ORIGINAL RESEARCH

Resolving genotype-phenotype discrepancies of the Kidd blood group using nanopore sequencing

Authors: M Gueuning[1], GA Thun[1], L Schneider[2], N Trost[2], S Sigurdardottir[2], C Engström[3], G Rizzi[3], Y Merki[2], K Neuenschwander[2], B Frey[1,2,3], MP Mattle-Greminger[1]*, S Meyer[2]*

Affiliations

[1]Department of Research and Development, Blood Transfusion Service Zurich, Swiss Red Cross, Schlieren, Switzerland

[2]Department of Molecular Diagnostics and Cytometry, Blood Transfusion Service Zurich, Swiss Red Cross, Schlieren, Switzerland

[3]Department of Immunohematology, Blood Transfusion Service Zurich, Swiss Red Cross, Schlieren, Switzerland

Correspondence

Maja P. Mattle-Greminger, Department of Research and Development, Blood Transfusion Service Zurich, Swiss Red Cross, Rütistrasse 19, 8952 Schlieren, Switzerland, Phone +41 58 272 52 22, E-mail: m.mattle@zhbsd.ch

*These authors contributed equally to this work

Short title: Nanopore sequencing of JK discrepancies

Keywords: Kidd blood group, structural variant, Oxford Nanopore sequencing, haplotype, long-read sequencing

Abstract

Due to substantial improvement in read accuracy, third-generation long-read sequencing holds great potential in blood group diagnostics, particularly in cases where traditional genotyping or sequencing techniques, primarily targeting exons, are unable to explain serologic phenotypes. In this study, we employed Oxford Nanopore sequencing to resolve all genotype-phenotype discrepancies in the Kidd blood group system (JK, *SLC14A1*) observed over seven years of routine high-throughput donor genotyping using a mass spectrometry based platform at Blood Transfusion Service Zurich. Discrepant results of standard serological typing and donor genotyping were confirmed by commercial PCR-SSP kits. To resolve discrepancies, we amplified the entire coding region of *SLC14A1* (~24 kb, exons 3 to 10) in two overlapping long-range PCRs in all samples. Amplicons were barcoded and sequenced on a MinION flow cell. Sanger sequencing and bridge-PCRs were used to confirm findings. Among 11,972 donors who had both serology and genotypic data available for the Kidd system, we identified 10 cases with unexplained conflicting results. Five were linked to known weak and null alleles caused by variants not included in the routine donor genotyping. In two cases, we identified novel null alleles on the *JK*01* (Gly40Asp; c.119G>A) and *JK*02* (Gly242Glu; c.725G>A) haplotype, respectively. Remarkably, the remaining three cases were linked to a yet unknown deletion of ~5 kb spanning over exon 9-10 of the *JK*01* allele, which other molecular methods had failed to detect. Overall, nanopore sequencing demonstrated reliable and accurate performance for detecting both single nucleotide and structural variants. It possesses the potential to become a robust tool in the molecular diagnostic portfolio, particularly for addressing challenging structural variation such as hybrid genes, deletions and duplications.

Introduction

The latest advancements in third-generation long-read sequencing (TGS) technologies offer notable advantages, including the capacity to elucidate extensive haplotypes and characterize genomic regions that pose challenges to Sanger sequencing or next-generation short-read sequencing[1]. In the particular field of transfusion medicine, TGS is a strong candidate for complementing traditional approaches used for the blood group assessment of samples with complex or discordant serological and genotypic results[2–5]. These cases include serological reactions exhibiting unexpected weak agglutination or null phenotypes, often resulting from rare or unknown genetic variation not typed in routine genotyping workflows[6].

Traditionally, Sanger sequencing has been used to resolve such elusive cases. However, it has several limitations that can impede its usefulness. One drawback is the lack of phase information, i.e. the inability to reconstruct allele haplotypes, due to overlapping signals from maternal and paternal alleles. This makes functional interpretation challenging, such as allelic assignment of silencing or weakening variants in heterozygous samples[7,8]. Additionally, in the presence of frameshift mutations or small structural variants (SVs), the overlapping signals hamper the readability of the sequences. Length restrictions of Sanger sequencing is another drawback, which limits the number of genomic regions that can realistically be targeted. Consequently, sequencing of non-coding regions of a gene, even those known to play a role in gene expression and phenotype determination (i.e. promoter regions, enhancer elements, and transcription factor binding sites) is rarely done. The limited length of the sequenced region also hinders the detection of SVs. Large deletions, for example, may stretch over several exons. Importantly, such deletions may also simply go unnoticed with Sanger sequencing in case of an unaffected second allele without the deletion in the background as this one will get sequenced (allelic dropout). Some of the limitations of Sanger sequencing could be overcome using short-read sequencing technologies. However, their intrinsic limitation, the short-read length, still hampers haplotype reconstruction and the analysis of SVs.

The long-read sequencing technology by Oxford Nanopore Technologies (ONT) represents a promising solution to tackle limitations of existing approaches. It allows sequencing long reads of blood group genes at the kilobase (kb) scale in a low throughput setting at high quality. This greatly facilitates haplotype reconstruction and the inclusion of introns and proximal regulatory elements[5,9]. Moreover, it lowers the risk of misinterpreting loss of heterozygosity caused by allelic dropout as the long reads are much more likely to cover the respective breakpoint sites. Previous constraints of ONT sequencing,

in particular the low single-read accuracy and the need for advanced bioinformatics skills [10–12], have been greatly alleviated by latest developments.

