The Journal of Molecular Diagnostics, Vol. 22, No. 10, October 2020

the Journal of Molecular Diagnostics
jmdjournal.org

ACKR1 Alleles at 5.6 kb in a Well-Characterized Renewable US Food and Drug Administration (FDA) Reference Panel for Standardization of Blood Group Genotyping

Kshitij Srivastava,* Pavel P. Khil,† Emilia Sippert,‡ Evgeniya Volkova,‡ John P. Dekker,§ Maria Rios,‡ and Willy A. Flegel*

From the Department of Transfusion Medicine and Laboratory Medicine,† NIH Clinical Center, NIH, Bethesda; the Office of Blood Research and Review,‡ Center for Biologics Evaluation and Research, US Food and Drug Administration, Silver Spring; and the Laboratory of Clinical Immunology and Microbiology,§ National Institute of Allergy and Infectious Diseases, Bethesda, Maryland*

Accepted for publication June 26, 2020.

Address correspondence to Willy A. Flegel, M.D., Laboratory Services Section, Department of Transfusion Medicine, NIH Clinical Center, NIH, 10 Center Drive, Bethesda, MD 20892. E-mail: waf@nih.gov.

The glycoprotein encoded by the *ACKR1* gene expresses the Duffy blood group antigens and is a receptor for malaria parasites. We recently described 18 long-range *ACKR1* alleles in an autochthonous population of a malaria endemic region. Extending this work, we sequenced the gene in a 53-sample repository established by the US Food and Drug Administration (FDA) as reference reagents for blood group genotyping. The FDA samples have been characterized for 19 genes; however, long-range haplotype information for these genes, including *ACKR1*, was lacking. We used a hybrid approach, novel for this type of gene, to characterize *ACKR1* by combining two next-generation sequencing technologies, the short-read massively parallel sequencing and the long-read nanopore sequencing. The expedient integration of data from both next-generation sequencing systems were necessary and sufficient to allow determination of all 25 long-range *ACKR1* alleles found in the 53 samples accurately. All 25 alleles identified in our current FDA cohort were novel and, unexpectedly, none had been observed among the 18 alleles in our previous study. The alleles will be useful for validation, calibration, and proficiency testing of red cell genotyping. The lack of any overlap between the *ACKR1* alleles in the two studies documents differences in mutation rate and recombination frequency among populations. The exact haplotype and their interethnic or interpopulation dissimilarities can influence disease susceptibility and therapy. *(J Mol Diagn 2020, 22: 1272—1279; https://doi.org/10.1016/j.jmoldx.2020.06.014)*

Use of lymphoblastoid cell lines as an unlimited renewable DNA source for external quality assessment schemes has been proposed since 2001[1] and has been implemented sporadically for red cell, human platelet, and neutrophil antigen genotyping.[2] The US Food and Drug Administration (FDA) recently developed a panel of 18 DNA reference reagents from Epstein-Barr virus—transformed cell lines, which can be used as a validated reference for standardization of blood group genotyping, and these reagents were added to the existing collection of 4 World Health Organization International Reference Reagents for blood group genotyping.[3—5] This FDA reference panel is designed to comprise the least number of samples representing the greatest number of genotypes for use as controls for prediction of blood group antigens. In 2018, these

Supported in part by the Intramural Research Program Z99 CL999999 of the NIH Clinical Center (W.A.F.); the National Institute of Allergy and Infectious Diseases at the National Institutes of Health (J.P.D.); and an appointment to the Research Participation Program at the Center for Biologics Evaluation and Research, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration, and the Center for Biologics Evaluation and Research/Office of Blood Research and Review and Office of Minority Health US US Food and Drug Administration (FDA) intramural funds: FDA Critical Path FY2014 and FDA Modernizing Science FY2015-2017 (M.R.).

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

Disclosures: None declared.

Published by Elsevier Inc. on behalf of the Association for Molecular Pathology and American Society for Investigative Pathology.
https://doi.org/10.1016/j.jmoldx.2020.06.014

18 reagents were tested in an international collaborative validation study designed to determine genotypes of predefined genomic loci associated with distinct red cell antigens, thus providing limited sequence information.[3] For example, in the Duffy blood group system, the antigens are predicted based on genotyping only three positions of the *ACKR1* gene, denoted as *c.−67T>C*, *c.125G>A*, and *c.265C>T*.

Since 1992,[6] various studies have been published using red cell genotypes for extended blood group typing,[7] and since 2010[8] also using next-generation sequencing (NGS) chemistries.[9] Reference sequences for blood group genes are important for effective red cell genotyping using NGS.[10] Unlike HLA,[11] there is a lack of experimentally confirmed allele information for blood group genes, which is needed to improve the inference accuracy.[12] A database of full-length alleles of all blood group genes among the samples used by the FDA to develop reference panels will aid in the development, validation, and proficiency testing of new blood group genotyping assays using NGS.[12]

The *ACKR1* gene encodes a multipass transmembrane glycoprotein that carries the five antigens of the Duffy (Fy) blood group system.[13] The two major and clinically most significant antithetical antigens, Fy^a and Fy^b, have been implicated in severe hemolytic transfusion reactions and hemolytic disease of the fetus and newborn.[14–16] Our previous work on the Duffy blood group system established long-range *ACKR1* reference alleles in a native East-African population at 5178 nucleotides using Sanger sequencing.[17] Recently, Fichou et al[18] defined 19 haplotypes at 2488 nucleotides using the third-generation, single-molecule, real-time (SMRT; Pacific Biosciences, Menlo Park, CA) sequencing platform to aid in the imputation and phasing of high-throughput sequencing data.

