THE FUTURE OF NEXT-GENERATION SEQUENCING FOR BLOOD GROUP GENOTYPING AND ITS IMPLICATIONS IN TRANSFUSION MEDICINE

by

AMR JAMAL HALAWANI

A thesis submitted to Plymouth University
in partial fulfillment for the degree of

DOCTOR OF PHILOSOPHY

School of Biomedical and Healthcare Sciences

July 2016

Acknowledgements

My first thank to God who blessed with the ability to complete this PhD project successfully. My special gratitude is to my parents and all the members of my family for their endless support.

I would like to express my sincere gratitude to Jazan University for funding my research and support me all the period of my scholarship in the United Kingdom.

I am extremely grateful to my lovely supervisors, Professor Neil Avent and Dr. Tracey Madgett for their infinite support. I have been privileged to work in Professor Neil Avent's lab to build my future as a professional researcher. I would like to thank Dr. Tara Meran, Dr. Michele Kiernan, Dr. Wondwossen Abate Woldie, Dr. Kris Jeremy and Dr. Paul Waines for their kind assistance in the laboratory. Finally, I would like to extend my thanks to my friends who care about me especially Dr. Majed Alghoribi and to anyone who helped me directly or indirectly all the way through my scholarship.

Author Declaration

At no time during the registration for the degree of *Doctor of Philosophy* has the author been registered for any other University award, without prior agreement of the Graduate Sub-Committee.

Name: Amr Jamal Halawani.

Signature:

Date:

Word count of the main body of thesis: 48,106 words.

Award

Margaret Kenwright Young Scientist Award 2014 from British Blood Transfusion Society (BBTS) for the following work:

Halawani, A. J., Altayar, M. A., Kiernan, M., Reynolds, A. J., Madgett, T. E. & Avent, N. D. 2014. Human Erythrocyte Antigens and Human Platelet Antigens Panel: A Genotyping Protocol Based on Next-Generation Sequencing. *Transfusion Medicine,* **24,** suppl. 2**,** 1-32.

Publications

Avent, N. D., Madgett, T. E., **Halawani, A. J**., Altayar, M. A., Kiernan, M., Reynolds, A. J. & Li, X. 2015. Next-generation sequencing: academic overkill or high-resolution routine blood group genotyping? *ISBT Science Series,* **10,** 250-256.

Conference proceedings

1. **Halawani, A. J.**, Altayar, M. A., Kiernan, M., Kaushik, N., Reynolds, A. J., Madgett, T. E. & Avent, N. D. 2013. Comprehensive Genotyping for Kell and Rh Blood Group Systems by Next-generation DNA Sequencing. *Transfusion Medicine,* **23,** suppl. 2**,** 30-71.

2. Altayar, M. A., **Halawani, A. J**., Kiernan, M., Kaushik, N., Reynolds, A. J., Madgett, T. E. & Avent, N. D. 2013. Next Generation Sequencing of ABO, Duffy and Kidd Blood Group Genotyping. *Transfusion Medicine,* **23,** suppl. 2**,** 30-71.

3. **Halawani, A. J.**, Altayar, M. A., Kiernan, M., Reynolds, A. J., Kaushik, N., Madgett, T. E. & Avent, N. D. 2014. Can Next-generation DNA Sequencing Solve the RH Complexity for Genotyping? *Vox Sang,* **107,** suppl. 1**,** 57-248.

4. Altayar, M. A., **Halawani, A. J.**, Kiernan, M., Reynolds, A. J., Kaushik, N., Madgett, T. E. & Avent, N. D. 2014 Extensive Genotyping of Blood Groups Duffy, Kidd and ABO by Next-generation Sequencing *Vox Sang*, **107,** suppl. 1**,** 57-248.

5. Avent, N. D., Madgett, T. E., **Halawani, A. J.**, Altayar, M. A., Kiernan, M. & Reynolds, A. J. 2014. Next Generation Sequencing: Academic Overkill or High-resolution Blood Group Genotyping? . *Vox Sang*, **107,** supple. 1**,** 1-56.

6. **Halawani, A. J.**, Altayar, M. A., Kiernan, M., Li, X., Madgett, T. E. & Avent, N. D. 2015. High Resolution Genotyping of the Rh Blood Group System by Next-generation Sequencing *Vox Sang,* **109,** suppl. 1**,** 1-379.

7. Altayar, M. A., **Halawani, A. J.**, Kiernan, M., Madgett, T. E. & Avent, N. D. 2015. Complete Gene Sequencing of ABO Blood Group by Next-generation Sequencing *Vox Sang,* **109,** suppl. 1**,** 1-379.

