Defining Blood Group Gene Reference Alleles by Long-Read Sequencing: Proof of Concept in the ACKR1 Gene Encoding the Duffy Antigens

Yann Fichou\textsuperscript{a, b}, Isabelle Berlivet\textsuperscript{a}, Gaëlle Richard\textsuperscript{a}, Christophe Tournamille\textsuperscript{b, c}, Lilian Castilho\textsuperscript{d}, Claude Férec\textsuperscript{a, e}

\textsuperscript{a}EFS, Inserm, Univ Brest, UMR 1078, GGB, Brest, France; \textsuperscript{b}Laboratoire d’Excellence GR-Ex, Paris, France; \textsuperscript{c}IMRB-Inserm U955 Equipe 2 Transfusion et Maladies du Globule Rouge, EFS Ile-de-France, Créteil, France; \textsuperscript{d}Hemocentro Campinas-unicamp, Campinas, Brazil; \textsuperscript{e}Laboratoire de Génétique Moléculaire et d’Histocompatibilité, CHU Morvan, Brest, France

Keywords
ACKR1 · Alleles · Blood group gene · Duffy · Haplotype · Long-read sequencing

Abstract

Background: In the novel era of blood group genomics, (re-)defining reference gene/allele sequences of blood group genes has become an important goal to achieve, both for diagnostic and research purposes. As novel potent sequencing technologies are available, we thought to investigate the variability encountered in the three most common alleles of ACKR1, the gene encoding the clinically relevant Duffy antigens, at the haplotype level by a long-read sequencing approach. Materials and Methods: After long-range PCR amplification spanning the whole ACKR1 gene locus (∼2.5 kilobases), amplicons generated from 81 samples with known genotypes were sequenced in a single read by using the Pacific Biosciences (PacBio) single molecule, real-time (SMRT) sequencing technology. Results: High-quality sequencing reads were obtained for the 162 alleles (accuracy > 0.999). Twenty-two nucleotide variations reported in databases were identified, defining 19 haplotypes: four, eight, and seven haplotypes in 46 ACKR1*01, 63 ACKR1*02, and 53 ACKR1*02N.01 alleles, respectively. Discussion: Overall, we have defined a subset of reference alleles by third-generation (long-read) sequencing. This technology, which provides a “longitudinal” overview of the loci of interest (several thousand base pairs) and is complementary to the second-generation (short-read) next-generation sequencing technology, is of critical interest for resolving novel, rare, and null alleles.

Introduction

For years, blood group genotyping has become routine practice in most immunohematology laboratories to circumvent and/or complement the current limitations of serological analysis [1]. To this aim, access to accurate, annotated reference gene/allele sequences is definitely critical to molecular biologists in order to engineer robust and “universal” tests for targeting and genotyping regions of interest and to prevent inappropriate design leading to mistyping. In the human GRCh38.p13/hg38 reference genome, for example, the reference sequence of the RHD gene (assembly GRCh38.p13; chromosome 1, NC_000001.11(25272393..25330445)), which is involved in the expression of the D antigen in the polymorphic, clinically relevant Rh blood group system [2], is currently a rare variant allele, i.e., RHD*DAU0 (or RHD*10.00), mostly found in populations of African origin [3]. While exonic sequences are well defined, variations in introns are barely studied. Genomic variability in these latter regions, which is poorly reported in spite of the overgrowing
number of available datasets due to next-generation sequencing (NGS) projects (www.ncbi.nlm.nih.gov/pubmed/?term=ngs), is globally unknown and may result in improper positioning of amplification primers with a subsequent potential “mistyping” risk that is not assessable. Therefore, (re)defining reference gene/allele sequences appears to be an important goal to achieve, notably in the field of molecular immunohematology.