In this study, we propose a cost-effective and reliable approach to resolve the Kidd blood group (JK) in samples with discordant serological and genotypic data. The workflow is based on long-range PCR (LR-PCR) of the gene encoding the JK blood group, *SLC14A1*, followed by nanopore sequencing of the amplicons. The approach can easily be customized to target various other blood group genes as only the LR-PCR primer design is gene-specific. We used here the Kidd system as a proof-of-concept due to a combination of its clinical relevance and its low antigen diversity, enabling straightforward phenotyping. Kidd only comprises three antigens: the antithetical Jk(a) and Jk(b) antigens (also called JK1 and JK2), as well as the high incidence JK3[13]. Antibody formation by allo-immunisation is particularly common in patients with sickle cell disease [14] and Kidd antibodies are a source of delayed hemolytic transfusion reactions as well as hemolytic disease of the newborn[15]. The *SLC14A1* gene, spanning over approximately 30 kb on chromosome 18, consists of 10 exons and translates into a 389 amino acid glycoprotein. The International Society of Blood Transfusion (ISBT) currently recognizes over 60 different *JK* alleles, of which most represent null alleles caused by missense or nonsense single nucleotide variants (SNVs)[13].

Here, we nanopore-sequenced all unresolved genotype-phenotype discrepant cases collected over seven years of routine donor genotyping at Blood Transfusion Service Zurich (Switzerland) using Matrix-Assisted Laser Desorption Ionization – Time-of-Flight mass spectrometry (MALDI-TOF MS)[6,16]. Overall, our results demonstrate that ONT sequencing can effectively and accurately resolve genotype-phenotype discrepancies, even those not resolvable by Sanger sequencing. This cost-effective approach holds general promise for accurate blood group determination in challenging cases.

Material and methods

An overview of how discrepant samples were identified and processed is provided in Figure 1.

Routine high-throughput donor genotyping

DNA of blood donors was extracted automatically on a Chemagic MSM I (Chemagen, Perkin Elmer). High-throughput genotyping was performed by MALDI TOF MS on 35,954 donor samples (between 2015 and 2021) as described previously [16,17]. Genotyping included discrimination of *JK* alleles based on genetic variants of interest in a Swiss and international donor pool. *JK*01* and *JK*02* were

discriminated using the SNV c.838G>A, whereas the null alleles *JK*01N.03* (Swiss), *JK*02N.01* and *JK*02N.02* (Asian) were typed using the c.582C>G, c.342-1G>A, and c.342-1G>C variant[18], respectively. All coding DNA positions provided are based on reference transcript NM_015865.7.

Of all donors genotyped by MALDI-TOF MS, 11,972 donors were additionally serologically typed on a follow-up donation using Erytra automated system (Grifols) for seven blood groups, among which was Kidd (JK). In case of genotype-phenotype discrepancy between MALDI-TOF MS genotyping and phenotyping on the Erytra, the sample was analyzed further using other standard serological techniques. MALDI-TOF MS results of *JK*01/02* genotyping were validated using commercial PCR-SSP kits (sequence-specific-priming PCR; inno-train). DNA was manually re-extracted from the follow-up donation using the Nucleon BACC 3 kit (Gen-Probe Life Sciences). DNA concentrations were measured using Nanodrop 2000 (Thermo Fisher Scientific).

Nanopore sequencing of genotype-phenotype discrepancies

In case the genotype-phenotype discrepancy was confirmed, the entire coding region of the *SLC14A1* gene (~24 kb, exon 3 to 10) was amplified and sequenced using ONT. Amplifications were done with two LR-PCRs. The first fragment (12,981 bp) was amplified using the forward primer 5'-TGCCACTTGAGTGTTTTCATTTGATGCTGC-3' and reverse primer 5'-CACTATCCCTCCTCCTTTTTGTTCCCAAGC-3'. PCR primers for the second fragment (13,112 bp) were 5'-AAGTGACGTCCCCTCTCTGAGAGCATTAAA-3' (forward) and 5'-AACATTCTGACAAGTGGCTGGTCCTAGAGA-3' (reverse). To facilitate phasing in the downstream analyses, both LR-PCR fragments were designed to have a large overlap (~1,550 bp). PCR amplifications were done in duplicates using the PrimerSTAR GXL polymerase (Takara Bio) and 250 ng of genomic DNA per reaction, following the manufacturer's protocol. To increase amplification success, we added 1 M of Betaine enhancer (VWR) per reaction. The PCR's profile followed a 2-step approach with a 10 second denaturation step at 95°C and a 10 minute extension step at 68°C for 30 cycles. Amplification success was verified on a 0.8% agarose gel stained with GelRed Nucleic Acid Gel (Biotium). After verification, PCR replicates were pooled and purified with 1x Agencourt AMPure XP magnetic beads (Beckman Coulter). Purified PCR products were eluted in 25 µl sterile H_2O and quantified using a dsDNA broad range assay kit on a Qubit fluorometer 3.0 (Invitrogen). Sequencing libraries were constructed following ONT's 'Amplicon barcoding with Native Barcoding' protocol (version: NBA_9102_v109_revF_09Jul2020). As starting material, we pooled 50 fmol of both purified amplicons for each sample. After completion of the protocol, 44 fmol of the final barcoded library was loaded on a MinION (R9.4.1) flow cell. Sequencing was stopped after no more pores were active (~72 h).

138 **Nanopore bioinformatics processing**

139 Raw nanopore reads were demultiplexed and basecalled using ONT's *Guppy* (v.4.4.2). After
140 demultiplexing, reads in FASTQ format were filtered based on expected length and observed read
141 length distributions. More specifically, we filtered out all reads shorter than 12,500 bp and longer than
142 13,600 bp. For three samples (s02, s03 and s07) harboring a second shorter peak in read length
143 distribution (~8000 bp; Figure 2c) we additionally kept all reads having a length between 7500 bp and
144 8500 bp. After size selection, *PoreChop* (v.0.2.3) was used to trim remaining adapter and barcode
145 sequences.