Genotype phasing is a process that determines if variants found in a gene sequence constitute an allele or a haplotype (*in cis*, on the same chromosome) or belong to two separate alleles (*in trans*, on the two chromosomes of an individual). Genotype phasing is critical for diagnostic purposes. Genotype phasing is also an objective of the analysis for data derived from NGS platforms, especially to detect rare allele combinations.

Phasing of variant sites in long-range nucleotide sequences has been accomplished previously using a combination of massively parallel sequencing (MiSeq; Illumina, San Diego, CA) with low error rate and single-molecule sequencing (GridION; Oxford Nanopore Technologies, Oxford, UK) with a higher error rate.[19] We integrated the application of both platforms and rapidly established long-range *ACKR1* alleles without ambiguity in the 18 samples of the FDA reference panel and 35 additional samples.

Materials and Methods

Samples

The Center for Biologics Evaluation and Research of the FDA provided genomic DNA from 53 volunteer blood donor samples used for the production of B-lymphoblastoid cell lines (Supplemental Table S1). A subset of 18 cell lines was grouped as a reference panel from a renewable source of genomic DNA that was evaluated in a collaborative study[3] and established by the World Health Organization Expert Committee for Biological Standardization as additional International Reference Reagents for Blood Group Genotyping.[20] This FDA reference panel[3] is a publicly available resource and was designed to encompass 41 genetic variants associated with 17 blood group systems present in the original group of 53 blood donors.

The blood samples were collected previously with written informed consent[3] in protocol BC12-15 approved by the Institutional Review Board of the BloodCenter of Wisconsin (Milwaukee, WI) and by the Research Involving Humans Subject Committee of the FDA (protocol number: 11-089B). The blood samples are unlinked from any personal donor information other than the donor's blood group, age, sex, and race, if available. Under the approved protocol, all methods of genetic characterization for blood group genes could be used to study these samples. Thus, no additional approval or consent were needed for the current study.

ACKR1 Long-Range PCR

An *ACKR1* amplicon, 5782-bp long, encompassing the whole *ACKR1* gene including upstream and downstream noncoding regions, was amplified by PCR using the universally tailed primers, 5′-TTTCTGTTGGTGCTGATATTGC-CAACCACTCCTCCCATGGCATT-3′ and 5′-ACTTGCCTGTCGCTCTATCTTC-GATGAGGAGGGGTTTCTGTCC-3′ (Eurofins MWG Operon, Louisville, KY) as described previously.[17] The PCR products were purified using the Agencourt AMPure XP (Beckman Coulter, Brea, CA) and quantified using the Qubit double-stranded DNA high-sensitivity Quantification Kit on a Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA).

Massively Parallel Library Preparation and Sequencing

The nucleotide sequencing covered 5615 nucleotides of the *ACKR1* amplicon. At either end of the *ACKR1* amplicon, 118 nucleotides at the 5′ end and 49 nucleotides at the 3′ end were missing because of insufficient sequencing coverage. Each sample was normalized to a concentration of 300 ng in 30 μL purified water to ensure an equal depth of coverage across the *ACKR1* amplicon. Libraries were prepared (Nextera DNA Flex Library Preparation Kit with Nextera DNA CD Indexes, 96 plex; Illumina) and sequenced (MiSeq Reagent Kit v2; Illumina) using a read length of 2 × 150 bp.

Nanopore Library Preparation and Sequencing

Each sample was normalized to a concentration of 45 ng in 24 μL purified water to ensure an equal depth of coverage across the *ACKR1* amplicon. Libraries were prepared (Ligation Sequencing Kit 1D, SQK-LSK109 with the PCR Barcoding

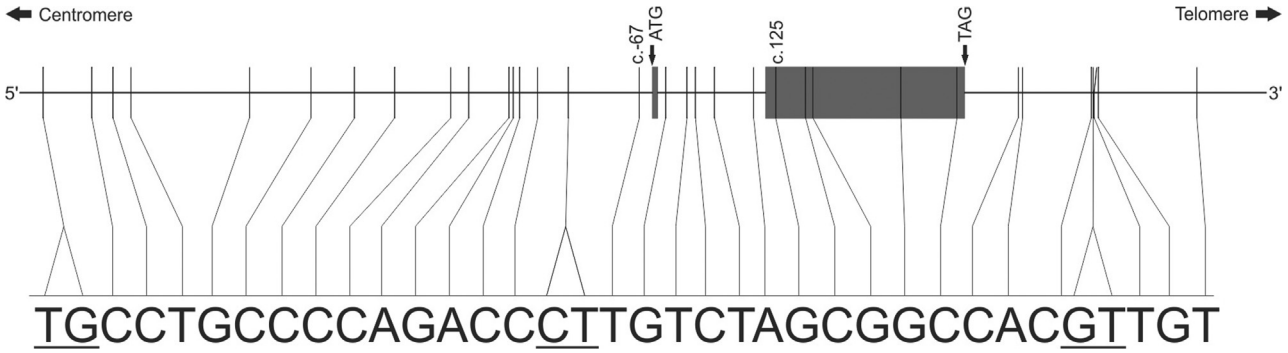

Figure 1 Structure of the *ACKR1* amplicon analyzed and single-nucleotide variants (SNVs) found. Among the 33 variant sites (SNVs, **black bars**) (Table 1), 5 were located in the coding and 28 in the noncoding regions of the nucleotide sequence at chromosomal position 1q23.2 (chromosome 1: 159,201,978 to 159,207,592). The ATG start and TAG stop codons of the coding sequence (**gray box**) are located in exon 1 and exon 2, respectively. The positions for the *GATA box* mutation (c.-67) and *FY*A/FY*B* polymorphism (c.125) are labeled. The nucleotides (**bottom**) represent the alignment of the 30 SNVs and three dinucleotide repeats (**underlined**) (Table 1) distributed throughout the 5615-bp *ACKR1* nucleotide sequence.