Recently, Tounsi and colleagues [4] investigated the whole RHD gene locus by an NGS technology-based approach in a total of 69 blood donors with either D-positive or weak D phenotype, as defined by routine serological analysis. The authors took advantage of the various zygosities of the gene (i.e., n = 0, 1, or 2 RHD copies) to generate data mostly in hemizygous samples (n = 1 copy). Short-read libraries generated by fragmenting pools of six overlapping long-range PCR (LR-PCR) amplicons spanning the whole gene were sequenced by using the Ion PGM Sequencing technology (Life Technologies). As expected, high-quality data were generated, reported and unknown variants in exons and intronic regions were identified. Reference RHD allele sequences associated with R1 (DCE), R0 (Dce), and R2 (DCE) haplotypes on the one hand, and R3 (DcE) haplotype on the other hand, were finally defined. Overall, this work is definitely successful and valuable to the scientific community.

However, in those conditions, two main issues remain for conducting similar studies in other genes: gene zygosity and short-read sequencing. Indeed, contrary to RHD, which may be found at a single copy number (= hemizygous state) in the human genome, blood group genes are mostly located on autosomes and thus found with two copies in nonpathological conditions. This implies that phasing of heterozygous markers is impaired, unless a time-consuming and technically demanding subcloning strategy is used. Although this drawback can be overcome theoretically by the NGS approach, short-read sequencing also limits the extent to which heterozygous markers may be phased to 150–500 base pairs (bp) typically. Use of computational approach may definitely help [5], but haplotype is basically predicted and thus remains unsure.

Pioneer works in the field of blood group genetics reported the successful use of a long-read sequencing approach for determining unambiguously haplotypes of the ACKR1 gene [6, 7], formerly known as DARC [8]. This gene encodes the transmembrane Atypical Chemokine Receptor 1 protein, which carries the clinically relevant red blood cell antigens in the Duffy blood group system first reported in 1950 [9–11]. The Duffy system is currently defined by five antigens: the polymorphic Fya and Fyb (or Fy1 and Fy2, respectively) antigens, as well as the high-prevalence Fy3, Fy5, and Fy6 antigens (ISBT website: http://www.isbtweb.org/fileadmin/user_upload/Working_parties/WP_on_Red_Cell_Immunogenetics_and/008_FY_Alleles_v4.1.pdf). In terms of molecular genetics, ACKR1 (GRCh37.p13/hg19, NC_000001.10(159173803.159176290)) is a small gene in size (~2.5 kb) composed of two exons (GenBank accession number: NM_002036.3). Exon 1 consists of a ~1-kb 5′-UTR and a short 21-bp coding sequence, while exon 2 consists of a ~1-kb coding sequence followed by a short 50–bp 3′-UTR. Exons 1 and 2 are separated by a single 480-bp intron. The codominant, antithetical Fya and Fyb antigens are encoded, respectively, by the ACKR1*01 and ACKR1*02 (or *01 and *02) alleles (formerly Fy*A and Fy*B), which differ by a single c.125G>A nucleotide change in their coding sequence in exon 2, resulting in a glycine-to-aspartate substitution at position 42 of the protein (p.Gly42Asp, or p.G42D) [12, 13]. The ACKR1 protein is well known to be a receptor for infection by the devastating malaria parasite Plasmodium vivax at the surface of red blood cells [14]. As a result, the ACKR1*02N.01 (or *02N.01) null allele, which is characterized by the c.–67T>C substitution in the 5′-UTR of the gene abolishing a GATA1 erythroid transcription factor binding site [15], is found at high frequency (up to 100%) in sub-Saharan African populations, thus conferring a selective advantage against P. vivax infection [16].

In the previous works carried out by Flegel and collaborators, the authors resolved ACKR1 haplotypes after extensive sequencing of LR-PCR products encompassing the whole gene locus by the Sanger method [6, 7]. In 54 African American blood donors, Schmid et al. [6] reported comprehensively frequent and rare variant distribution in both coding and noncoding regions of the gene (~2.5 kb) in the 108 haplotypes successfully and unambiguously sequenced, accounting for a total of eleven distinct alleles. Later on, Yin et al. [7] used a similar approach supported by computational phasing to resolve haplotypes in a cohort of 57 Ethiopian donors from a malaria-endemic area, as well as in three healthy donors. In this second report, a larger region (~5.2 kb) was successfully sequenced from the 60 individuals, yielding a total of 18 distinct alleles. Those reports are definitely critical and provide comprehensive and accurate datasets for future studies [7]. However, although a long-read sequencing strategy was used by the authors, the conventional Sanger approach is extremely labor-intensive, but also use of computational phasing only allows prediction of haplotypes, which may be erroneous [7].