146 To reduce computational burden, reads were randomly downsampled to 1000 reads per fragment
147 using *seqkit* (v.0.15). To differentiate between fragment 1 and 2, which were very similar in read
148 length, all size-selected reads were first mapped to the *SLC14A1* reference sequence (NG_011775.4)
149 using *minimap2* (v.2.17) and separated according to their mapped genomic location. For the three
150 samples showing the secondary shorter peak, mapping showed that reads responsible for this peak
151 partially corresponded to fragment 2. Therefore, we further downsampled to 500 reads for both the
152 long and short fragment 2. Finally, uniformly downsampled datasets were remapped to NG_011775.4
153 using *minimap2*, allowing the presence of supplementary alignments (i.e. split-read mapping).
154 Mapped reads were sorted, indexed and converted to BAM format using *samtools* (v.1.15).

155 Variant calling and phasing of called variants was performed using *Medaka* (v.1.2.3). As a consequence
156 of the high coverage, threshold quality for calling indels and SNVs (default: Q9 and Q8, respectively)
157 was increased to Q10. Presence of SVs was investigated with *cuteSV* (v.1.0.11). Command options
158 were set as recommended for nanopore data. *BCFtools* (v.1.9) was finally used to output phased
159 sequences in FASTA format.

160 **Variant confirmation by Sanger sequencing and bridge-PCR**

161 Nanopore sequencing results were confirmed using Sanger sequencing. Specific PCR primers for each
162 newly detected variant were designed, and PCRs were carried out using AmpliTaq DNA polymerase
163 (Life Technologies). Primers and PCR protocols are available on request. Amplicon purification and
164 Sanger sequencing were outsourced to an external company (Microsynth AG).

165 Finally, a bridge-PCR assay was designed to confirm a large deletion found with nanopore sequencing
166 in three samples. One primer pair was designed to amplify over the deletion using primers located at
167 the flanking regions, with the forward primer being located upstream of the 5' breakpoint and the
168 reverse primer downstream of the 3' breakpoint. Another reverse primer was designed to amplify the

169 wild type allele, enabling assessment of the heterozygosity status of the deletion. This reverse primer
170 was designed to bind within the region of the deletion and amplify with the upstream primer targeting
171 the 5' breakpoint. For both primer pairs, three wild type controls were used.

172 Results

173 Observed genotype-phenotype discrepancies

174 Congruence of Kidd phenotypes deduced from genotypes of seven years of high-throughput donor
175 screening and serological typing was very high. In detail, of the 35,954 donors genotyped by MALDI-
176 TOF MS , 35,937 (99.95%) passed quality control and were assigned to *JK*01* and/or *JK*02* alleles using
177 the c.838G>A SNV. Among them, 50.43% (n = 18,124) were *JK*01/02* heterozygous, 26.30% (n = 9,452)
178 *JK*01/01* and 23.27% (n = 8,361) *JK*02/02* homozygous, as expected under Hardy-Weinberg
179 equilibrium. Out of 11,972 donors with available serological data, there were 17 cases in which
180 *JK*01/02* genotyping did not agree with the observed phenotype (considering weak phenotypes as
181 normal). In seven instances, deviant results could be explained by null alleles routinely typed in the
182 MALDI-TOF MS module (n = 4 for *JK*01N.03*, and n = 3 for *JK*02N.01*). All remaining donors (n = 10;
183 0.08%) showed phenotypes that could not be explained by available genotype information. MALDI-
184 TOF MS results for the *JK*01/02* discrimination (c.838G>A) were confirmed for all 10 discrepant
185 samples by PCR-SSP.

186 Nanopore sequencing output

187 We obtained very high coverage for each amplified fragment in all 10 samples (Figure 3). Overall, the
188 nanopore sequencing run produced ~1 million reads (≥ Q7) with a median PHRED score of 12.6. After
189 demultiplexing and filtering reads according to fragment sizes, median number (± SD) of reads per
190 barcode (i.e. sample) was 38,688 (± 15,182). Mapping of filtered reads against *SLC14A1* reference
191 sequence showed a drop in coverage over a length of 5 kb in the second fragment for three samples
192 (Figure 3). For these same samples (s02, s03 and s07), the gel electrophoresis for verifying PCR
193 amplification successes prior to sequencing had also already revealed the presence of shorter
194 fragments in the second LR-PCR (Figure 2c).

195 Resolving discrepant cases

196 Serology and genetic characterization of the discrepant cases are provided in Table 1. In seven cases
197 we detected SNVs in coding regions that could explain the observed phenotype. In five donors, these

198 variants had previously been reported causing weak *(JK*01W.05*, *JK*02W.04*) and null alleles
199 (*JK*02N.06/08/09*), respectively. In the other two donors, we identified exonic SNVs that had so far
200 not been linked to Kidd phenotypes. In one case, the SNV was located in exon 3 (c.119G>A;
201 NG_011775.4:g.48415G>A) on a *JK*01* allele, altering the protein sequence close to the N terminus
202 (Gly40Asp). The other novel SNV was located in exon 7 (c.725G>A, NG_011775.4:g.57200G>A) on a
203 *JK*02* allele and corresponded to a missense mutation changing the codon 242 from glycine to
204 glutamic acid (Gly242Glu). Sanger sequencing confirmed the presence of all SNVs.

205 Three samples did not show any promising candidates when calling SNVs and short indels. However,
206 SV calling identified a large 5,095 bp deletion in all three samples (Figure 2a). Using NG_011775.4 as
207 reference, the software suggested breakpoints located before position 64,898 bp in intron 8 and after
208 position 69,992 bp in exon 10 (g.64898_69992del, Figure 2b). Bridge-PCR followed by Sanger
209 sequencing confirmed these exact breakpoint positions (Figure 2d). In all three samples, the deletion
210 was located on a *JK*01* allele (c.838G), and donors were *JK*01*/*JK*02* heterozygotes.