(See Appendix A for the abstracts).

Abstract

The Future of Next-generation Sequencing for Blood Group Genotyping and its Implications in Transfusion Medicine

Amr Jamal Halawani

Alloimmunisation becomes a problem when serological discrepancies occur in matching antigens between donors and patients for blood transfusion. The rate of alloimmunisation has been increased especially in multiply transfused patients. Blood group genotyping (BGG) is a DNA-based assay that aids in reducing this situation. Currently, many platforms of BGG have become available, in which every technique has its own advantages and disadvantages. All these platforms lack the ability to identify novel alleles that might have an unknown clinical significance. The advent of next-generation sequencing (NGS) offers identification of the unprecedented alleles due to its basis of sequence-based typing. Moreover, it provides an extreme high-throughput which may be able to screen many donors and patients in a single run. In this project, two approaches have been developed in generating sequencing libraries followed by sequencing on the Ion Torrent Personal Genome Machine™ platform (Ion PGM™). The first approach was amplicon-based target selection using Ion Ampliseq™ Custom Panel, designated as Human Erythrocyte Antigens and Human Platelet Antigens Panel (HEA and HPA Panel). This panel assay screens 11 blood group systems, as well as 16 human platelet antigens. The outcome was extraordinary, in particular four novel alleles had been identified out of 28 samples, one in the *RHCE* gene 208C>T (Arg70Trp) in exon 2 and three in the *KEL* gene. The first SNP was 331G>A (Ala111Thr) in exon 4. The second SNP was 1907C>T (Ala636Val) in exon 17 and the third SNP was 2165T>C (Leu722Pro) in exon 19. However, some issues occurred regarding co-amplification of homologous genes. The second approach was a long-range polymerase chain reaction (LR-PCR) based approach. This method provided a high resolution assay by amplification of entire genes, including the non-coding area, of the Kell and Rh blood group systems. The Kell blood group was initially utilised as a model in order to apply the same approach on the Rh system. Most alleles encoding the antigens of the Kell blood group, especially the high prevalence ones, were identified. The Rh LR-PCR approach was carried out by amplification of the *RHD* and *RHCE* genes with seven amplicons. For five RhD-positive samples no mutations were observed within the coding areas. On the other hand, five serotyped weak D samples were genotyped as; two weak D type 1, two weak D type 2 and one DAR3.1 weak partial D 4.0 (*RHD*DAR3.01*). Regarding the *RHCE*, the following antigens (C, c, E, c) were predicted properly from the sequencing data. Moreover, the *RHCE*ceVS.02* was identified. 64 and 39 intronic SNPs were identified in *RHD* and *RHCE* genes, respectively. The intronic SNPs assisted the genotyping process by identifying the haplotype of interest. Interestingly, the novel allele identified in the *RHCE* gene by the HEA and HPA Panel was confirmed to belong to the *RHCE* gene by the LR-PCR approach, indicating the panel misaligned it to the *RHD* gene. In conclusion, NGS paves the way to be an alternative substitution to the previous molecular techniques. It would supplant the conventional serology for typing blood for transfusion.

Table of Contents

Chapter 1 : Literature Review **1**
1.1 Introduction to blood groups *1*
1.2 Nomenclature of the blood groups *1*
1.3 The molecular basis for blood group polymorphism *5*
1.4 Blood group immunogenicity *6*
1.4.1 Haemolytic transfusion reactions 6
1.4.2 Haemolytic disease of the foetus and newborn 7
1.5 Rh blood group system *8*
1.5.1 Rh proteins 8
1.5.2 RhD polymorphism 11
1.6 Kell blood group system *19*
1.6.1 Kell glycoprotein 19
1.6.2 Molecular basis of Kell system 23
1.7 Other blood groups *26*
1.8 Human platelet antigens (HPAs) *28*
1.9 Typing of blood groups *29*
1.9.1 Serological methods 29
1.9.2 Genotyping of blood groups 31
1.10 Conventional sequencing 'Sanger sequencing' *34*
1.11 Next generation DNA sequencing *37*
1.11.1 Platforms of NGS 38
1.12 Thesis aims and objectives *44*
Chapter 2 : Materials and Methods **46**
2.1 Sourcing of blood samples *46*
2.2 Genomic DNA extraction *48*
2.2.1 Routine genomic DNA extraction 48
2.2.2 Genomic DNA extraction for long amplicons 49
2.2 Assessment of DNA quality and quantity *50*
2.3 Genotyping using microarray beads *51*
2.3.1 Multiplex PCR 51
2.3.2 Hybridisation 51
2.3.3 Labelling 52
2.3.4 Quantification 52
2.5 NGS libraries construction *53*
2.5.1 NGS libraries construction using a panel of short amplicons 53
2.5.5 Data analysis for NGS 73
Chapter 3 : Blood Group Genotyping by Microarrays **79**
3.1 Introduction *79*
3.2 Aims *80*
3.3 Results *81*
3.4 Discussion *86*
3.4.1 Cost of the microarray platforms 91
Chapter 4 : Human Erythrocyte Antigens and Human Platelet Antigens Panel (HEA and HPA Panel) **94**
4.1 Introduction *94*
4.2 Aims *95*
4.3 Results *96*
4.3.1 Sequencing report 98
4.3.2 Quality Control 100
4.3.3 Sequencing visualisation 104
4.3.4 Variant analysis using Ion Reporter™ software 115
4.3.5 Genotyping results of HEA and HPA Panel 117
4.4 Discussion *133*