Expansion Pack 1-96, EXP-PBC096; Oxford Nanopore Technologies); the samples were combined in equal proportions, loaded onto a single FLOMIN-106 R9 flow cell (Oxford Nanopore Technologies), and sequenced on the GridION X5 platform (Oxford Nanopore Technologies).

Data Analysis

The MiSeq reads were demultiplexed and pairs of FASTQ files were generated (MiSeq software version 2.5.0.5; Illumina). Nanopore reads were base-called in real time on a GridION X5 system and data subsequently were demultiplexed (qcat, version 1.0.7; *https://github.com/nanoporetech/qcat*).[21] BWA-MEM[22] was used to align the data to the human reference genome assembly 38 (hg38) with default settings for the MiSeq reads and modified settings (-x ont2d) for the GridION reads.

Variant Detection and Phasing

Aligned reads (BAM files) were examined visually for variants (Integrative Genomics Viewer; Broad Institute, Boston, MA).[23] In samples containing more than one heterozygote variant, nanopore reads spanning the full-length amplicons were explored visually in Integrative Genomics Viewer to detect phase information (cis/trans relationship) of single-nucleotide variants (SNVs).

Sanger Sequencing

We used Sanger sequencing[17] at three variant positions, *c.-1896G>A*, *c.-1606C>T*, and *c.125G>A*, to determine the accuracy of MiSeq data.

Computational Phasing

The unphased *ACKR1* genotype data from the 53 samples was used with Markov chain-based haplotyper MaCH software version 1.0[24] to statistically infer alleles.[17] Because of the inherent uncertainty of computational phasing, the analysis was performed with MaCH program settings of 2000 rounds and 500 states.[17]

Statistical Analysis

The 95% CIs for allele frequencies were calculated using Poisson distribution.[25] The observed genotype frequencies were examined for deviation from the Hardy–Weinberg equilibrium using a goodness-of-fit χ^2-test with 1 df.

Results

ACKR1 variants, including SNVs and repeats, were detected and their phase information was accurately determined at 5615 bp in 53 DNA samples from the FDA repository. The sequencing covered 1011 nucleotides of the coding sequence, 480 nucleotides of the single intron, 947 nucleotides of the 5′-untranslated region, 50 nucleotides of the 3′-untranslated region, 2035 nucleotides of the 5′-flanking region, and 1092 nucleotides of the 3′-flanking region (Figure 1).

SNV Detection in the *ACKR1* Gene

MiSeq sequencing identified 30 SNVs and 3 dinucleotide repeats in the 53 samples (Table 1). One SNV (rs55872368) was observed as tri-allelic. One SNV was novel, whereas the remaining 29 variants already were listed in the dbSNP database (National Library of Medicine, Bethesda, MD). Besides the *GATA box* mutation (*c.-67T>C*), no other SNV indicative of a nonfunctional allele was detected. Sanger sequencing showed 100% concordance with the MiSeq data for the three variant positions: *c.-1896G>A*, *c.-1606C>T*, and *c.125G>A*.

Table 1 Genetic Variations Detected in the *ACKR1* Gene

Location	Nucleotide change*	dbSNP reference number†	Protein residue change‡	Observations (n = 53)			VAF	HWE (P)
				Homozygote reference	Heterozygote	Homozygote variant		
5′ Flanking region	*-2872_-2871TG>del*	rs5778112	NA	41	6	6	0.170	<0.001
	-2640C>T	rs41313908	NA	45	8	0	0.076	0.552
	-2539C>T	NA	NA	52	1	0	0.009	0.945
	-2456T>G	rs35432289	NA	52	1	0	0.009	0.945
	-1896G>A	rs35333710	NA	47	6	0	0.057	0.662
	-1606C>T	rs6676002	NA	39	14	0	0.132	0.268
	-1400C>T	rs2746047	NA	51	1	1	0.028	<0.001
	-1211C>T	rs3027008	NA	39	14	0	0.132	0.268
5′ UTR	*-947delC*	rs11364458	NA	39	14	0	0.132	0.268
	-863A>G	rs3027009	NA	45	8	0	0.075	0.552
	-673G>A	rs41264467	NA	52	1	0	0.009	0.945
	-655A>G	rs3027011	NA	49	3	1	0.047	0.007
	-627C>T	rs3027012	NA	39	14	0	0.132	0.268
	-541C>T	rs3027013	NA	45	8	0	0.075	0.552
	-399_-398CT>del	rs71782098	NA	49	3	1	0.047	0.007
	-67T>C	rs2814778	NA	39	1	13	0.255	<0.001
Intron 1	*+17G>T*	rs200907215	NA	52	1	0	0.009	0.945
	+115T>C	rs7550207	NA	36	15	2	0.179	0.781
	+150C>T	rs863002	NA	29	18	6	0.283	0.235
	-243T>del	rs17838198	NA	4	15	34	0.783	0.224
	-58A>G	rs3027016	NA	47	6	0	0.057	0.662
Exon 2	*125G>A*	rs12075	Gly42Asp	10	18	25	0.642	0.057
	265C>T	rs34599082	Arg89Cys	51	2	0	0.190	0.889
	298G>A	rs13962	Ala100Thr	39	12	2	0.151	0.396
	714G>A	rs36007769	Gly238=	52	1	0	0.009	0.945
	977C>T	rs17851570	Ser326Phe	52	1	0	0.009	0.945
3′ Flanking region	*+250C>T*	rs12042349	NA	42	11	0	0.104	0.399
	+268A>G	rs863003	NA	36	14	3	0.188	0.318
	+591C>T	rs863004	NA	18	20	15	0.472	0.077
	+596_+597delGT	rs72387739	NA	52	1	0	0.009	0.945
	+599T>C	rs2281301	NA	43	10	0	0.094	0.448
	+616G>T or A	rs55872368	NA	31	13	9	0.292	0.003
	+1083T>C	rs863005	NA	5	15	33	0.764	0.118