211 Discussion

212 By employing a combination of LR-PCR and long-read nanopore sequencing, we successfully resolved
213 all genotype-phenotype discrepancies in the Kidd blood group system observed over seven years of
214 routine high-throughput donor genotyping. Our developed protocol allowed sequencing the complete
215 coding region of the *SLC14A1* gene including introns as haplotypes, greatly facilitating the
216 identification of underlying genetic basis for the observed genotype-phenotype discrepancies and
217 directly producing high-quality haplotype reference sequences for newly described *JK* alleles.

218 The majority of *JK* weak and null alleles listed in the ISBT tables are known to occur at very low
219 frequencies across different populations[4,19,20], indirectly supporting the low number of discrepancies
220 identified in our donor population. Specifically, we detected the rare SNVs c.742G>A (rs763095261),
221 c.871T>C (rs78242949), c.956C>T (rs565898944), and c.191G>A (rs114362217) in our discrepant
222 samples, as part of *JK*01W.05*, *JK*02N.06*, *JK*02N.08* and *JK*02N.09* alleles, respectively (Table 1).
223 All these variants show minor allele frequencies (MAFs) in large-scale sequencing projects[21] that are
224 in close agreement with the exceeding rarity observed in our study (MAFs < 0.001). It is important to
225 acknowledge that these variants (alleles) were only detected in *JK*01/02* heterozygotes of our donor
226 population since discordant serology would not have appeared in homozygotes due to the masking
227 effect of the wildtype allele.

228 Of the ten observed genotype-phenotype discrepancies in our study, two were caused by known *JK*
229 alleles classified as weak by the ISBT (*JK*02W.04, JK*01W.05*) but appearing as putative null alleles in
230 our dataset. One case showed the c.130G>A (rs2298720) variant on a *JK*02* background (*JK*02W.04*
231 allele) as likely reason for the discrepancy. This variant itself is a common SNV (MAF > 0.05) which also
232 defines the most frequent weak allele on the *JK*01* background (*JK*01W.01*) along with several other
233 weak alleles[13]. Hence, this variant is expected to be frequent in our donor population (e.g. Jk(a) in s01
234 and s08, Table 1). However, it would normally not present as a discrepancy since weak Jk(a) or Jk(b)
235 phenotypes were not considered as discrepancy as long as the genotype information suggested the
236 presence of the respective allele. Weak phenotypes were neglected as clinically of low relevance.
237 Transfusion recommendations remain unchanged from those carrying the respective wildtype allele.
238 The present case of *JK*02W.04*, however, appeared as discrepancy as the c.130G>A variant seems to
239 have led to a null allele, i.e. a more complete suppression of the affected allele, or at least to very
240 weak expression as suggested by different serological methods (s04, Table 1). As a second case, our
241 data revealed a *JK*01W.05* allele as cause of the observed phenotype-genotype discrepancy. The
242 variant rs763095261, as part of the *JK*01W.05* allele, manifested as causing a null rather than a weak
243 allele (s05, Table 1). This is in line with previous publications which also described the *JK*01W.05* allele
244 in the context of null alleles underlying genotype-phenotype discrepancy cases.

245 Among the remaining cases that displayed genotype-phenotype discrepancies in our study, two were
246 found with exonic SNVs likely causing/defining new null alleles not yet listed by the ISBT. One of these
247 variants, c.725G>A (rs1197896884), is located in exon 7 and has been previously reported in two
248 Europeans in the gnomAD[22] sequence collection of more than 125,000 sequenced exomes and whole
249 genomes. The other variant, c.119G>A, is entirely novel. It appears to be a tri-allelic position. A
250 singleton with a different base change (G>C) at the same position resulting in a different amino acid
251 change (Gly40Ala) has been found in the gnomAD[22] sequence collection (rs1253002394). It is
252 noteworthy that a change of the identical amino acid (Gly40Ser, c.118G>A, rs145283450) leads to the
253 presence of a null alleles *(JK*01N.17, JK*02N.21*)[13,23].

254 Most remarkably, we discovered that three of the observed genotype-phenotype discrepancies in our
255 study were caused by an identical large structural variant. Our nanopore sequencing strategy revealed
256 a large deletion of approximately 5 kb spanning from exon 9 to 10, linked to the *JK*01* allele. This
257 deletion represents to the best of our knowledge the largest SV ever documented within the Kidd
258 blood group system. Currently, the Kidd ISBT allele table only encompasses one large deletion (~1.6
259 kb) defining the *JK*01N.01* allele. First reported in Tunisian women[24], it has since then been
260 sporadically reported in several other populations[25,26]. This deletion (c.1_341del) spans over exons 3

261 and 4 of the *SLC14A1* gene[24] and results in the lack of translation of the Kidd protein as the start codon
262 is located in exon 3. Considering the complete absence of exon 9 and a portion of exon 10 in alleles
263 carrying our newly discovered deletion, it is highly probable that the resulting protein is not functional.
264 Like for the two aforementioned SNVs causing *JK*02W.04, JK*01W.05*, a in-depth validation of these
265 variants causing true null rather than very weak alleles will require additional analyses using for
266 instance adsorption-elution or flow cytometric experiments.

267 The apparent lack of described SVs in blood group genes is not restricted to the Kidd system but
268 extends to almost all blood groups[13]. With the continuing accumulation of high-coverage whole
269 genome sequencing datasets and improved algorithms for SV discovery, it's becoming clear that SVs
270 appear to be more common than previously thought. Depending on the population targeted and the
271 sequencing technology used, between 7,500 SVs and 22,600 per genome were observed[27,28].
272 Additionally, SVs are responsible for an estimated ~30% of rare heterozygous (MAF < 1%) gene
273 inactivation events per individual[27]. Therefore, the extremely rare reports of SVs in blood group genes
274 could be linked to the lack of adequate tools for their detection, rather than being driven by
275 evolutionary forces.