4.4.1 Genotyping by the HEA and HPA Panel 133
4.4.2 Missed regions in the designed panel 135
4.4.3 Low depth of coverage 136
4.4.4 The issue of unspecific primers 136
4.4.5 Population screened 138
4.4.6 Scalability of the number of samples 138
4.4.7 The cost of HEA and HPA 141
Chapter 5 : Genotyping the Kell Blood Group by Next-generation Sequencing 143
5.1 Introduction 143
5.2 Aims 143
5.3 Results 145
5.3.1 Long range PCR for *KEL* gene 145
5.3.2 Sequencing Libraries 149
5.3.3 Analysis of NGS sequencing data 154
5.3.4 Quality Control 156
5.3.5 Sequencing visualisation 159
5.3.6 Variant analysis 161
5.3.7 Genotyping of the *KEL* 163
5.3.8 Validation of the Kell Antigens 166
5.4 Discussion 168
5.4.1 Quality control of the NGS 168
5.4.2 NGS data visualisation 168
5.4.3 Genotyping of the Kell antigens 168
5.4.4 Genotyping of the *KEL* 170
Chapter 6 : Genotyping the Rh Blood Group by Next-generation Sequencing 175
6.1 Introduction 175
6.2 Aims 176
6.3 Results 177
6.3.1 Long range PCR for Rh genes 177
6.3.2 Sequencing Libraries 181
6.3.3 Analysis of NGS sequencing data 190
6.3.4 Quality Control 192
6.3.5 Sequencing visualisation 198
6.3.6 Variant analysis 206
6.3.7 Genotyping of the Rh blood group system 206
6.4 Discussion 216
6.4.1 Primer design 216
6.4.2 LR-PCR Polymerase 217
6.4.3 The depth of coverage 218
6.4.4 Genotyping results of the Rh blood group system 219
6.4.5 Zygosity results 221
6.4.6 The cost of the NGS for the Rh blood group system 222
Chapter 7 : General Discussion and Conclusion 225
7.1 The HEA and HPA Panel 225
7.2 LR-PCR approach 227
7.3 Intronic SNPs 229
7.4 Small read length and cloning 231
7.5 Data analysis and storage 231
7.6 The cost 231
7.7 The future of the NGS technology 232
7.8 Future work 232
References 235
Appendices 256