The Pacific Biosciences (PacBio) single molecule, real-time (SMRT) sequencing technology [17] has been used successfully for accurate sequencing of both targeted and shotgun, long libraries ranging from >500 bp up to 30 kilobases (kb) in very various fields of research, but also for human diagnostic purposes [18]. We then thought to take advantage of SMRT sequencing for conducting a study comparable to those carried out by Dr. Flegel’s group, and
similarly studied the genomic sequences of the three most common ACKR1 alleles, namely *01, *02, and *02N.01 alleles. To this aim, we amplified an LR-PCR product encompassing the whole ACKR1 gene locus in selected samples with known genotypes and sequenced the products by the SMRT sequencing technology. Here we report the genomic variability in the subset of heterogeneous samples and highlight the current advantages of the approach for conducting such investigations.

**Materials and Methods**

**Samples and LR-PCR Amplification**

Genomic DNA (gDNA) samples for external quality assessment (EQA) (Transfusion Medicine No. 235 – Survey Immunohaematology 06, Instand e.V., Düsseldorf, Germany) were selected for the study (EUR subset). In parallel and in order to have access to the ACKR1*02N.01 allele in individuals of African ancestry, blood samples from unrelated donors and patients with sickle cell disease of Brazilian origin were collected after informed consent, and genomic DNA was extracted by routine techniques (BRA subset).

ACKR1 genotypes at positions of interest (NM_002036.3: c.–67, c.125, and c.265) were investigated by the BioArray HEA (human erythrocyte antigen) BeadChip Kit (Immucor, Paris, France). Genotypes were confirmed by the manufacturer in EQA samples. Although the precise ethnic origin of the EQA samples is unknown, only two samples in the EUR subset carry the ACKR1*02N.01 allele, suggesting that this subset is mostly of non-African origin.

PCR primers for LR-PCR amplification of the whole ACKR1 locus (Fig. 1a), as well as proximal 5′- and 3′-flanking regions, were designed with PrimerQuest Tool (www.idtdna.com/PrimerQuest/) by taking into account the genomic variability in reference database (variant allele frequency < 10⁻²), as reported in the Genome Aggregation Database (gnomAD v2.1.1 browser) [19], which can potentially result in allele dropout. LR-PCR amplification was carried out in 1X PrimerSTAR GXL Buffer (TaKaRa, purchased from Ozyme, Saint-Cyr-l’École, France) with 0.2 µM forward (ACKR1_e1-2F: 5′-ACAAATGAAACAGCGTCCCC-3′) and reverse (ACKR1_e1-2R: 5′-TCTCCCTTGAGATGGCCGTG-3′) primers, 1.25 U PrimeSTAR GXL DNA Polymerase (TaKaRa), and ∼200 ng gDNA as a template in a 50-µL final volume to generate a ∼3-kb product. PCR conditions were 30 cycles at 98 °C for 10 s and 68 °C for 3 min. PCR products were controlled on a 1% agarose gel stained with ethidium bromide for visualization of LR-PCR amplification.

