4 algorithms flagged 79 of 109 (72%) listed blood group variants that result in antigenic or enzymatic activity changes. Prediction varied widely across algorithms, which unanimously agreed on a possibly damaging prediction only 19% of the time. On the other hand, 4 of the 12 enzymatic blood group variants listed in Table S1 that do not affect the RBC phenotype, such as *FUT3* c.478C>T and *FUT2* c.718G>A, were flagged as damaging by at least one of these algorithms. Some well-

documented blood group variants that were missed by all 4 prediction algorithms include those associated with the prediction of the Fy^a/Fy^b , Jk^a/Jk^b , and Co^a/Co^b phenotypes.

Discussion

We present the genomic coordinates, nucleotide interpretation key, and continental frequencies of 120 blood group SNVs and short indels reported by the 1000 Genomes Project (Table S1). This study is limited to the analysis of the public 1000 Genomes database through the UCSC browser, and the reader is referred to the original publication for a discussion of its false discovery rate and genotyping accuracy estimates.¹ Representation of the 1000 Genomes database may vary across different genome browser platforms. All 1000 Genomes samples are available through the NHGRI and the NIGMS Repositories at the Coriell Institute for validation studies of the genotypes reported here, but phenotype information of the research participants is not available. Analysis of structural variations (SVs) from the 1000 Genomes Project has been published,² but its precision is limited.^{17,57} Homologous genes also pose the risk of misalignment of the short NGS reads produced by this project, and their accurate immunohematology interpretation may require locus-specific adjustment of filter thresholds, as illustrated by recent publications.^{10,16} Thus, in the absence of experimental validation data and targeted filter adjustment, structural variants >50bp and homologous blood group genes with known recombination events (*RHD*, *RHCE*, *GYP A*, *GYP B*, and *GYPE*) were deferred from this study.

This study focuses on individual genomic variants. These variants may be associated with more than one allele (all listed in Table S1), and the final RBC phenotype of an individual ultimately depends on the precise collection of blood group variants and their distribution into haplotypes. For example, an adenine in position 838 of the *SLC14A1* transcript is associated with asparagine in position 280 and the Jk^b epitope, but an accompanying c.871T>C variant on the same chromosome would define the *JK*02N.06* allele and nullify the expected Jk^b expression for that haplotype. Our approach is similar to current RBC genotyping approaches where only a portion of a gene is sequenced, but allowing for a substantially broader scope and for detection of novel variants. The novel variants reported here are predicted purely from genomic approaches, and thus require experimental and serologic confirmation.

The resolution of multiple variants into haplotypes, known as ‘phasing’, represents one advantage of NGS over other genotyping approaches. Haplotypes can be determined directly from NGS (coexistence in a single read or across reads with overlapping heterozygous calls, or experimentally through fosmid cloning) or inferred from reference haplotype panels and known population frequencies.^{58,59} Thus, NGS computational physical phasing depends on the read length and density of heterozygous calls in each genomic area, and prediction algorithms may be limited by the scope of the reference panel and choice of inferential method.^{58,59} Haplotypes from the 1000 Genomes are available (see the ErythroGene database for an in-depth analysis of blood group phased alleles from this dataset).¹⁷ These haplotypes were constructed through placement of bi-allelic, multiallelic, and SVs onto an initial haplotype scaffold constructed from microarray data from the research subjects, and when available, first degree relatives.^{1,2,60} Although this is a highly accurate approach, a

level of imprecision likely remains, given that imputation is part of the process and trio data was not available for 49% of the dataset.¹ Additionally, phase resolution of rare variants, such as several blood group changes described here, is known to be more challenging.⁵⁸ Haplotype accuracy metrics from fosmid cloning in a limited number of six 1000 Genomes samples reports a haplotype concordance rate that ranges from 90.17% to 99.4%, with the 2 samples lacking trio data demonstrating the lowest values.¹

As previously reported,^{7,17} we found multiple discrepancies with current RBC blood group allele databases when mapping alleles to the GRCh37 assembly. This highlights the need for a revised RBC allele database and for incorporating unique chromosomal coordinates into the literature to help resolve uncertainties. Coordinates are indispensable for the interpretation of the most typical NGS analysis output file: a vcf (variant call format) file. Although a number of NGS applications and a blood group software algorithm have been published, publicly available blood group chromosomal coordinates and their interpretation logic are lacking. Möller et al¹⁷ provide a comprehensive list of GRCh37 and GRCh38 coordinates for known and novel blood group variants, and we provide a straightforward nucleotide interpretation key in both assemblies for a set of blood group variants found worldwide. GRCh38 differs from the preceding human genome assembly by including 252 additional alternate loci, the mitochondrial genome, and many rectifications in nucleotide calls, misassembled regions, and previous gaps. Although GRCh37 has been used more extensively in large-scale sequencing projects, GRCh38 provides higher accuracy and richer diversity features.⁶¹ Immunohematologists will notice that blood group variant coordinates often shift in the newest assembly, that *C4A/C4B* gene prediction tracks are more accurate, and that there is an alternate haplotype for the region that encodes for *KEL*. Regardless of the reference build used, a critical item is to maintain consistency throughout the entire NGS analysis pipeline.