276 The utilization of long-read sequencing has proven to be particularly advantageous for detecting and
277 characterizing SVs that often go undetected by traditional Sanger sequencing and short-read
278 sequencing methods[28]. Indeed, using Sanger sequencing alone, the sequencing results for exon 9 and
279 10 of these three samples harbouring the deletion would have appeared as perfectly homozygous
280 since only the wild-type allele would have been amplified and sequenced. Hence, results solely based
281 on Sanger sequencing would not have allowed us to determine the origin of discrepancy in three out
282 of 10 cases. In contrast, nanopore sequencing enabled us to precisely identify the breakpoints, which
283 were subsequently verified using bridge-PCR. This study serves as a compelling example for illustrating
284 the power of nanopore sequencing in resolving complex SVs that would have remained elusive using
285 conventional sequencing methods. With the continuing reduction in cost and increase in accessible
286 and easy-to-use bioinformatic tools, we dare to forecast that this and other TGS technology will be
287 more and more used to overcome challenging diagnostic cases encompassing hybrid genes, deletions,
288 and duplications.

289 **Conflict of Interest**

290 None of the authors declares any conflicts of interest.

291 **Funding Information**

292 This work was financially supported by the Blood Transfusion Service Zurich SRC (Switzerland).

293 **Data Availability Statement**

294 The identified 5,095 bp deletion NM_015865.7(SLC14A1):c.947-1453_*2066del on the *JK*01*
295 haplotype has been submitted to ClinVar (accession number VCV001202625.1).

296 # Tables

297 **Table 1. Summary table of serology, genotyping, and nanopore sequencing for all genotype-**
298 **phenotype discrepant cases.** Discrepant antigens, based on comparison of observed with deduced
299 phenotype, as well as causative alleles are highlighted in red.

Sample	Observed serology	Deduced phenotype based on MALDI-TOF MS genotyping	Nanopore Sequencing	
			Haplotype 1	Haplotype 2
Known weak and null alleles				
s01	Jk(a+^{weak} b-)	Jk(a+b+)	*JK*01W.01*	*JK*02N.08*
s04	Jk(a+b-)	Jk(a+b+)	*JK*01*	*JK*02W.04*
s05	Jk(a-b+)	Jk(a+b+)	*JK*01W.05*	*JK*02*
s06	Jk(a+b-)	Jk(a+b+)	*JK*01*	*JK*02N.09*
s08	Jk(a+^{weak} b-)	Jk(a+b+)	*JK*01W.01*	*JK*02N.06*
Novel null alleles				
s09	Jk(a+b-)	Jk(a+b+)	*JK*01*	*JK*02(G242E)Null*[§]
s10	Jk(a-b+)	Jk(a+b+)	*JK*01(G40D)Null*[§]	*JK*02*
Novel structural variant				
s02	Jk(a-b+)	Jk(a+b+)	*JK*01(Ex9_10del)*[§]	*JK*02*
s03	Jk(a-b+)	Jk(a+b+)	*JK*01(Ex9_10del)*[§]	*JK*02*
s07	Jk(a-b+)	Jk(a+b+)	*JK*01(Ex9_10del)*[§]	*JK*02*

300 § Novel alleles not yet described

Figures

Figure 1. Flowchart depicting selection and processing of genotype-phenotype discrepancies. Main bioinformatics tools are provided in green.