**Library Preparation and SMRT Sequencing**

LR-PCR reaction products were purified with 0.6× AMPure PB Beads (Pacific Biosciences, Menlo Park, CA, USA) in conditions recommended by the manufacturer and eluted in 20 µL Elution Buffer (EB, PacBio) prior to DNA quantitation with the Qubit ds-DNA HS (High Sensitivity) Assay Kit (Thermo Fisher Scientific, Montigny-le-Bretonneux, France). All other reagents are from PacBio, unless indicated. Libraries were prepared by following the Preparing Amplicon Libraries using PacBio Barcoded Adapters for Multiplex SMRT sequencing procedure. Briefly, end-repair and ligation reaction for multiplexing was carried out with equimolar ratios of input amplicons with the SMRTbell Barcoded Adapter Prep Kit and Barcoded Adapter Plate-96. Products were pooled together, and purification was carried out as described above. DNA damage repair and exonuclease digestion were successively carried out by using 500 ng DNA with the SMRTbell DNA Damage Repair Kit before two successive rounds of purification and then quantitation of SMRTbell Templates in conditions described above.
Final libraries (EUR1 and EUR2), consisting of the circular templates, polymerase and Sequencing Primer v3 or v4, were prepared with the Sequel Binding Kit 2.0 or v3, loaded onto SMRT cells 1M v2 or v3 LR at a 4 pM final concentration by diffusion, and sequenced with the Sequel Sequencing Kit 2.0 or 3.0 in the Sequel System, respectively. Sequencing data were collected from either 10-hour or 20-hour movie time. Alternative versions of the kits, reagents, and movie time were used for the BRA library: Sequencing Primer v4, Sequel Binding Kit 3.0, SMRT cells 1M v3 LR, Sequel Sequencing Kit 3.0, and 20-hour movie time.

Data Quality Control and Analysis
Primary analysis of output data was carried out with SMRT-Link v5.1.0 (EUR1 and EUR2) or v6.0.0 (BRA) software (PacBio). Briefly, barcode demultiplexing was carried out automatically at the end of the runs. Data were analyzed by the Long Amplicon Analysis (LAA) module, which is dedicated to phasing information by generating long haploid reads [20], with a minimum predicted accuracy of 0.999 (LAA parameters: minimum read length = 2,500 bp; maximum read length = 3,200 bp; maximum allele/haplotype read fraction = 0.2; maximum phasing reads = 500). Fastq files were exported to SeqOne (Montpellier, France) for processing. Briefly, fastq files were aligned on reference genome GRCh37/hg19 with BWA-MEM [21], variant calling was carried out with freebayes software (setting: –C 1) [22]. Variants were normalized with vt [23], phased with WhasTeP [24], filtered on ACKR1 locus with BCFtools [25], and annotated with the Ensemble Variant Effect Predictor (VEP) [26]. Output data were finally compiled and individual haplotypes were generated with in-house software module.

Results

LR-PCR Amplification and Sample Genotype
Forty-four and 53 samples in the EUR and BRA subsets, respectively, were selected for ACKR1 LR-PCR amplification on the basis of previous genotyping data: *01 and/or *02 carriers for EUR subset, and *02N.01 carriers for BRA subset. LR-PCR amplification was successfully carried out in 38/44 samples of European origin, as well as in 43/53 samples (26 donors and 17 patients) of Brazilian origins, as revealed by a major ~3-kb PCR amplification product visualized on a 1% agarose gel (Fig. 1b).

Those 81 samples, organized in three different subsets (EUR1, n = 10; EUR2, 28; BRA, 43) were selected for subsequent SMRT sequencing. All samples were previously genotyped by a routine molecular approach targeting nucleotide positions c.–67, c.125, and c.265, which are functionally relevant. A total of 46 ACKR1*01, 63 ACKR1*02, and 53 ACKR1*02N.01 alleles were found by genotyping, while nucleotide variation c.265C>T (ACKR1*02W.01 allele, formerly known as FYwX), which results in a weak Fyβ antigen expression [27, 28], was not identified in the subsets of samples (Table 1).

SMRT Sequencing of the ACKR1 Gene
After library preparation of the 81 LR-PCR amplification products, SMRT sequencing was carried out in three separate runs: 10 samples in run 1 (EUR1), 28 samples in run 2 (EUR2), and 43 samples in run 3 (BRA) (online suppl. Table S1; for all online suppl. material, see www.karger.com/doi/10.1159/000504584; mean estimated accuracy = 99.9964%). For reporting the data, we chose to restrict our study to the 2,488 bp defining ACKR1 reference sequence in GRCh37.p13/hg19 (i.e., NC000001.10 (159173803..159176290)).