Thirty-four of the 120 blood group variants described here have at least one novel population distribution (Figure 2, marked in Tables 2 and S1), with a total of 41 novel blood group continental frequencies – 56% of them observed in South Asia for the first time. Seven blood group variants are described for the first time in Africa and may be of significance in the frequently-transfused sickle cell disease population. Like the general findings of the 1000 Genomes Project,¹ blood group variants with a frequency <0.5% are more often restricted to a single population. With the exception of the SNVs associated with *KEL**02M.05, *ABCB6**01W.04, and *DI**02.03, blood group variants identified in the Americas are shared with another continental group. Known variants that were documented for the first time in the Americas but shared with another continent were not included in the novel population distribution tally, given the known recent, mixed ancestry of this continent and the expected high number of shared variants.¹

Distinguishing neutral passenger variants from deleterious mutations has challenged genomics from its inception, and as NGS is applied to transfusion medicine, novel blood group variants have been a recurrent finding.^{6,9,12,13,15,17} Several prediction tools used widely in genomics have not been tested for immunohematology. Four prediction algorithms failed to properly flag 28% of the clinically significant blood group variants described in this study. This may reflect the discrepancy between prediction of a protein's *function* –for

which these algorithms are designed - and *antigenicity*, the critical concern for transfusion medicine. One basic premise of these algorithms is that evolution selects against damaging mutations in functional residues. However, RBC alloantigenicity depends primarily on residues/modifications exposed to the extracellular milieu, regardless of their role in expression, localization, or function. Many blood group antigenic sites will not be subject to strong evolutionary pressure, as described previously for Diego⁶² and Kell.⁶³ Biochemical differences in amino acid substitutions may also not be as critical for immunohematology, since several proven antigenic variations (Wu/DISK, KEL:–22, KUCI-, Co(b+), etc) result from amino acid substitutions that are considered “conservative”.^{64,65} Möller et al. filtered variants with extracellular localization for possible antigenicity,¹⁷ and Howe et al. reported that Meta-SNP scores can be helpful.⁶⁵ The prediction of the clinical importance of novel blood group variants will likely require algorithms tailored to incorporate expression predictors, protein structure, antigenicity scores, and epitope determinants.^{66,67}

This study contributes toward creation of a transfusion medicine NGS pipeline, integrating blood group polymorphisms into the genome-based medical record for optimal personalized medicine. Similar studies on larger NGS datasets that are now available, such as the ExAC,⁶⁸ gnomAD, and TOPMed projects, will continue to expand information about worldwide genetic variation of blood group genes. Complete sequencing of the germline genome of complex patients could be employed, not only for diagnostic and pharmacologic purposes, but also to predict a comprehensive RBC phenotype that will be useful for lifelong transfusion support. Obtaining unbiased, complete gene sequences would also allow ongoing detection and study of rare and novel variants.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This research was supported by the Intramural Research Program of the National Institutes of Health Clinical Center.

The views expressed do not necessarily represent the view of the National Institutes of Health, the Department of Health and Human Services, or the U.S. Federal Government.

Source of Support: This research was supported by the Intramural Research Program of the National Institutes of Health Clinical Center

References

- 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. *Nature* 2015;526: 68–74. [PubMed: 26432245]
- Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH, Konkel MK, Malhotra A, Stutz AM, Shi X, Casale FP, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJP, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HYK, Mu XJ, Alkan C, Antaki D, Bae T, Cerveira E, Chines P, Chong Z, Clarke L, Dal E, Ding L, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Y, Lameijer EW, McCarthy S, Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R, Shabalin