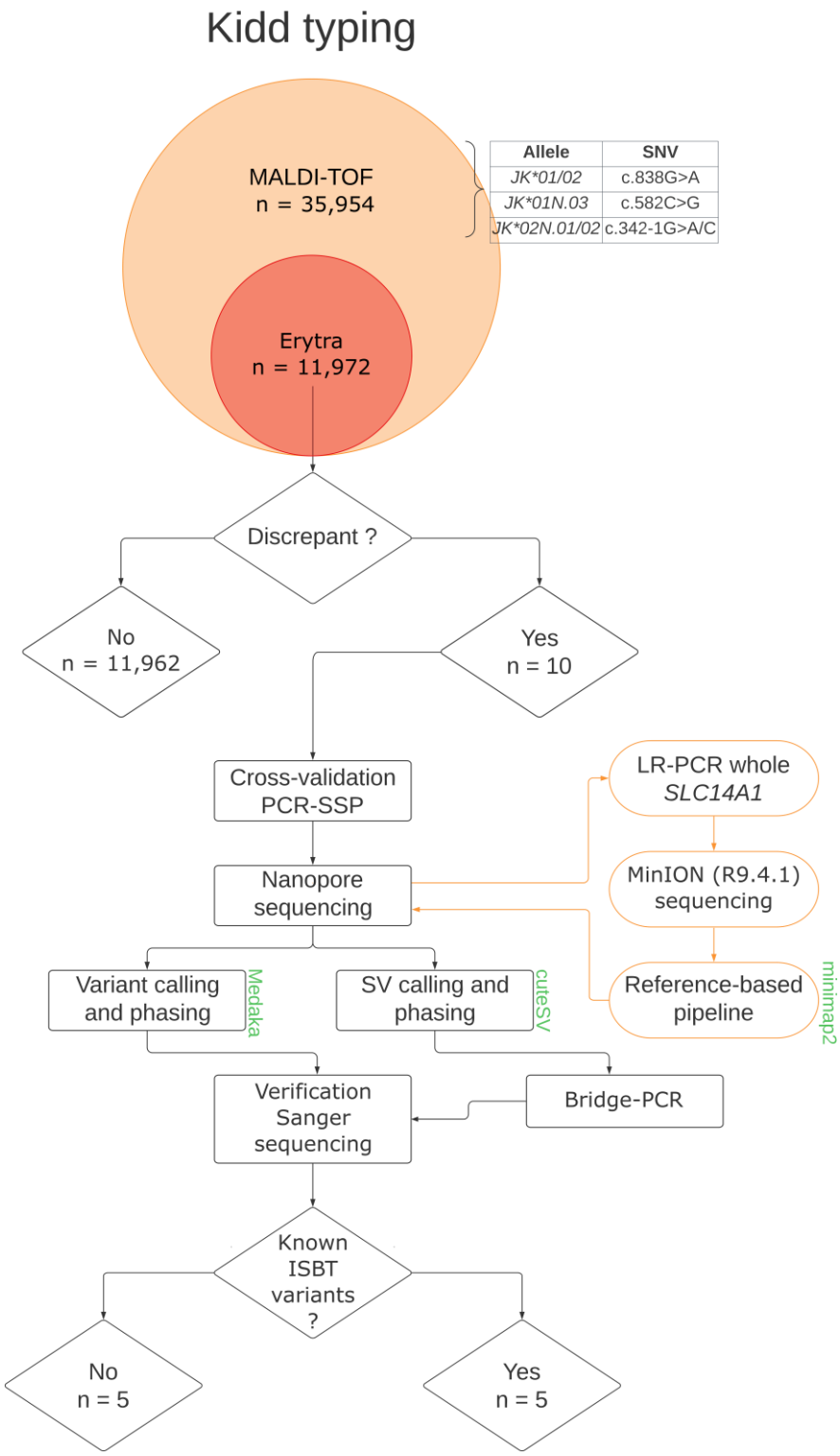


Figure 2. Genetic structure of *SLC14A1* gene encoding JK blood group and identification of a large deletion. (a) *SLC14A1* gene showing exons, coding DNA sequence (CDS), newly identified ~5 kb deletion (red), positions of long-range PCRs (LR1/LR2), as well as overlap between both amplicons. (b) Region of ~5 kb deletion with breakpoints and primer positions used for bridge-PCR. (c) Gel electrophoresis of long-range PCR products (LR1 and LR2) of all samples. (d) Gel electrophoresis of bridge-PCR products for the three samples with the novel deletion and wild-type control reactions.

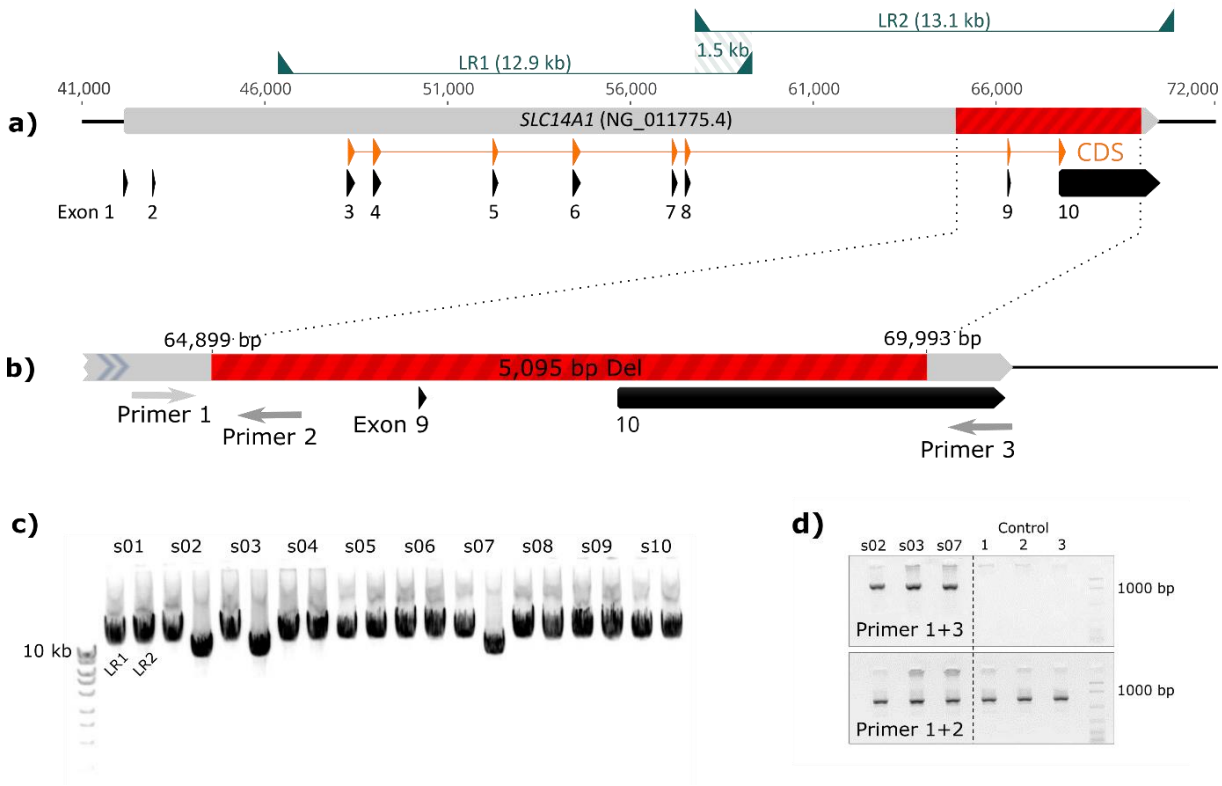
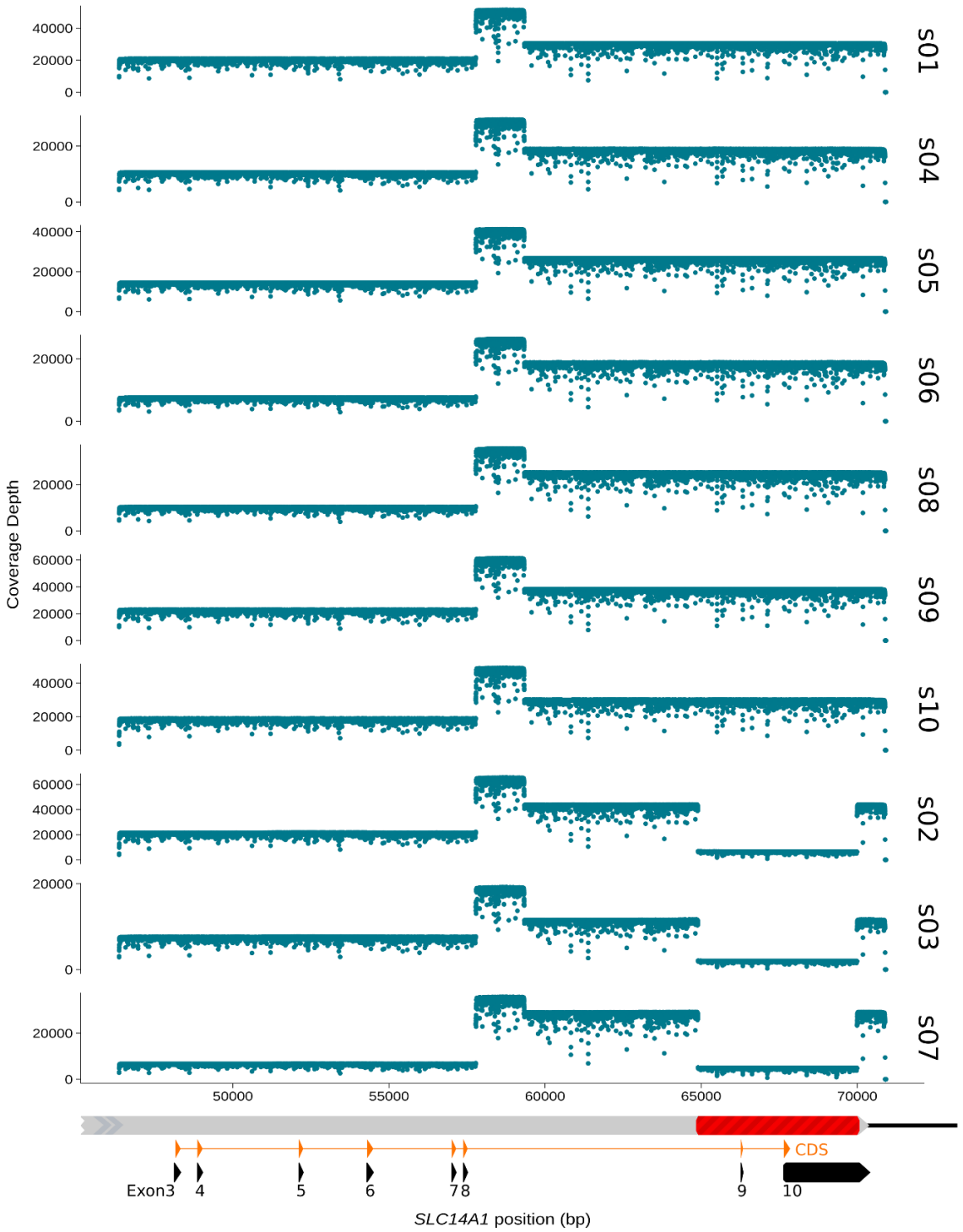