High-quality haplotypes were obtained successfully with no ambiguity for the 162 alleles (online suppl. Table S2). Genotypes at positions of interest (i.e., c.–67, c.125, and c.265) were in full concordance with those obtained by routine genotyping procedure (data not shown). Also, in the BRA subset, phenotypes predicted on the basis of SMRT sequencing data correlated perfectly with donor and patient phenotypes (online suppl. Table S2). Twenty-two single nucleotide variations (SNVs), all of them referenced in the gnomAD v2.1.1 database [20], were identified in a total of 19 different haplotypes unofficially annotated (Table 1): 46 ACKR1*01, 63 ACKR1*02, and 53 ACKR1*02N.01-based alleles defining four, eight,
and seven haplotypes, respectively. Of the 22 SNVs, 13 are located within the 5'-UTR (including the functional c.–67T>C variation) and five within the single intron of the gene. The remaining four variants are in exon 2: the polymorphic c.125G>A SNV, which determines Fya/Fyb antigen expression; the missense c.298G>A variant, which has been reported on both *01 and *02 allele backgrounds and has long been known to be nonfunctional [27]; the synonymous c.714G>A variant; and the missense c.796G>A (p.Val266Met) variant carried by a *02N.01 allele. Interestingly and although the trend has to be confirmed in a larger study, if excluding the three common c.–67T>C, c.125G>A, and c.298G>A SNVs, 16/19 SNVs were found only on a single allelic background: one with *01, eight with *02, and seven with *02N.01 (Table 2). This result suggests that linkage disequilibrium may be used directly for haplotyping in future studies.

Although it cannot be concluded unambiguously, because the regions sequenced in the respective studies are not exactly the same, as many as eight ACKR1 haplotypes reported in our study, including *01_01, *01_02, *02_01, *02_02, *02_03, *02_05, *02N.01_01, and *02N.01_03, may have been identified before (online suppl. Table S3)

Table 2. ACKR1 variants and haplotypes identified by long-read sequencing

<table>
<thead>
<tr>
<th>Genomic coordinates</th>
<th>Nucleotide change</th>
<th>Consequence</th>
<th>dbSNP ID</th>
<th>Global MAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>159204017</td>
<td>c.–945delC</td>
<td>5'-UTR</td>
<td>rs1164458</td>
</tr>
<tr>
<td>B</td>
<td>159204907</td>
<td>c.–863A-G</td>
<td>5'-UTR</td>
<td>rs3027009</td>
</tr>
<tr>
<td>C</td>
<td>159204113</td>
<td>c.–847C&gt;T</td>
<td>5'-UTR</td>
<td>rs14349581</td>
</tr>
<tr>
<td>D</td>
<td>159204287</td>
<td>c.–735G&gt;A</td>
<td>5'-UTR</td>
<td>rs41246467</td>
</tr>
<tr>
<td>E</td>
<td>159204305</td>
<td>c.–655A-G</td>
<td>5'-UTR</td>
<td>rs3027011</td>
</tr>
<tr>
<td>F</td>
<td>159204326</td>
<td>c.–643C&gt;T</td>
<td>5'-UTR</td>
<td>rs76624005</td>
</tr>
<tr>
<td>G</td>
<td>159204333</td>
<td>c.–627C&gt;T</td>
<td>5'-UTR</td>
<td>rs3027012</td>
</tr>
<tr>
<td>H</td>
<td>159204403</td>
<td>c.–557G&gt;C</td>
<td>5'-UTR</td>
<td>rs90558956</td>
</tr>
<tr>
<td>I</td>
<td>159204419</td>
<td>c.–541C&gt;T</td>
<td>5'-UTR</td>
<td>rs3027013</td>
</tr>
<tr>
<td>J</td>
<td>159204561</td>
<td>c.–399C–398</td>
<td>delCT</td>
<td>rs17873098</td>
</tr>
<tr>
<td>K</td>
<td>159204676</td>
<td>c.–284C&gt;T</td>
<td>5'-UTR</td>
<td>rs89350800</td>
</tr>
<tr>
<td>L</td>
<td>159204690</td>
<td>c.–270C&gt;T</td>
<td>5'-UTR</td>
<td>rs37450877</td>
</tr>
<tr>
<td>M</td>
<td>159204895</td>
<td>c.–267C&gt;T</td>
<td>5'-UTR</td>
<td>rs2614778</td>
</tr>
<tr>
<td>N</td>
<td>159205095</td>
<td>c.–251±51T&gt;C</td>
<td>intron</td>
<td>rs2750107</td>
</tr>
<tr>
<td>O</td>
<td>159205130</td>
<td>c.–211±51T&gt;C</td>
<td>intron</td>
<td>rs863002</td>
</tr>
<tr>
<td>P</td>
<td>159205179</td>
<td>c.–199±51T&gt;C</td>
<td>intron</td>
<td>rs36784395</td>
</tr>
<tr>
<td>Q</td>
<td>159205218</td>
<td>c.–22±51T&gt;C</td>
<td>intron</td>
<td>rs27673742</td>
</tr>
<tr>
<td>R</td>
<td>159205403</td>
<td>c.–11±51T&gt;C</td>
<td>intron</td>
<td>rs3027016</td>
</tr>
<tr>
<td>S</td>
<td>159205546</td>
<td>c.125G&gt;A</td>
<td>p.Gly42Asp</td>
<td>rs12075</td>
</tr>
<tr>
<td>T</td>
<td>159205737</td>
<td>c.129G&gt;A</td>
<td>p.Ala43Thr</td>
<td>rs39692</td>
</tr>
<tr>
<td>U</td>
<td>159205737</td>
<td>c.129G&gt;A</td>
<td>p.Ala43Thr</td>
<td>rs39692</td>
</tr>
<tr>
<td>V</td>
<td>159206235</td>
<td>c.129G&gt;A</td>
<td>p.Ala43Thr</td>
<td>rs39692</td>
</tr>
</tbody>
</table>