*Nucleotide substitutions are shown relative to the reference sequence (*NG_011626.3*). Nucleotide positions are defined using the first nucleotide of the coding sequence of the *NM_002036.3* isoform as nucleotide position 1.

†Publicly available nucleotide sequences, as reported in the National Center for Biotechnology Information Nucleotide database (*https://www.ncbi.nlm.nih.gov/nuccore*, last accessed June 29, 2020).

‡Relative to the National Center for Biotechnology Information Reference Sequence *NP_002027.2*.

HWE, Hardy-Weinberg equilibrium; NA, not applicable; UTR, untranslated region; VAF, variant allele frequency.

ACKR1 Alleles by a Hybrid Sequencing Approach

Detecting phase information in long amplicons with several distant heterozygous variants may be impossible through Sanger sequencing with primer walking. MiSeq sequencing, allowing short read lengths with an average of 300 bp in our study, cannot phase variants that are more than 300 bp apart. By using GridION, full-length, 5.6-kb *ACKR1* amplicons could be sequenced as single reads. Long-range phased *ACKR1* data were thus obtained, and the data sets from the two NGS technologies were analyzed in combination. This hybrid approach, novel for red cell genotyping, integrating MiSeq and GridION data, allowed accurate determination of 25 *ACKR1* alleles at 5615 nucleotides each in the 106 chromosomes analyzed (Table 2).

ACKR1 Results by the Hybrid Approach versus Computational Phasing

By using *ACKR1* genotype information (Supplemental Table S2) as input data, the MaCH software predicted 24 *ACKR1* alleles (Supplemental Table S3). The confirmed *ACKR1* alleles were compared with the simulated results by computational phasing. Of the 25 experimentally confirmed alleles, only 19 alleles (76.0%) were predicted correctly by MaCH, whereas 6 alleles (*MN813502*, *MN813504*,

Table 2 *ACKR1* Allele Distribution in the FDA Reference Samples

GenBank number*	Alignment of variant positions†	Observations, *n*	Allele frequency, %		Ethnicity
			Mean‡	95% CI§	
NG_011626.3	TGCCTGCCCCAGACCCTTGTCTAGCGGCCACGTTGT	NA	NA	NA	NA
MN813501	.....	21	19.8	12.1–29.9	Caucasian, Hispanic
MN813502	.....A.	1	0.9	0.05–5.02	Caucasian
MN813503	...T-.....C	4	3.8	1.3–9.1	Caucasian
MN813504	...T.....T-.A.....T....C	1	0.9	0.05–5.02	Caucasian
MN813505	....-.....T.....C	1	0.9	0.05–5.02	Caucasian
MN813506	....-.....T....C.C	8	7.5	3.1–14.1	Caucasian, African American
MN813507	....-....A.T....C.C	1	0.9	0.05–5.02	Caucasian
MN813508	....-.....TT....C.C	1	0.9	0.05–5.02	Caucasian
MN813509	....-.A.....T...TC	1	0.9	0.05–5.02	Caucasian
MN813510	.....**C**...-.A.....T...TC	4	3.8	1.3–9.1	African American
MN813511	–.....**C**...-.A.....T...TC	14	13.2	7.6–21.6	Caucasian, African American
MN813512	–..G.....**C**...-.A.....T...TC	1	0.9	0.05–5.02	African American
MN813513	–.....T.....**C**...-.A.....T...TC	3	2.8	0.8–7.6	African American
MN813514	.....C.-GA.....T...TC	4	3.8	1.3–9.1	Caucasian, Native American
MN813515	.....A.....C.-GA.....T...TC	1	0.9	0.05–5.02	Caucasian
MN813516	.....TC.-GA.....T...TC	1	0.9	0.05–5.02	Caucasian
MN813517	...T-.A.A...GT–.TC	1	0.9	0.05–5.02	Caucasian
MN813518	...T-.A.....GT....C	4	3.8	1.3–9.1	Caucasian
MN813519	...T-.A.A...GT....C	13	12.3	6.3–20.2	Caucasian
MN813520	...T-.ATA...GT....C	2	1.9	0.3–6.3	Caucasian
MN813521	.....AT.T-...T.....T-.A.....C	3	2.8	0.8–7.6	Caucasian, Hispanic, African American
MN813522	.....AT.T-...T.....T-.A.....	2	1.9	0.3–6.3	Caucasian
MN813523	.....AT.T-...T.....	1	0.9	0.05–5.02	Caucasian
MN813524	..T...T.T-G..TT...C.-.A.....C	8	7.5	3.1–14.1	Caucasian
MN813525	.....G..–**C**.C.-.A.....C	5	4.7	1.8–10.5	African American
Total		106	100	NA	

*Publicly available nucleotide sequences, as reported in GenBank (*https://www.ncbi.nlm.nih.gov/genbank*, last accessed June 29, 2020).