- AA, Untergasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T, Sebat J, Batzer MA, McCarroll SA, Genomes Project C, Mills RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korb J. An integrated map of structural variation in 2,504 human genomes. *Nature* 2015;526: 75–81. [PubMed: 26432246]
3. Green ED, Guyer MS, National Human Genome Research I. Charting a course for genomic medicine from base pairs to bedside. *Nature* 2011;470: 204–13. [PubMed: 21307933]
  4. Kalia SS, Adelman K, Bale SJ, Chung WK, Eng C, Evans JP, Herman GE, Hufnagel SB, Klein TE, Korf BR, McKelvey KD, Ormond KE, Richards CS, Vlangos CN, Watson M, Martin CL, Miller DT. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet Med* 2017;19: 249–55. [PubMed: 27854360]
  5. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR, Martin CL, Nussbaum RL, Plon SE, Ramos EM, Sherry ST, Watson MS, ClinGen--the Clinical Genome Resource. *N Engl J Med* 2015;372: 2235–42. [PubMed: 26014595]
  6. Fichou Y, Mariez M, Le Marechal C, Ferec C. The experience of extended blood group genotyping by next-generation sequencing (NGS): investigation of patients with sickle-cell disease. *Vox Sang* 2016;111: 418–24. [PubMed: 27442304]
  7. Lane WJ, Westhoff CM, Uy JM, Aguad M, Smeland-Wagman R, Kaufman RM, Rehm HL, Green RC, Silberstein LE, MedSeq P. Comprehensive red blood cell and platelet antigen prediction from whole genome sequencing: proof of principle. *Transfusion* 2016;56: 743–54. [PubMed: 26634332]
  8. Johnsen JM. Using red blood cell genomics in transfusion medicine. *Hematology Am Soc Hematol Educ Program* 2015;2015: 168–76. [PubMed: 26637717]
  9. Schoeman EM, Lopez GH, McGowan EC, Millard GM, O'Brien H, Roulis EV, Liew YW, Martin JR, McGrath KA, Powley T, Flower RL, Hyland CA. Evaluation of targeted exome sequencing for 28 protein-based blood group systems, including the homologous gene systems, for blood group genotyping. *Transfusion* 2017;57: 1078–88. [PubMed: 28338218]
  10. Lane WJ, Westhoff CM, Gleadall NS, Aguad M, Smeland-Wagman R, Vege S, Simmons DP, Mah HH, Lebo MS, Walter K, Soranzo N, Di Angelantonio E, Danesh J, Roberts DJ, Watkins NA, Ouwehand WH, Butterworth AS, Kaufman RM, Rehm HL, Silberstein LE, Green RC, MedSeq P. Automated typing of red blood cell and platelet antigens: a whole-genome sequencing study. *Lancet Haematol* 2018;5: e241–e51. [PubMed: 29780001]
  11. Orzinska A, Guz K, Mikula M, Kulecka M, Kluska A, Balabas A, Pelc-Kopotowska M, Ostrowski J, Brojer E. A preliminary evaluation of next-generation sequencing as a screening tool for targeted genotyping of erythrocyte and platelet antigens in blood donors. *Blood Transfusion* 2018;16: 285–92. [PubMed: 28287381]
  12. Avent ND, Madgett TE, Halawani AJ, Altayar MA, Kiernan M, Reynolds AJ, Li X. Next-generation sequencing: academic overkill or high-resolution blood group genotyping? *ISBT Science Series* 2015;10: 250–6.
  13. Wu PC, Lin YH, Tsai LF, Chen MH, Chen PL, Pai SC. ABO genotyping with next-generation sequencing to resolve heterogeneity in donors with serology discrepancies. *Transfusion* 2018.
  14. Liu Z, Liu M, Mercado T, Illoh O, Davey R. Extended blood group molecular typing and next-generation sequencing. *Transfus Med Rev* 2014;28: 177–86. [PubMed: 25280589]
  15. Jakobsen MA, Dellgren C, Sheppard C, Yazer M, Sprogø U. The use of next-generation sequencing for the determination of rare blood group genotypes. *Transfus Med* 2017.
  16. Chou ST, Flanagan JM, Vege S, Luban NLC, Brown RC, Ware RE, Westhoff CM. Whole-exome sequencing for RH genotyping and alloimmunization risk in children with sickle cell anemia. *Blood Adv* 2017;1: 1414–22. [PubMed: 29296782]
  17. Möller M, Joud M, Storry JR, Olsson M. ErythroGene: a database for in-depth analysis of the extensive variation in 36 blood group systems in the 1000 Genomes Project. *Blood Advances* 2016;1: 240–9. [PubMed: 29296939]
  18. McLellan T, Jorde LB, Skolnick MH. Genetic distances between the Utah Mormons and related populations. *Am J Hum Genet* 1984;36: 836–57. [PubMed: 6591796]
  19. Klein HG, Flegel WA, Natanson C. Red Blood Cell Transfusion: Precision vs Imprecision Medicine. *JAMA* 2015;314: 1557–8. [PubMed: 26355383]