Respective major genotypes are indicated by a “–” in the table. a Genomic coordinates in the human GRC38.p13/hg38 reference genome. b In accordance with GenBank accession numbers NM_002036.3 and NP_002037.2. c NCBI database of genetic variation (URL: www.ncbi.nlm.nih.gov/snp). d Minor allele frequency in the general populations (gnomAD v2.1.1 [20]). e Occurrence (EUR + BRA; total = 162). N/A, not available.
Importantly, the reference ACKR1 allele in the human GRCh38/hg38 reference genome was found to be the most common *01 allele (*01_01 in Table 2: 27/46, 58.7%) in both subsets. Also, while a single haplotype accounts for 75.5% (40/53) in the *02N.01 subgroup, three alleles have a frequency >20% in the *02 subgroup. The highest variability observed in the latter subgroup is definitely in favor of the hypothesis that ACKR1*02 is the ancestral allele already supported by previous studies [15, 29, 30].

Discussion

So far, molecular studies have been carried out by various methods in the field of blood group genetics. The identification of the molecular bases of blood group antigen expression has allowed targeted SNV genotyping in numerous genes by standard approaches, such as PCR-restriction fragment length polymorphism (PCR-RFLP) [12, 13, 28, 31], sequence (allele)-specific primer PCR (SSP/ASP-PCR) [27, 32, 33], (high-resolution) melting-curve analysis (HRM) [34, 35], and many other medium- and high-throughput technologies [36], supported by the input of conventional Sanger sequencing for the direct identification of rare SNVs, illustrated by hundreds of reports in the literature. For the past decade, NGS of short reads, namely second-generation sequencing, has contributed to extend dramatically the catalogue of molecular variations in the human genome, including blood group genes, by targeted exome, whole exome, and even whole genome strategies in various populations [37–49]. Therefore, data mining and computational biology have boosted the characterization of novel blood group systems and the identification of the molecular basis of blood antigen expression [50–55]. The data generated by these investigations are definitely valuable for molecular biologists/geneticists both in diagnostics and research. However, at this point, a current limitation is that interpretation of those data in subsets of samples is mostly “transversal,” i.e., each variation may be considered generally on an individual basis, independently of others located several hundred base pairs away.

In order to get insights into the “longitudinal” interpretation of variants, we thought to take advantage of the long-read sequencing technology, namely third-generation sequencing, to investigate and characterize the most common haplotypes in the three main alleles (*01, *02, and *02N.01) of the gene encoding the Duffy blood group antigens, i.e., ACKR1. As indicated above, such a strategy using long-range haplotype analysis has been previously used successfully [6, 7]. In those reports, sequencing data were generated by the conventional Sanger method, and haplotypes were reconstructed by either haplotype-specific sequence analysis, which is labor-intensive and time-consuming, or computational phasing, which is a predictive approach. The SMRT sequencing technology allows direct generation of the haplotype in a single read, which is a significant advent for conducting such a study.