†The nucleotides at the 30 SNV and three dinucleotide repeat (rs5778112, rs71782098, and rs72387739; underlined) positions are shown in 5′- to 3′-orientation (Table 1). The remaining 5579 nucleotide positions had no variation relative to the reference sequence *NG_011626.3*. All nucleotide variants in the *ACKR1* reference are shown. For all other alleles, only nucleotides that differed from the reference are shown. The nucleotide in bold is the *GATA box* mutation (*c.-67T>C*). The dot symbols underneath represent nucleotide positions conserved in all alleles.

‡Number of observed alleles × 100/total number of alleles.

§95% CI, Poisson distribution, two sided.

FDA, Food and Drug Administration; NA, not applicable.

MN813508, *MN813517*, *MN813520*, and *MN813523*) were missed (Supplemental Table S3). Another five alleles (*MaCH-01* to *MaCH-05*) (Supplemental Table S3), not present in any of the 53 samples, were predicted incorrectly by MaCH as single occurrences (Supplemental Tables S3 and S4). Relying on only computerized allele calling would result in 3.8% incorrect allele calls, potentially affecting 1 of 27 individuals (Table 3).

Table 3 *ACKR1* Alleles: Computer Prediction by MaCH Compared with Physical Sequencing

Two alleles per individual	Computational allele prediction by MaCH		
	Predicted alleles, *n*	Individuals, *n*	Rate, %
Both correct	102	51	96.2
Both incorrect	4	2	3.8
Total	106	53	100

Comparison with Ethiopian Alleles

With the exception of rs863005 (*c.+1083T>C*), the other 32 SNV positions were sequenced in our previous study among 60 autochthonous Ethiopian individuals (Table 1).[17] However, all *ACKR1* alleles detected in the present study were novel and not observed among the 18 Ethiopian alleles (Supplemental Table S5). As expected, the present study, with a mixed population, had a lower frequency (20%) of the Duffy-null allele (*FY*02N.01*) than the Ethiopian study population (89%) (Supplemental Table S5).

When comparing the number of variant sites in the *ACKR1* alleles between the Ethiopian and FDA samples, the FDA allele *MN813501* was found to have no difference from the *ACKR1* reference allele (*NG_011626.3*) (Figure 2 and Supplemental Table S6). The FDA allele *MN813501* was found to be closest to the Ethiopian allele *MG932635*, with only two differences at positions *c.-243T>del* (rs17838198) and *c.125G>A* (rs12075).