20. Chapuy CI, Nicholson RT, Aguad MD, Chapuy B, Laubach JP, Richardson PG, Doshi P, Kaufman RM. Resolving the daratumumab interference with blood compatibility testing. *Transfusion* 2015;55: 1545–54. [PubMed: 25764134]
21. Nedelcu E, Hall C, Stoner A, Eichbaum Q, Meena-Leist C. Interference of Anti-CD47 Therapy with Blood Bank Testing. *Transfusion* 2017;57: 2.
22. Velliquette RW, Degtyaryova D, Hong H, Lomas-Francis C, Shakarian G, Westhoff CM. Serological Observations in Patients Receiving Hu5F9-G4 Monoclonal Anti-CD47 Therapy. *Transfusion* 2017;57: 1. [PubMed: 28097701]
23. 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. [PubMed: 22084196]
24. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res* 2002;12: 996–1006. [PubMed: 12045153]
25. Daniels G *Human Blood Groups*. 3 ed: Wiley-Blackwell, 2013.
26. Reid M, Lomas-Francis C, Olsson M. *The Blood Group Antigen Facts Book*. 3 ed: Elsevier Ltd, 2012.
27. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res* 2012;40: W452–7. [PubMed: 22689647]
28. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;7: 248–9. [PubMed: 20354512]
29. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010;7: 575–6. [PubMed: 20676075]
30. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* 2011;39: e118. [PubMed: 21727090]
31. Karamatic Crew V, Thornton N, Burton N, Poole J, Search S, Daniels G. Two Heterozygous Mutations in an Individual Result in the Loss of a Novel High Incidence Lutheran Antigen Lurc. *Transfus Med* 2009;19: 10.
32. Hamilton HB, Nakahara Y. The rare Kell blood group phenotype KO in a Japanese family. *Vox Sang* 1971;20: 24–8. [PubMed: 5553615]
33. Abdelaal MA, Anyaegbu CC, al Sobhi EM, al Baz NM, Hodan K. Blood group phenotype distribution in Saudi Arabs. *Afr J Med Med Sci* 1999;28: 133–5. [PubMed: 11205816]
34. Jones J, Reid ME, Oyen R, Harris T, Moscarelli S, Co S, Leger R, Beal C, Cardillo K. A novel common Kell antigen, TOU, and its spatial relationship to other Kell antigens. *Vox Sang* 1995;69: 53–60. [PubMed: 7483493]
35. Lee S, Debnath AK, Wu X, Scofield T, George T, Kakaiya R, Yogore MG 3rd, Sausais L, Yacob M, Lomas-Francis C, Reid ME. Molecular basis of two novel high-prevalence antigens in the Kell blood group system, KALT and KTIM. *Transfusion* 2006;46: 1323–7. [PubMed: 16934067]
36. Mollicone R, Reguigne I, Kelly RJ, Fletcher A, Watt J, Chatfield S, Aziz A, Cameron HS, Weston BW, Lowe JB. Molecular basis for Lewis alpha(1,3/1,4)-fucosyltransferase gene deficiency (FUT3) found in Lewis-negative Indonesian pedigrees. *J Biol Chem* 1994;269: 20987–94. [PubMed: 8063716]
37. Elmgren A, Mollicone R, Costache M, Borjeson C, Oriol R, Harrington J, Larson G. 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. [PubMed: 9268337]
38. Soejima M, Munkhtulga L, Iwamoto S, Koda Y. Genetic variation of FUT3 in Ghanaians, Caucasians, and Mongolians. *Transfusion* 2009;49: 959–66. [PubMed: 19175549]
39. Pang H, Liu Y, Koda Y, Soejima M, Jia J, Schlaphoff T, Du Toit ED, Kimura H. Five novel missense mutations of the Lewis gene (FUT3) in African (Xhosa) and Caucasian populations in South Africa. *Hum Genet* 1998;102: 675–80. [PubMed: 9703429]
40. Howes RE, Patil AP, Piel FB, Nyangiri OA, Kabaria CW, Gething PW, Zimmerman PA, Barnadas C, Beall CM, Gebremedhin A, Menard D, Williams TN, Weatherall DJ, Hay SI. The global distribution of the Duffy blood group. *Nat Commun* 2011;2: 266. [PubMed: 21468018]

