Allele-level, whole gene blood group typing with PacBio HiFi sequencing to identify new alleles

Maaike Rijkers[1], Loes A.L. van de Pasch[1], Daniel Alvarez[1], Bart Valkenburg[1], Masja de Haas[2], Peter C. Ligthart[2], Barbera Veldhuisen[2], C. Ellen van der Schoot[3], Erik H. Rozemuller[1], Maarten Penning[1]

GenDx, Utrecht, The Netherlands & Sanquin Diagnostic Services

Introduction

Blood group typing is currently predominantly performed by serology or arrays identifying known SNP variants. Employing long-read sequencing, unambiguous and fully phased allele level sequencing data can be obtained. This allows for the identification of new blood group alleles in an unbiased manner.

Methods

For nine blood group genes, primers were designed up and downstream of the 5' and 3' UTR to obtain amplicons covering the whole genes. Sequencing was performed on a PacBio Sequel IIe sequencer, generating long HiFi reads which could cover the longest amplicon of 22kb. DNA samples of 70 healthy donors and 26 patients were used for sequencing of the nine blood groups and data analysis was performed in NGSengine with a customized blood group reference allele database based on ISBT alleles (https://www.isbtweb.org/, Oct 2021).

Results

High quality, fully phased PacBio sequencing data was obtained for all nine blood groups. Many samples contained alleles with mismatches compared to the alleles described by ISBT (October 2021). These alleles were annotated as new (Table 1), and submitted to GenBank.

Conclusion

By whole gene PacBio sequencing of blood groups, we were able to obtain fully phased, high quality sequencing data and thereby uniquely identify new alleles of nine minor blood groups. This demonstrates that the genetic diversity of blood groups in a relatively small sample panel, including the exons, is higher than anticipated.

Table 1. Blood group assays.
Whole gene amplification and sequencing was performed for nine blood group genes.

Blood group	Gene	Gene Length	New alleles	Exon variants	Non-synonymous
Kell	KEL	22 kb	51	5	4
Lewis	FUT3	10 kb	65	21	18
Duffy	ACKR1	4 kb	15	4	3
Diego	SLC4A1	20 kb	133	10	4
Colton	AQP1	14 kb	71	3	2
Landsteiner-Wiener	ICAM4	2 kb	4	1	1
H	FUT1	8 kb	38	0	0
Gerbich	GYPC	42 kb*	89	2	1
Vel	SMIM1	4 kb	23	0	0
		Total	489	46	33

* For Gerbich, a two amplicon strategy was designed, resulting in two 21 kb amplicons with an overlapping region.
All alleles found in the 96 samples for the nine blood group genes were compared to the known alleles described by the ISBT (database October 2021). Alleles not present in the database were identified as new.

Figure 1: Data example 1: Duffy (4 kb)
In the alignment view of NGSengine, the UTRs (blue), exons (yellow) and introns (black line) are depicted. The coverage bar indicates equal coverage (read depth) of the whole gene. The red line indicates full phasing of the first till the last heterozygous position. The blue and red triangles show the mismatches in comparison to the best matches known alleles (FY*01), meaning this sample contains 2 new alleles. Data quality is high, with low noise levels below 1% and an equal 50/50% allele balance, as shown in the base variation plot.

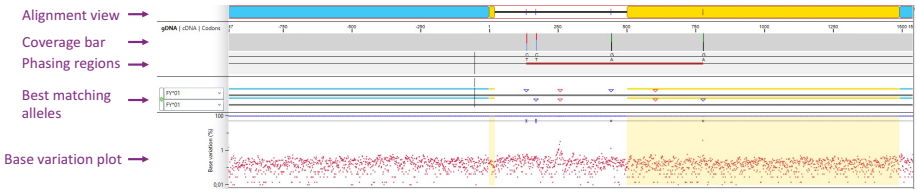
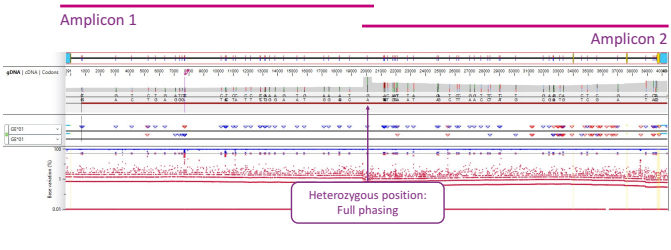


Figure 2: Data example 2: Gerbich (42 kb)
Gerbich is covered by 2 amplicons (both 21 kb), including an overlapping region. Fifty-four out of 96 samples were fully phased (dependent on presence of heterozygous position in overlapping region).



References

1 GenDx, Utrecht, The Netherlands 2 Department of Immunohaematology Diagnostics, Sanquin Diagnostic Services, Amsterdam, The Netherlands 3 Department of Experimental Immunohematology, Sanquin, Amsterdam, The Netherlands