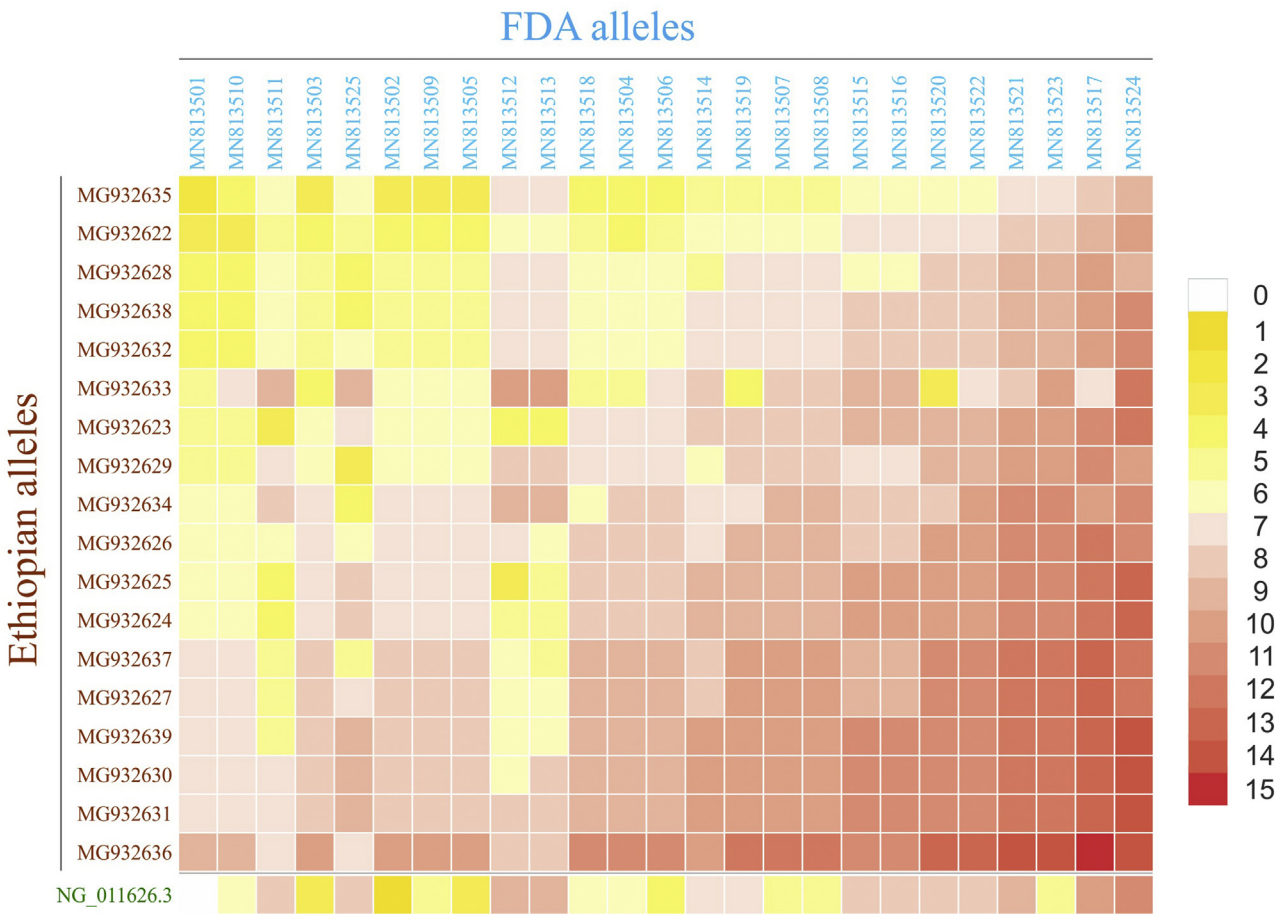

Figure 2 Comparison of the number of variant sites in the *ACKR1* alleles between the Food and Drug Administration (FDA) and African samples. There were 25 alleles among the 53 FDA samples (*MN813501* to *MN813525*) and 18 alleles among the 60 Ethiopian individuals (*MG932622* to *MG932639*).[17] The National Center for Biotechnology Information reference sequence *NG_011626.3* is shown for comparison (green). The colors correspond to the number of SNVs in the *ACKR1* alleles differing between the two studies (Supplemental Table S6).

Distribution of SNVs in the *ACKR1* Gene

The density of SNVs in the coding compared with the noncoding regions did not differ for the 25 alleles present in the FDA samples ($P = 0.670$; χ^2 test, 2-sided), the 18 Ethiopian alleles[17] ($P = 0.642$), and the 43 alleles combined ($P = 0.671$). The Kolmogorov-Smirnov test indicated a normal distribution for the 33 variant positions over the 5.6-kb amplicon in the current 53 FDA samples ($P = 0.96$), for the 18 variant positions in the Ethiopian study[17] ($P = 0.92$), for the 39 variant positions combining the variants in the FDA and Ethiopian samples ($P = 0.98$), and for the 12 variant positions that are shared between the FDA and Ethiopian samples ($P = 0.45$).

Discussion

The Oxford Nanopore Technologies MinION/GridION systems allow large fragments to be sequenced in their entirety, but are known to be error prone,[26] similar to the other long-read sequencing platform from Pacific Biosciences.[27] The Illumina MiSeq system allows only short fragments to be sequenced, although with high accuracy.[28] DNA polymerase errors during amplification such as single-nucleotide substitutions and small insertions or deletions can affect the accuracy of sequencing.[29] In this study, SNVs in the *ACKR1* gene were assembled in 53 DNA samples to confirm 25 alleles at 5.6 kb each without ambiguity. This result was accomplished by using a high-fidelity DNA polymerase on ample genomic DNA template, independent amplification reactions, and a hybrid approach for long-range sequencing, combining MiSeq short-read and GridION long-read sequencing.

The 5.6-kb sequence of the *MN813501* allele, most prevalent among the 53 FDA samples, was identical to the National Center for Biotechnology Information reference sequence *NG_011626.3*. This allele was not found in a previously published study among 60 individuals in Ethiopia.[17] In fact, none of the 25 alleles of the current study matched any of the 18 alleles in the Africans of Ethiopia (Figure 2). The high prevalence of the *GATA box* mutation *c.-67T>C*, common among Africans in malaria-endemic regions, partially can explain the discrepancy of alleles between the two studies, with 40

samples of the current study representing non-African individuals. However, even when the *GATA box* mutation is discounted, no allele was identical between the two studies (Figure 2) and most were highly divergent, differing by up to 15 SNVs. The Ethiopian alleles with *GATA box* mutation (Supplemental Table S5) differed at 4 to 10 SNV positions from the five FDA alleles with *GATA box* mutation (*MN813525*, *MN813511*, *MN813512*, *MN813510*, and *MN813513*). Some SNVs in noncoding sequences, where the majority of our SNVs were found (Figure 1), may impact *ACKR1* gene expression similar to the *GATA box* mutation, which has been studied in detail and is well understood. These data on genetic diversity (Figure 2) thus may have biological relevance and are promising for studies in related African populations, where malaria is not endemic.

A sequencing approach that combined two NGS technologies was used, and it showed the feasibility of accurately identifying more than 5-kb long alleles for a blood group gene. The use of either technology alone is insufficient in a clinical setting to determine long-range alleles with confidence: the second-generation sequencing platforms, such as HiSeq and MiSeq from Illumina, cannot phase distant SNV sites, whereas third-generation sequencing platforms, such as Oxford Nanopore MinION or GridION and Pacific Biosciences RS II or Sequel II, are limited by high rates of indel typing errors.[30] Computerized allele prediction, reliable in predicting common alleles, is inadequate when rare alleles are encountered (Table 3). In addition to this advance in technology, the data drawn from our study can be useful to identify[12] additional *ACKR1* alleles that have an effect on complex diseases such as benign ethnic neutropenia,[31] osteoporosis,[32] and various cancers.[33]