Figure 3. Nanopore sequencing coverage by nucleotide position for each sample. Coverages were computed by mapping length-filtered reads to *SLC14A1* reference sequence (NG_011775.4) using minimap2. For readability, samples showing a drop in coverage in LR2 were grouped together. Graphical representation of mapping location on *SLC14A1* gene is given underneath the coverages plots. Coverages scales are not fixed.



References

1. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze A V., Mikheenko A, et al. The complete sequence of a human genome. Science [Internet]. 2022 Apr;376(6588):44–53. Available from: https://www.science.org/doi/10.1126/science.abj6987

2. Lane WJ, Gleadall NS, Aeschlimann J, Vege S, Sanchis-Juan A, Stephens J, et al. Multiple GYPB gene deletions associated with the U− phenotype in those of African ancestry. Transfusion [Internet]. 2020 Jun 30;60(6):1294–307. Available from: https://onlinelibrary.wiley.com/doi/10.1111/trf.15839

3. Zhang Z, An HH, Vege S, Hu T, Zhang S, Mosbruger T, et al. Accurate long-read sequencing allows assembly of the duplicated RHD and RHCE genes harboring variants relevant to blood transfusion. Am J Hum Genet [Internet]. 2022 Jan 6;109(1):180–91. Available from: https://doi.org/10.1016/j.ajhg.2021.12.003

4. Montemayor C, Simone A, Long J, Montemayor O, Delvadia B, Rivera R, et al. An open-source python library for detection of known and novel Kell, Duffy and Kidd variants from exome sequencing. Vox Sang [Internet]. 2021 Apr 10;116(4):451–63. Available from: https://onlinelibrary.wiley.com/doi/10.1111/vox.13035

5. Thun GA, Gueuning M, Sigurdardottir S, Meyer E, Gourri E, Schneider L, et al. Novel regulatory variant in ABO intronic RUNX1 binding site inducing A3 phenotype. bioRxiv [Internet]. 2023;1–11. Available from: http://biorxiv.org/content/early/2023/05/04/2023.05.04.539366.abstract

6. Meyer S, Vollmert C, Trost N, Brönnimann C, Gottschalk J, Buser A, et al. High-throughput Kell, Kidd, and Duffy matrix-assisted laser desorption/ionization, time-of-flight mass spectrometry-based blood group genotyping of 4000 donors shows close to full concordance with serotyping and detects new alleles. Transfusion [Internet]. 2014 Dec;54(12):3198–207. Available from: https://onlinelibrary.wiley.com/doi/10.1111/trf.12715

7. Miller DE, Lee L, Galey M, Kandhaya-Pillai R, Tischkowitz M, Amalnath D, et al. Targeted long-read sequencing identifies missing pathogenic variants in unsolved Werner syndrome cases. J Med Genet [Internet]. 2022 May 9;jmedgenet-2022-108485. Available from: https://jmg.bmj.com/lookup/doi/10.1136/jmedgenet-2022-108485

8. Watson CM, Dean P, Camm N, Bates J, Carr IM, Gardiner CA, et al. Long-read nanopore sequencing resolves a TMEM231 gene conversion event causing Meckel–Gruber syndrome. Hum Mutat [Internet]. 2020 Feb 11;41(2):525–31. Available from: https://doi.org/10.1002/humu.23940

9. Gueuning M, Thun GA, Wittig M, Galati AL, Meyer S, Trost N, et al. Haplotype sequence collection of ABO blood group alleles by long-read sequencing reveals putative A1 -diagnostic variants. Blood Adv [Internet]. 2023 Mar 28;7(6):878–92. Available from: https://doi.org/10.1182/bloodadvances.2022007133

10. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 2021;39(11):1348–65.