41. Vege S, Hue-Roye K, Velliquette RW, Lomas-Francis C, Westhoff CM. A New Duffy Allele, FY*A 395G>A (p.Gly132Asp), Associated with Silencing Fy(a) Expression. *Transfusion* 2013;53: 2. [PubMed: 23294206]
42. Hoher G, Fiegenbaum M, Almeida S. Molecular basis of the Duffy blood group system. *Blood Transfus* 2018;16: 93–100. [PubMed: 28151395]
43. Deal T, Adamski J, Hue-Roye K, Vege S, Lomas-Francis C, Westhoff CM. Two Novel JKA Alleles in a Jk(a+b-) Patient with Anti-Jka. *Transfusion* 2011;51: 2. [PubMed: 21219319]
44. Hue-Roye K, Chaudhuri A, Velliquette RW, Fetis S, Thomas R, Balk M, Wagner FF, Flegel WA, Reid ME. STAR: a novel high-prevalence antigen in the Scianna blood group system. *Transfusion* 2005;45: 245–7. [PubMed: 15660834]
45. Flegel WA, Chen Q, Reid ME, Martin J, Orsini LA, Poole J, Moulds MK, Wagner FF. SCER and SCAN: two novel high-prevalence antigens in the Scianna blood group system. *Transfusion* 2005;45: 1940–4. [PubMed: 16371048]
46. Karamatic Crew V, Poole J, Marais I, Needs M, Wiles D, Daniels G. DOLG, a novel high incidence antigen in the Dombrock blood group system. *Vox Sang* 2011;101: 263.
47. Bell JA, Johnson ST, Moulds M, Tello F, Gutsell NS, Gottschall J. Clinical Significance of anti-Tc(âb) in the second example of a Tc(a-b-) individual. *Transfusion* 1989;29: 1. [PubMed: 2643208]
48. Law J, Judge A, Covert P, Lewis N, Sabo B, McCreary J. A new low frequency factor proposed to be the product of an allele to Tc(a). *Transfusion* 1982;22: 1.
49. Lublin DM, Kompelli S, Storry JR, Reid ME. Molecular basis of Cromer blood group antigens. *Transfusion* 2000;40: 208–13. [PubMed: 10686005]
50. Karamatic Crew V, Poole J, Mathlouthi R, Wall L, Daniels G. A novel Cromer blood group system antigen, CRUE, arising from two heterozygous *DAF* mutations in one individual with the corresponding anti-CRUE. *Vox Sang* 2012;103: 211–2.
51. Zelinski T, Coghlan G, Liu XQ, Reid ME. ABCG2 null alleles define the Jr(a-) blood group phenotype. *Nat Genet* 2012;44: 131–2. [PubMed: 22246507]
52. Saison C, Helias V, Ballif BA, Peyrard T, Puy H, Miyazaki T, Perrot S, Vayssier-Taussat M, Waldner M, Le Pennec PY, Cartron JP, Arnaud L. Null alleles of ABCG2 encoding the breast cancer resistance protein define the new blood group system Junior. *Nat Genet* 2012;44: 174–7. [PubMed: 22246505]
53. Hue-Roye K, Lomas-Francis C, Coghlan G, Zelinski T, Reid ME. The JR blood group system (ISBT 032): molecular characterization of three new null alleles. *Transfusion* 2013;53: 1575–9. [PubMed: 23066723]
54. Smith KJ, Coonce LS, South SF, Troup GM. Anti-Cra: family study and survival of chromium-labeled incompatible red cells in a Spanish-American patient. *Transfusion* 1983;23: 167–9. [PubMed: 6836696]
55. Telen MJ, Rao N, Udani M, Thompson ES, Kaufman RM, Lublin DM. Molecular mapping of the Cromer blood group Cra and Tca epitopes of decay accelerating factor: toward the use of recombinant antigens in immunohematology. *Blood* 1994;84: 3205–11. [PubMed: 7524769]
56. Lucien N, Sidoux-Walter F, Roudier N, Ripoche P, Huet M, Trinh-Trang-Tan MM, Cartron JP, Bailly P. Antigenic and functional properties of the human red blood cell urea transporter hUT-B1. *J Biol Chem* 2002;277: 34101–8. [PubMed: 12093813]
57. Baker M Structural variation: the genome's hidden architecture. *Nat Methods* 2012;9: 133–7. [PubMed: 22290183]
58. Snyder MW, Adey A, Kitzman JO, Shendure J. Haplotype-resolved genome sequencing: experimental methods and applications. *Nat Rev Genet* 2015;16: 344–58. [PubMed: 25948246]
59. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. *Nat Rev Genet* 2011;12: 703–14. [PubMed: 21921926]
60. Delaneau O, Marchini J, Genomes Project C, Genomes Project C. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun* 2014;5: 3934. [PubMed: 25653097]