An example of the clinical importance of a correct allele determination is the *GATA box* mutation in the *ACKR1* gene. This SNV (*c.-67T>C*) typically, but not always, occurs *in cis* with the SNV *c.125G>A* and determines a *FY*B* allele (*FY*02N.01*). This allele does not express the Fy^b antigen on red cells, although it induces Fy^b expression on cells of other tissues.[34] As a consequence, individuals carrying at least 1 copy of the *FY*02N.01* allele are not at risk of developing anti-Fy^b[35] and can be transfused with Fy^b antigen−positive red cell units.[15,36] The *FY*02N.01* allele is common in individuals of African ancestry.[36] However, the *GATA box* mutation *in cis* to a *FY*A* allele (*FY*01N.01*) also has been observed, such as in individuals from Papua New Guinea,[37] Sudan,[38] Brazil,[39] and Greece[40]; these individuals can develop anti-Fy^b and should not be transfused with Fy^b-positive red cell units. Although these individuals have a Fy(a−) phenotype, they are tolerized by the Fy^a antigen expressed in their nonerythroid tissues and can be transfused with Fy(a+) blood without risk of developing anti-Fy^a.[40] The two SNVs, *c.-67T>C* and *c.125G>A*, are 972-bp apart at the genomic level and cannot be phased using the Illumina paired reads[41] without parental genotypic data,[42] allele-specific nested PCR,[17] or supplemental statistical phasing that is error prone.[10,17] The long-read nanopore sequencer is capable of producing very long reads to resolve haplotypes without the need for computational phasing. Hence, we applied nanopore sequencing on GridION, which allowed us to detect *ACKR1* alleles successfully from the long-reads *de novo* without parental alleles,[42] nested PCR,[17] or computational phasing.[43]

The *ACKR1* alleles observed in our previous study on the Ethiopian population[17] were not found in the current study, even though 13 African American individuals were among the 53 FDA samples (Supplemental Table S1). The discrepancy of *ACKR1* alleles between the predominantly Caucasian FDA samples and the Ethiopian population stressed the importance of characterizing blood group gene sequences at full length in different populations.[44] Exact long-range alleles can be used to evaluate the influence of genetic variation on the risk of transfusion reactions or diseases.

Red cell genotyping is moving from single locus−based to NGS-based genotyping of red cell antigens. Our data will be useful for NGS platform applications, when their calibration, validation, and proficiency testing are attempted.

Authors Contributions

W.A.F. and M.R. designed the study; J.P.D. provided equipment set-up and discussed the study; K.S. and P.P.K. designed and performed molecular experiments; E.V. and E.S. provided and characterized DNA samples; and W.A.F. and K.S. wrote the manuscript.

Supplemental Data

Supplemental material for this article can be found at *http://doi.org/10.1016/j.jmoldx.2020.06.014*.

References

1. Carl B, Kroll H, Bux J, Bein G, Santoso S: B-lymphoblastoid cell lines as a source of reference DNA for human platelet and neutrophil antigen genotyping. Transfusion 2000, 40:62−68
2. Flegel WA, Chiosea I, Sachs UJ, Bein G: External quality assessment in molecular immunohematology: the INSTAND proficiency test program. Transfusion 2013, 53:2850−2858
3. Volkova E, Sippert E, Liu M, Mercado T, Denomme GA, Illoh O, Liu Z, Rios M: Validated reference panel from renewable source of genomic DNA available for standardization of blood group genotyping. J Mol Diagn 2019, 21:525−537
4. Boyle J, Thorpe SJ, Hawkins JR, Lockie C, Fox B, Matejtschuk P, Halls C, Metcalfe P, Rigsby P, Armstrong-Fisher S, Varzi AM, Urbaniak S, Daniels G: International reference reagents to standardise blood group genotyping: evaluation of candidate preparations in an international collaborative study. Vox Sang 2013, 104:144−152
5. Kroll H, Carl B, Santoso S, Bux J, Bein G: Workshop report on the genotyping of blood cell alloantigens. Transfus Med 2001, 11:211−219
6. Ugozzoli L, Wallace RB: Application of an allele-specific polymerase chain reaction to the direct determination of ABO blood group genotypes. Genomics 1992, 12:670−674
7. St-Louis M: Molecular blood grouping of donors. Transfus Apher Sci 2014, 50:175−182