List of Figures

Figure 1.1 The structure of the Rh proteins. 10
Figure 1.2 The genomic structure of the human Rh locus. 13
Figure 1.3 The molecular background of partial D phenotypes. 15
Figure 1.4 The structure of the Kell glycoprotein. 22
Figure 1.5 Sanger sequencing. 36
Figure 1.6 The sequencing reaction on the Ion PGM™. 43
Figure 2.1 The different blood samples were used in this PhD project. 47
Figure 2.2 An illustration of the indicated wells of the High Sensitivity DNA chip regarding the loading positions for the quantification assay. 64
Figure 2.3 The sample port that was used to prepare the sequencing template for the clonal amplification. 68
Figure 2.4 The eight wells for the enrichment of the template using the Ion OneTouch Enrichment System. 70
Figure 2.5 A brief summary of NGS data analysis workflow. 75
Figure 2.6 Masking the Rh genes. 76
Figure 4.1 Schematic diagram for constructing a sequencing library using Ion Ampliseq™ Custom Panel. 97
Figure 4.2 An overview of a report on the sequencing run of the HEA and HPA Panel. 99
Figure 4.3 Phred quality scores across all the bases for a single sample of the HEA and HPA Panel. 101
Figure 4.4 Quality scores per sequencing count for a single sample of the HEA and HPA Panel. 103
Figure 4.5 A homozygous SNP (hemizygous in this case for weak D sample). 106
Figure 4.6 A heterozygous SNP for HPA-15a/15b genotype in the *CD109* gene. 107
Figure 4.7 Good mapping quality for the sequencing reads 108
Figure 4.8 Poor mapping quality in exon 10 of the *RHD* gene. 109
Figure 4.9 Poor mapping quality for the sequencing reads for exon 8 of the *RHD* gene. 110
Figure 4.10 Missed regions within the exon 7 of the *ABO* gene 111
Figure 4.11 Allelic imbalance in exon 1 of the *RHCE* gene. 112
Figure 4.12 A comparison between two samples of the *GYPA* gene in the MNS blood group system. 113
Figure 4.13 A nucleotide insertion with *ABO* gene in two samples with A and B phenotypes. 114
Figure 4.14 An Excel sheet for the genotyping analysis report using the Ion Reporter™ software 116
Figure 4.15 A novel SNP that was misaligned to the *RHD* gene instead of the *RHCE* gene with an allelic imbalance ratio of (60:40). 129
Figure 4.16 A novel SNP found in exon 4 of the *KEL* gene. 130
Figure 4.17 A novel SNP found in exon 17 of the *KEL* gene. 131
Figure 4.18 A novel SNP found in exon 19 of the *KEL* gene. 132
Figure 5.1 Two LR-PCR products covered the whole *KEL* gene. 147
Figure 5.2 Amplification of the entire *KEL* gene by two products of LR-PCR. 148
Figure 5.3 An electropherogram of the fragmented *KEL* sequencing library (a pool of two *KEL* amplicons) 150
Figure 5.4 An electropherogram of the ligated products of the *KEL* sequencing library. 152
Figure 5.5 An electrophergram of the size selected *KEL* sequencing library. 153
Figure 5.6 An overview of a report on the sequencing run of the *KEL* gene. 155
Figure 5.7 Phred quality scores across all the bases for a single sample of the *KEL* gene. 157
Figure 5.8 Quality scores per sequencing count for a single sample of the *KEL* gene 158
Figure 5.9 A heterozygous SNP in the *KEL* gene for the *KEL*01.01/02* genotype. 160
Figure 5.10 The genotyping results of the annotated SNP for the *KEL*01.01/02* allele using SeattleSeq Annotation 141 website. 162
Figure 5.11 Validation of the SNP for the *KEL*01.01* allele encoding the K antigen 167
Figure 6.1 Seven LR-PCR products covered the entire of the *RHD* and *RHCE* genes. 179
Figure 6.2 LR-PCR products for the Rh blood group system. 180
Figure 6.3 An electropherogram of the fragmented *RHD* LR-PCR products. 182
Figure 6.4 An electropherogram of the fragmented *RHCE* LR-PCR products. 183
Figure 6.5 An electropherogram of the ligation of the *RHD* sequencing library 185

Figure 6.6 An electropherogram of the ligation of the *RHCE* sequencing library. 186
Figure 6.7 An electropherogram of the size-selected *RHD* sequencing library. 188
Figure 6.8 An electropherogram of the size-selected *RHCE* sequencing library. 189
Figure 6.9 An overview of a report on the sequencing run of the genes of the Rh blood group system. 191
Figure 6.10 Phred quality scores across all the bases for a single sample of the *RHD* gene. 193
Figure 6.11 Phred quality scores across all the bases for a single sample of the *RHCE* gene. 194
Figure 6.12 Quality scores per sequencing count for a single sample of the *RHD* gene.....196
Figure 6.13 Quality scores per sequencing count for a single sample of the *RHCE* gene.....197
Figure 6.14 An overview image of the coverage by sequencing for the entire *RHD* gene. 200
Figure 6.15 Regions with low depth of coverage for the first amplicon of the *RHCE* gene. ...201
Figure 6.16 High depth of coverage for the regions that covered exon 3 and exon 4 of the *RHCE* region by the second amplicon. 202
Figure 6.17 Very low depth of coverage in the *RHCE* gene in the regions between exon 5 and exon 10. 203
Figure 6.18 A misalignment in data visualisation of exon 8 of the *RHD* gene due to the different reference allele used by the IGV software. 205
Figure 6.19 A hemizygous SNP in exon 9 of the *RHD* gene indicating a weak D type 2 sample. 209
Figure 6.20 Five SNPs in the R_1r sample. 214
Figure 6.21 Low depth of coverage for *RHCE*e* allele. 215