61. Guo Y, Dai Y, Yu H, Zhao S, Samuels DC, Shyr Y. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics* 2017;109: 83–90. [PubMed: 28131802]
62. Jarolim P, Rubin HL, Zakova D, Storry J, Reid ME. Characterization of seven low incidence blood group antigens carried by erythrocyte band 3 protein. *Blood* 1998;92: 4836–43. [PubMed: 9845551]
63. Lee S, Debnath AK, Redman CM. Active amino acids of the Kell blood group protein and model of the ectodomain based on the structure of neutral endopeptidase 24.11. *Blood* 2003;102: 3028–34. [PubMed: 12842980]
64. Castro-Chavez F The rules of variation: amino acid exchange according to the rotating circular genetic code. *J Theor Biol* 2010;264: 711–21. [PubMed: 20371250]
65. Howe JG, Stack G. Structural and functional impacts of amino acid substitutions that create blood group antigens: implications for immunogenicity. *Transfusion* 2017;57: 541–53. [PubMed: 28164302]
66. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, Wheeler DK, Gabbard JL, Hix D, Sette A, Peters B. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res* 2015;43: D405–12. [PubMed: 25300482]
67. Duquesnoy RJ. Antibody-reactive epitope determination with HLAMatchmaker and its clinical applications. *Tissue Antigens* 2011;77: 525–34. [PubMed: 21410655]
68. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won HH, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG, Exome Aggregation C. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;536: 285–91. [PubMed: 27535533]

Figure 1.
Schematic flowchart of methods, data acquisition, and analysis.

Figure 2. Alleles/phenotypes associated with group variants that demonstrate novel superpopulation distributions in the 1000 Genomes Database. Colored stars represent populations of European (blue) or African (yellow) descent sampled within the USA. See 1000 Genomes Project reference manuscript¹ for population sample details. Table S1 provides the full dataset and specifies gene and nucleotide variant queried.

Table 1.

Complete list of genes evaluated and associated blood groups. Genes marked with an asterisk have associated recombination events and were excluded from study.

ISBT No.	Name	Symbol	Gene(s)	Excluded alleles [†]	Mapped blood group variants [‡]	Mapped variants reported by IKG [§]	Total nonsynonymous variants reported by IKG	Nonsynonymous variant density [¶]
1	ABO	ABO	<i>ABO*</i>	-	-	-	-	-
2	MNS	MNS	<i>GYPA*</i>	-	-	-	-	-
			<i>GYPB*</i>	-	-	-	-	-
			<i>GYPE*</i>	-	-	-	-	-
3	PIPK	PIPK	<i>A4GALT</i>	4	32	1 (3%)	20	53.10
4	Rh	RH	<i>RHD*</i>	-	-	-	-	-
			<i>RHCE*</i>	-	-	-	-	-
5	Lutheran	LU	<i>BCAM</i>	2	23	8 (35%)	72	26.21
6	Kell	KEL	<i>KEL</i>	2	69	15 (22%)	59	37.27
7	Lewis	LE	<i>FUT3</i>	30	21	5 (24%)	57	19.05
			<i>FUT6</i>	15	3	0	38	28.42
			<i>FUT7</i>	0	1	1 (100%)	36	27.19
8	Duffy	FY	<i>ACKR1</i>	3	20	4 (20%)	14	72.21
9	Kidd	JK	<i>SLC14A1</i>	3	28	10 (36%)	21	55.71
10	Diego	DI	<i>SLC4A1</i>	1	20	3 (15%)	62	44.13
11	Cartwright	YT	<i>ACHE</i>	0	1	1 (100%)	22	84.27
12	Xg	XG	<i>XG</i>	0	0	0	6	90.5
			<i>CD99</i>	0	0	0	17	32.82
13	Scianna	SC	<i>ERMAP</i>	0	7	3 (43%)	40	35.70
14	Dombrock	DO	<i>ART4</i>	1	13	5 (38%)	26	36.35
15	Colton	CO	<i>AQPI</i>	0	9	1 (11%)	21	38.57
16	Landsteiner-Wiener	LW	<i>ICAM4</i>	0	2	1 (50%)	11	74.18
17	Chido/Rodgers	CH/RG	<i>C4A</i>	1	1	0	18	290.83
			<i>C4B</i>	1	2	0	16	327.1875
18	H	H	<i>FUT1</i>	1	55	5 (9%)	27	40.67
			<i>FUT2</i>	6	28	18 (64%)	55	18.76