8. Stabentheiner S, Danzer M, Niklas N, Atzmuller S, Proll J, Hackl C, Polin H, Hofer K, Gabriel C: Overcoming methodical limits of standard RHD genotyping by next-generation sequencing. Vox Sang 2011, 100:381−388
9. Orzinska A, Guz K, Brojer E: Potential of next-generation sequencing to match blood group antigens for transfusion. Int J Clin Transfus Med 2019, 7:11−22
10. Srivastava K, Lee E, Owens E, Rujirojindakul P, Flegel WA: Full-length nucleotide sequence of ERMAP alleles encoding Scianna (SC) antigens. Transfusion 2016, 56:3047−3054
11. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG: The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res 2015, 43:D423−D431
12. Srivastava K, Wollenberg KR, Flegel WA: The phylogeny of 48 alleles, experimentally verified at 21 kb, and its application to clinical allele detection. J Transl Med 2019, 17:43
13. Tournamille C, Le Van Kim C, Gane P, Cartron JP, Colin Y: Molecular basis and PCR-DNA typing of the Fya/fyb blood group polymorphism. Hum Genet 1995, 95:407−410
14. Poole J, Daniels G: Blood group antibodies and their significance in transfusion medicine. Transfus Med Rev 2007, 21:58−71
15. Meny GM: The Duffy blood group system: a review. Immunohematology 2010, 26:51−56
16. Meny GM: An update on the Duffy blood group system. Immunohematology 2019, 35:11−12
17. Yin Q, Srivastava K, Gebremedhin A, Makuria AT, Flegel WA: Long-range haplotype analysis of the malaria parasite receptor gene ACKR1 in an East-African population. Hum Genome Var 2018, 5:26
18. Fichou Y, Berlivet I, Richard G, Tournamille C, Castilho L, Férec C: Defining blood group gene reference alleles by long-read sequencing: proof of concept in the ACKR1 gene encoding the Duffy antigens. Transfus Med Hemother 2020, 47:23−32
19. Duke JL, Mosbruger TL, Ferriola D, Chitnis N, Hu T, Tairis N, Margolis DJ, Monos DS: Resolving MiSeq-generated ambiguities in HLA-DPB1 typing by using the Oxford Nanopore Technology. J Mol Diagn 2019, 21:852−861
20. WHO: Report of the international collaborative study to evaluate eighteen additional candidates for addition to the existing collection of four WHO international reference reagents for blood group genotyping. Edited by WHO Expert Committee on Biological Standardization. Geneva: World Health Organization, 2019
21. Munnink BBO, Nieuwenhuijse DF, Stein M, O'Toole Á, Haverkate M, Mollers M, Kamga SK, Schapendonk C, Pronk M, Lexmond P, van der Linden A, Bestebroer T, Chestakova I, Overmars RJ, van Nieuwkoop S, Molenkamp R, van der Eijk AA, GeurtsvanKessel C, Vennema H, Meijer A, Rambaut A, van Dissel J, Sikkema RS, Timen A, Koopmans M: Dutch-Covid-19 response team: rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. Nat Med 2020, 26:1405−1410
22. Li HW: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. [ePub] arXiv 2013:1303.3997v1301.
23. Thorvaldsdottir H, Robinson JT, Mesirov JP: Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013, 14:178−192
24. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR: MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 2010, 34:816−834
25. Sachs L: Angewandte Statistik - Anwendung statistischer Methoden. ed 7. Berlin, Springer-Verlag, 1992. pp. 446−447
26. Magi A, Giusti B, Tattini L: Characterization of MinION nanopore data for resequencing analyses. Brief Bioinform 2017, 18:940−953
27. Tedersoo L, Drenkhan R, Anslan S, Morales-Rodriguez C, Cleary M: High-throughput identification and diagnostics of pathogens and pests: overview and practical recommendations. Mol Ecol Resour 2019, 19:47−76
28. Fox EJ, Reid-Bayliss KS, Emond MJ, Loeb LA: Accuracy of next generation sequencing platforms. Next Gener Seq Appl 2014, 1:1000106
29. Eckert KA, Kunkel TA: DNA polymerase fidelity and the polymerase chain reaction. PCR Methods Appl 1991, 1:17−24
30. Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA: Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics 2012, 13:375
31. Donadieu J, Fenneteau O, Beaupain B, Mahlaoui N, Chantelot CB: Congenital neutropenia: diagnosis, molecular bases and patient management. Orphanet J Rare Dis 2011, 6:26
32. Edderkaoui B, Baylink DJ, Beamer WG, Wergedal JE, Porte R, Chaudhuri A, Mohan S: Identification of mouse Duffy antigen receptor for chemokines (Darc) as a BMD QTL gene. Genome Res 2007, 17:577−585
33. Massara M, Bonavita O, Mantovani A, Locati M, Bonecchi R: Atypical chemokine receptors in cancer: friends or foes? J Leukoc Biol 2016, 99:927−933
34. Peiper SC, Wang ZX, Neote K, Martin AW, Showell HJ, Conklyn MJ, Ogborne K, Hadley TJ, Lu ZH, Hesselgesser J, Horuk R: The Duffy antigen/receptor for chemokines (DARC) is expressed in endothelial cells of Duffy negative individuals who lack the erythrocyte receptor. J Exp Med 1995, 181:1311−1317
35. Le Pennec PY, Rouger P, Klein MT, Robert N, Salmon C: Study of anti-Fya in five black Fy(a-b-) patients. Vox Sang 1987, 52:246−249
36. Tournamille C, Colin Y, Cartron JP, Le Van Kim C: Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. Nat Genet 1995, 10:224−228
37. Zimmerman PA, Woolley I, Masinde GL, Miller SM, McNamara DT, Hazlett F, Mgone CS, Alpers MP, Genton B, Boatin BA, Kazura JW: Emergence of FY*A(null) in a Plasmodium vivax-endemic region of Papua New Guinea. Proc Natl Acad Sci U S A 1999, 96:13973−13977
38. Kempinska-Podhorodecka A, Knap O, Drozd A, Kaczmarczyk M, Parafiniuk M, Parczewski M, Ciechanowicz A: Analysis for genotyping Duffy blood group in inhabitants of Sudan, the fourth cataract of the Nile. Malar J 2012, 11:115
39. Langhi DM, Albuquerque SR, Covas DT, Perez CA, Bordin JO: The presence of FYA^{null} allele of Duffy blood group system in blood donors and individuals from a malarial endemic region of Brazil [abstract]. Blood 2004, 104:741a
40. Pisacka M, Marinov I, Kralova M, Kralova J, Koranova M, Bohonek M, Sood C, Ochoa-Garay G: FY*A silencing by the GATA-motif variant FY*A(-69C) in a Caucasian family. Transfusion 2015, 55:2616−2619
41. Duke JL, Lind C, Mackiewicz K, Ferriola D, Papazoglou A, Derbeneva O, Wallace D, Monos DS: Towards allele-level human leucocyte antigens genotyping - assessing two next-generation sequencing platforms: Ion Torrent Personal Genome Machine and Illumina MiSeq. Int J Immunogenet 2015, 42:346−358
42. Lin S, Chakravarti A, Cutler DJ: Haplotype and missing data inference in nuclear families. Genome Res 2004, 14:1624−1632
43. Browning SR, Browning BL: Haplotype phasing: existing methods and new developments. Nat Rev Genet 2011, 12:703−714
44. Belsare S, Levy-Sakin M, Mostovoy Y, Durinck S, Chaudhuri S, Xiao M, Peterson AS, Kwok PY, Seshagiri S, Wall JD: Evaluating the quality of the 1000 genomes project data. BMC Genomics 2019, 20:620