11. Makałowski W, Shabardina V. Bioinformatics of nanopore sequencing. J Hum Genet [Internet]. 2020 Jan 26;65(1):61–7. Available from: http://www.nature.com/articles/s10038-019-0659-4

12. Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, et al. PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. Cell Genomics [Internet]. 2022 May;2(5):100129. Available from:

https://linkinghub.elsevier.com/retrieve/pii/S2666979X22000581

13. International Society of Blood Transfusion [Internet]. red cell immunogenetics and blood group terminology. [cited 2023 Jan 20]. Available from: https://www.isbtweb.org/isbt-working-parties/rcibgt.html

14. da Cunha Gomes EG, Machado LAF, de Oliveira LC, Neto JFN. The erythrocyte alloimmunisation in patients with sickle cell anaemia: a systematic review. Transfus Med. 2019;29(3):149–61.

15. Dean L. chapter10: The Kidd blood group. In: Blood groups and red cell antigens [Internet]. 2005. Available from: https://www.ncbi.nlm.nih.gov/books/NBK2272/

16. Gassner C, Meyer S, Frey BM, Vollmert C. Matrix-Assisted Laser Desorption/Ionisation, Time-of-Flight Mass Spectrometry–Based Blood Group Genotyping—The Alternative Approach. Transfus Med Rev [Internet]. 2013 Jan;27(1):2–9. Available from: http://dx.doi.org/10.1016/j.tmrv.2012.10.001

17. Meyer S, Trost N, Frey BM, Gassner C. Parallel Donor Genotyping for 46 Selected Blood Group and 4 Human Platelet Antigens Using High-Throughput MALDI-TOF Mass Spectrometry BT - Molecular Typing of Blood Cell Antigens. In: Bugert P, editor. New York, NY: Springer New York; 2015. p. 51–70. Available from: https://doi.org/10.1007/978-1-4939-2690-9_5

18. Hamilton JR. Kidd blood group system: a review. Immunohematology. 2015;31(1):29–35.

19. Vorholt SM, Hamker N, Sparka H, Enczmann J, Zeiler T, Reimer T, et al. High-Throughput Screening of Blood Donors for Twelve Human Platelet Antigen Systems Using Next-Generation Sequencing Reveals Detection of Rare Polymorphisms and Two Novel Protein-Changing Variants. Transfus Med Hemotherapy [Internet]. 2020;47(1):33–44. Available from: https://www.karger.com/Article/FullText/504894

20. Dinardo CL, Oliveira TGM, Kelly S, Ashley-Koch A, Telen M, Schmidt LC, et al. Diversity of variant alleles encoding <scp>Kidd, Duffy, and Kell</scp> antigens in individuals with sickle cell disease using whole genome sequencing data from the <scp>NHLBI TOPMed Program</scp>. Transfusion [Internet]. 2021 Feb 24;61(2):603–16. Available from: https://onlinelibrary.wiley.com/doi/10.1111/trf.16204

21. Sherry ST, Ward M, Sirotkin K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res. 1999 Aug;9(8):677–9.

22. Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. bioRxiv [Internet]. 2022 Jan 1;2022.03.20.485034. Available from: http://biorxiv.org/content/early/2022/10/10/2022.03.20.485034.abstract

23. Ramsey G, Sumugod RD, Lindholm PF, Zinni JG, Keller JA, Horn T, et al. A Caucasian JK*A/JK*B woman with Jk(a+b-) red blood cells, anti-Jkb, and a novel JK*B allele c.1038delG. Immunohematology. 2016 Sep;32(3):91–5.

24. Lucien N, Chiaroni J, Cartron JP, Bailly P. Partial deletion in the JK locus causing a Jknull phenotype. Blood [Internet]. 2002 Feb 1;99(3):1079–81. Available from: https://ashpublications.org/blood/article/99/3/1079/53498/Partial-deletion-in-the-JK-locus-causing-a-Jknull

25. Irshaid NM, Eicher NI, Hustinx H, Poole J, Olsson ML. Novel alleles at the JK blood group locus explain the absence of the erythrocyte urea transporter in European families*. Br J Haematol [Internet]. 2002 Feb;116(2):445–53. Available from: http://doi.wiley.com/10.1046/j.1365-2141.2002.03238.x

410 26. Wester ES, Johnson ST, Copeland T, Malde R, Lee E, Storry JR, et al. Erythroid urea transporter
411 deficiency due to novel JKnull alleles. Transfusion. 2008;48(2):365–72.

412 27. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, et al. A structural variation
413 reference for medical and population genetics. Nature [Internet]. 2020 May
414 28;581(7809):444–51. Available from: https://www.nature.com/articles/s41586-020-2287-8

415 28. Beyter D, Ingimundardottir H, Oddsson A, Eggertsson HP, Bjornsson E, Jonsson H, et al. Long-
416 read sequencing of 3,622 Icelanders provides insight into the role of structural variants in
417 human diseases and other traits. Nat Genet [Internet]. 2021 Jun 10;53(6):779–86. Available
418 from: http://dx.doi.org/10.1038/s41588-021-00865-4