ISBT No.	Name	Symbol	Gene(s)	Excluded alleles [†]	Mapped blood group variants [‡]	Mapped variants reported by 1KG [§]	Total nonsynonymous variants reported by 1KG	Nonsynonymous variant density [¶]
19	KX	XK	<i>XK</i>	0	27	0	2	667.50
20	Gerbich	GE	<i>GYPC</i>	1	12	1 (8%)	15	25.80
21	Cromer	CROM	<i>CD55</i>	0	18	6 (33%)	20	57.30
22	Knops	KN	<i>CR1</i>	0	6	3 (50%)	115	53.22
23	Indian	IN	<i>CD44</i>	0	4	1 (25%)	38	28.58
24	Ok	OK	<i>BSG</i>	0	3	0	16	50.63
25	Raph	RAPH	<i>CD151</i>	0	3	1 (33%)	15	50.80
26	John Milton Hagen	JMH	<i>SEMA7A</i>	0	5	0	28	71.46
27	I	I	<i>GCNT2</i>	0	10	3 (30%)	27	44.78
28	Globoside	GLOB	<i>B3GALNT1</i>	0	13	2 (15%)	6	166
29	Gill	GIL	<i>AQP3</i>	0	1	0	8	109.875
30	Rh-associated glycoprotein	RHAG	<i>RHAG</i>	1	31	3 (10%)	23	53.48
31	FORS	FORS	<i>GBGT1</i>	1	3	2 (67%)	48	21.75
32	Junior	JR	<i>ABCG2</i>	4	24	7 (29%)	59	33.36
33	LAN	LAN	<i>ABCB6</i>	5	36	7 (19%)	53	47.72
34	VEL	VEL	<i>SMIMI</i>	0	3	0	2	118.5
35	CD59	CD59	<i>CD59</i>	1	2	0	3	129.00
36	Augustine	AUG	<i>SLC29A1</i>	0	2	1 (50%)	24	57.13
	T/Tn	N/A	<i>C1GALT1C1</i>	2	11	0	10	95.70
			<i>C1GALT1</i>	0	0	0	10	109.20
	In(Lu)	N/A	<i>KLF1</i>	3	54	2 (4%)	23	47.35
	X-linked Lu-mod	N/A	<i>GATA1</i>	0	1	0	15	82.80

[†]Excluded alleles: number of alleles with more than one associated variant, not included in this study

[‡]Mapped blood group variants: number of variants mapped to GRCh37

[§]Mapped variants reported by 1KG: number of mapped variants that were also reported by 1000 Genomes as present in at least one superpopulation

^{||}Total missense variants reported by 1KG: total number of missense variants reported by the 1000 Genomes in the corresponding gene transcript (refer to Table S1 for list of UCSC transcript IDs)

[¶]Missense variant density: 1 missense variant reported by 1000 Genomes per indicated number of coding base pairs.

Table 2.

Known variants for *BCAM* and *KEL* identified by the 1000 Genomes Project. Reference and variant nucleotides are presented in the plus strand. Refer to Table S1 for the complete database, UCSC transcript ID, GRCh38 coordinates, dbSNP/dbVar ID, and prediction analyses.