It is important to mention that, as we intended to use third-generation sequencing for the first time in this field, our aim was to generate a sufficient amount of genetic data in the three alleles of interest. Therefore, the selection of samples was evidently biased by a preliminary genotyping procedure to identify the corresponding samples of interest. Also, as indicated before, the precise origin of the samples in the EUR subset is unknown. So, contrary to the previous studies [6, 7], which report the analysis of individuals selected on the basis of both ethnic and geographic criteria, our work does not provide data in terms of epidemiological genetics but only depicts genetic variability in heterogeneous samples.

In the present study, we report the whole ACKR1 gene sequence in 81 samples (i.e., 162 haplotypes) obtained directly in a single read with high accuracy. It is worth mentioning that LR-PCR amplification failed in as many as 16/97 samples in our conditions, while actually amplification of a 3-kb product does not appear to be very difficult in such a genomic region with low complexity. This result evidently challenges the strategy that has been used and raises questions about the failure rate. Although we have not addressed those questions directly, at least two factors affecting LR-PCR amplification, additionally to allele dropout, which is a common feature of all PCR-based methods, may be suggested: genomic DNA quality and DNA polymerase. First, specific care was taken for designing primers to prevent from allele dropout. To this aim, the target regions selected for primer annealing actually contain polymorphic positions (online suppl. Table S4: two and five positions in the forward and reverse primers, respectively) but with a minor allele frequency <6.0 × 10⁻³ at most. Therefore, on the basis of those frequencies, it is unlikely that all 32 alleles exhibit one or more variation(s), and finally allele dropout may account only very partially for the 16.5% failure rate observed in our work. Second, the TaKaRa PrimeSTAR GXL DNA Polymerase, which has shown its potency in previous works for generating long amplicons [56], was recommended directly by PacBio (personal communication). In our hands, it has been used successfully also to generate amplicons of up to 25 kb in length (personal unpublished data). So, the DNA polymerase used in this work appears to be perfectly suitable to generate such a “short” 3-kb fragment. Third, genomic DNA samples used as templates in this work were not extracted initially for LR-PCR amplification.
Several factors, such as blood time storage, extraction methods (column, bead, precipitation), and DNA storage (temperature, freeze-thaw cycles, etc.), affect DNA quality [57]. As those parameters were not controlled through the whole procedure, i.e., from blood sampling to library production, it is tempting to suggest that DNA quality in our samples was probably not optimal. Subsequently, this low-quality (degraded) DNA may have impaired totally the efficiency of LR-PCR amplification, leading to failure in 16/97 samples. Very importantly in the field of long-read sequencing, this result actually stresses the crucial importance of optimizing and controlling the whole pre-analytical procedure for preserving genomic DNA integrity in order to maximize the chance of generating high-quality data.

Although consensus haplotypes were generated by using the LAA module from PacBio, which has been the recommended tool for phasing [20], LAA was challenged recently. In a report presenting SMRT sequencing data of long-range amplicons, the authors identify what they qualify as "unexpected numbers of amplicon sequences with substantial inaccuracies in the consensus sequences" [58, 59]. They developed another cluster and assembly pipeline based on circular consensus sequences (CCS) and showed more accurate sequence output [59]. In our study, as no confirmation of the variations by Sanger sequencing or another genotyping approach was carried out, accuracy of the final haplotypes is thus to be discussed. First, mean estimated accuracy was 99.9964%, suggesting very high-quality data. Second, there are several differences between those reports challenging LAA and ours. In addition to the longer (∼5-kb) amplicons (vs. 3-kb in our work) studied in their work, Francis and colleagues [58, 59] used previous versions of the sequencing instrument, SMRT Link software, chemistry, and movie time (RS II, v2.3, P6/C4, and 6 h, respectively, vs. Sequel, v5.1 or v6.0, v2.1 or v3, and 10 to 20 h). It is tempting to suggest that those different settings contribute definitely to generating lower-quality data. Third, genotypes at positions of interest (i.e., c.–67, c.125, and c.265) were fully concordant with previous genotyping data, which validate the data to some extent. Fourth, all variations identified in our study were previously reported and annotated. Overall, there are several technical, methodological and experimental data suggesting full consistency of our final data (haplotypes). However, it is worth recommending either (1) LAA-analyzed, third-generation sequencing data complemented with second-generation sequencing data, when possible [60], or (2) optimization of CCS reads to generate long high-fidelity reads [59], for phasing highly accurate long reads in future studies, especially as the size of the targets increase [61].