Variant	Associated antigen or phenotype	hg19 chromosome:nucleotide	Reference (plus strand)	Variant (plus strand)	East Asia	Americas	Africa	Europe	South Asia
<i>BCAM</i> c.230G>A	Lu ^a /Lu ^b	19:45315445	G=Lu ^b	A=Lu ^a	0.000000	0.021600	0.029500	0.028800	0.001000
<i>BCAM</i> c.326G>A	Lu5	19:45315541	G=Lu5	A=LU:-5	0.000000	0.000000	0.002300	0.000000	0.000000
<i>BCAM</i> c.824C>T	Lu6/Lu9	19:45317448	C=Lu6	T=Lu9	0.000000	0.000000	0.000000	0.001000	0.007200*
<i>BCAM</i> c.611T>A	Lu8/Lu14	19:45316704	T=Lu8	A=Lu14	0.000000	0.004300	0.000000	0.011900	0.000000
<i>BCAM</i> c.679C>T	Lu16	19:45316772	C=Lu16	T=LU:-16	0.000000	0.000000	0.006800	0.000000	0.000000
<i>BCAM</i> c.1615A>G	Lu18/Lu19	19:45322744	A=Lu18	G=Lu19	0.113100	0.230500	0.434900	0.300200	0.207600
<i>BCAM</i> c.223C>T	Lu22 (LURC)	19:45315438	C=Lu22	T=LU:-22	0.002000*	0.000000	0.000000	0.000000	0.000000
<i>BCAM</i> c.1495C>T	Lu26 (LUBI)	19:45322624	C=Lu26	T=LU:-26	0.000000	0.000000	0.000000	0.003000	0.009200*
<i>KEL</i> c.578C>T	K/k	7:142655008	G=k	A=K	0.000000	0.021600	0.002300	0.037800	0.006100
<i>KEL</i> c.841C>T	Kp ^a /Kp ^b	7:142651354	G=Kp ^b	A=Kp ^a	0.000000	0.015900	0.001500	0.011900	0.001000
<i>KEL</i> c.1790T>C	Js ^a /Js ^b	7:142640113	A=Js ^b	G=Js ^a	0.000000	0.008600	0.099100	0.000000	0.000000
<i>KEL</i> c.1481A>T	Ula (K10)	7:142641420	T	A=Ula	0.007900	0.000000	0.000000	0.003000	0.000000
<i>KEL</i> c.905T>C	K11/K17(Wka)	7:142651290	A=KEL11	G=KEL17	0.000000	0.000000	0.000000	0.001000	0.000000
<i>KEL</i> c.388C>T	K18	7:142658027	G=KEL18	A=KEL:-18	0.002000*	0.000000	0.000000	0.000000	0.002000*
<i>KEL</i> c.389G>A	K18	7:142658026	C=KEL18	T=KEL:-18	0.000000	0.001400	0.000000	0.000000	0.001000
<i>KEL</i> c.1475G>A	K19	7:142641426	C=KEL19	T=KEL:-19	0.000000	0.000000	0.004500	0.000000	0.000000
<i>KEL</i> c.1217G>A	TOU (KEL26)	7:142643391	C=KEL26	T=KEL:-26	0.000000	0.000000	0.000000	0.000000	0.001000*
<i>KEL</i> c.913G>A	KTIM (KEL30)	7:142651282	C=KEL30	T=KEL:-30	0.000000	0.000000	0.002300*	0.000000	0.000000
<i>KEL</i> c.875G>A	KYO (KEL31)/KEL38	7:142651320	C=KEL38	T=KEL31	0.003000	0.000000	0.000000	0.000000	0.000000
<i>KEL</i> c.1391T>C	KET1 (KEL36)	7:142641752	G=KEL36	A=KEL:-36	0.000000	0.001400	0.000000	0.001000	0.000000
<i>KEL</i> c.877C>T	KHUL(KEL37)	7:142651318	G=KEL37	A=KEL:-37	0.000000	0.000000	0.000000	0.000000	0.001000
<i>KEL</i> c.924+1g>t	Kell null	7:142651270	C= wild-type	A= null (splicing)	0.000000	0.000000	0.000000	0.000000	0.001000

* = novel population distribution of the corresponding known variant. Novel variants in the Americas that are shared with other continents are not marked.

Novel blood group gene variants identified by 1000 Genomes that warrant further study. Reference and variant nucleotides are presented in the plus strand. Refer to Table S2 for complete database, correlation with known variants, dbSNP/dbVar ID, and GRCh38 coordinates.

Table 3.

Gene	hg19 chromosome:nucleotide(s)	Reference	Variant	Predicted aminoacid change	East Asia	Americas	Africa	Europe	South Asia
<i>KEL</i>	7:142638431	C	G	Gly703Arg	0.003000	0.000000	0.000000	0.000000	0.000000
<i>KEL</i>	7:142641475	G	A	Gln476Ter	0.000000	0.000000	0.000000	0.001000	0.000000
<i>SLC14A1</i>	18:43310291–43310314		del of 24 bp	Asp3_Val10del	0.000000	0.000000	0.005300	0.000000	0.000000
<i>ERMAP</i>	1:43296594	C	T	Arg81Trp	0.001000	0.000000	0.001500	0.000000	0.000000
<i>ERMAP</i>	1:43296537	G	T	Glu62Ter	0.001000	0.000000	0.000000	0.000000	0.000000
<i>ERMAP</i>	1:43300805	G	A	Trp177Ter	0.002000	0.000000	0.000000	0.000000	0.000000
<i>FUT2</i>	19:49206298	C	T	Gln29Ter	0.000000	0.000000	0.001500	0.000000	0.000000
<i>FUT2</i>	19:49206445	G	T	Glu78Ter	0.000000	0.000000	0.000800	0.000000	0.000000
<i>FUT2</i>	19:49207104	C	A	Tyr297Ter	0.000000	0.000000	0.000000	0.000000	0.001000
<i>CD55</i>	1:207504467	G	T	Ala227Ser	0.000000	0.000000	0.000000	0.002000	0.000000
<i>ABCB6</i>	2:220083056	C	A	Glu114Ter	0.000000	0.000000	0.000000	0.001000	0.000000
<i>KLF1</i>	19:12995840	G	T	Cys316Ter	0.000000	0.000000	0.000000	0.000000	0.001000