Twenty-two SNVs defining 19 different haplotypes were characterized. While the boundaries of the reported haplotypes are not exactly the same in the previous studies carried out by Dr. Plegel’s group [6, 7] and ours, some of them display a number of similarities (online suppl. Table S3), suggesting that both methodologies and approaches yield comparable results. Replication studies, in well-defined populations (in terms of ethnicity, geographical origin, and phenotype) of other origins as carried out by this group, are to be done definitely. Indeed, as variant frequency is shown to differ between populations typically, little is known about the variability of “real” haplotype, which is generally “predicted.” One may hypothesize that analysis of “real” haplotypes may be very valuable to population genetics studies.

Although expected, our data also show experimentally, as previously reported [6, 7], that variations are not distributed “randomly.” In other words, some specific variations have not been found to be cis-associated, i.e., on the same alleles, in our experiment. Typically, c.21+150C was not found to be linked to c.21+115T or c.22–58G on the same *02 allele (Table 2). It is worth mentioning that this result may be achieved theoretically by short-read sequencing, which also allows phasing of physically close markers. However, other direct phasings are possible by using long-read sequencing only. For example, the missense c.796G>A variation was identified at the heterozygous state in a sample genotyped as *02/*02N.01. Direct phasing linked the SNV to the latter allele, suggesting that a “normal” Fyb antigen may be expressed by the *02 allele.

While SNVs, which are the most common variants in the human genome, are very well detected by second-generation sequencing approaches, identification of structural variants (SVs; >50 bp typically), which include the largest set of base pairs affected, remains challenging [62]. SVs, such as hybrid genes, duplications, and deletions, are commonly encountered in RHD/RHCE and GYPA/GYPB, the genes encoding the antigens in the Rh and MNS blood group systems, respectively (ISBT Red Cell Immunogenetics and Blood Group Terminology Working Party: http://www.isbtweb.org/working-parties/red-cell-immunogenetics-and-blood-group-terminology/). An interesting example is illustrated by the recent characterization of a complex haplotype, namely DUP4, involving both the GYPB and GYPB genes [63]. This complex allele, which is protective against severe malaria, was resolved by an extensive quantitative analysis of genome sequence data. By considering the phasing capacity of the long-read sequencing method [18], it is tempting to speculate that such an approach may have facilitated the resolution and characterization of the physical structure of the DUP4 haplotype. This report also suggests that other potential complex SV alleles/haplotypes remain to be discovered. There is no doubt that recent technologies, such as third-
generation sequencing, may play a critical role for identifying those alleles by providing valuable “longitudinal” sequence information in this novel blood group genomics era [64]. Overall, we have defined a subset of reference alleles/haplotypes in the ACKR1 gene by a long-read sequencing approach. We do think that, thanks to its efficiency for phasing markers separated by several hundred, and even thousand, base pairs with high accuracy, this technology, which is complementary of the conventional short-read sequencing approach, is to be considered as a method of choice for characterizing specific reference, novel, rare, and null alleles of interest.

Acknowledgments

Statement of Ethics

The study protocol number 3.349.490 was approved by the local institute’s committee of human research.

Disclosure Statement

The authors declare no conflict of interest.

Funding Sources

This study was supported by the Direction de la Recherche et de la Valorisation (DRV) of the Etablissement français du sang (EFS) (AAP2017-19).

References


DOI: 10.1159/000504584

Transfus Med Hemother

Fichou et al.