List of Tables

Table 1.1 Blood group systems. 4
Table 1.2 List of the antigens of the Kell blood group system. 25
Table 1.3 Polymorphisms in other blood group systems. 27
Table 1.4 NGS platforms. 40
Table 2.1 Coverage summary of the HEA and HPA Panel. 54
Table 2.2 Primer sequences used in LR-PCR to amplify the entire *KEL* gene. 58
Table 2.3 Primer sequences used in LR-PCR to amplify the whole *RHD* gene. 60
Table 2.4 Primer sequences used in LR-PCR to amplify the whole *RHCE* gene. 60
Table 2.5 Primer sequences used to validate the SNP encoding the K antigen from NGS data. 78
Table 3.1 The serological results provided by NHSBT Filton for 28 samples that were used for microarray. 83
Table 3.2 The results of the predicted phenotypes of 24 samples using the ID Core + kit. 84
Table 3.3 The predicted phenotypes of four samples that had some non-valid results using the ID Core+ kit. 85
Table 4.1 The serological results provided by NHSBT Filton for the 28 samples. 123
Table 4.2 The genotyping results of the ABO blood group system obtained by the HEA and HPA Panel. 124
Table 4.3 Genotyping results using NGS of weak D samples obtained by the HEA and HPA Panel. 125
Table 4.4 The predicted phenotypes of NGS genotyping results for other blood groups by the HEA and HPA Panel. 126
Table 4.5 The predicted phenotypes of the HPAs from the genotyping results using HEA and HPA Panel. 127
Table 4.6 The novel alleles detected by the HEA and HPA Panel. 128
Table 4.7 Number of samples can be performed on the HEA and HPA 140
Table 5.1 Serology information for the 20 samples regarding the Kell phenotypes. 146
Table 5.2 Genotyping of the main alleles of the Kell blood group system by NGS. 164
Table 5.3 A list of intronic SNPs with the *KEL* gene. 165
Table 5.4 Number of NGS samples can be achieved for *KEL* gene 173
Table 6.1 Serology information provided by NHSBT Filton for the Rh status. 178
Table 6.2 The genotyping results of five weak D samples using LR-PCR amplification followed by 210
Table 6.3 Intronic SNPs found in the *RHD* gene by NGS. 211
Table 6.4 The genotyping results of the *RHCE* alleles. 212
Table 6.5 The number of samples of the *RHD* and *RHCE* genes that can be obtained with a depth of coverage of 100× in addition to the cost of the sequencing run. 223

Abbreviations

BAI	binary alignment index
BAM	binary alignment/map
BED	browser extensible data
BGG	blood group genotyping
BIDS	BLOODchip ID Software
bp	base pair
DAT	direct antiglobulin test
dCTP	deoxycytidine triphosphate
ddNTPs	dideoxynucleotides triphosphates
DEL	D-elute
DNA	deoxyribonucleic acid
dNTPs	deoxynucleotide triphosphates
EDTA	ethylenediaminetetraacetate
FNAIT	Foetal and neonatal alloimmune thrombocytopaenia
HDFN	haemolytic disease of the foetus and newborn
HEA	human erythrocyte antigens
HEA and HPA Panel	human erythrocyte antigens and human platelet antigens panel
HGP	human genome project
HLA	human leukocyte antigens
HPAs	human platelet antigens
HPLC	high performance liquid chromatography
HTR	haemolytic transfusion reaction
IAT	indirect antiglobulin test
ICSH	International Committee for Standaradisation in Haematology
IgG	immunoglobulin G
IgM	immunoglobulin M
IGV	Integrative Genome Viewer
Ion PGM™	Ion Torrent Personal Genome Machine™
ISBT	International Society of Blood Transfusion
ISPs	ion sphere particles
Kb	kilobases
KDa	kilodalton
LIMS	laboratory information management system
LR-PCR	long-range polymerase chain reaction

MPR	multitransfusion platelet refractoriness
NCBI	National Centre for Biotechnology Information
NGS	next-generation sequencing
NHSBT	National Health Service Blood and Transplant
PCR	polymerase chain reaction
PTP	posttransfusion purpura
RBCs	red blood cells
RFLP-PCR	restriction fragment length polymorphism-polymerase chain reaction
RhAG	Rh-associated glycoprotein
RNA	ribonucleic acid
RT-PCR	real time-polymerase chain reaction
SAPE	streptavidin and phycoerythrin
SCD	sickle cell disease
SNPs	single nucleotide polymorphisms
SNV	single nucleotide variant
SSP-PCR	sequence specific primers-polymerase chain reaction
T_m	melting temperature
TMEM50A	transmembrane 50A
VCF	variant call format
WES	whole exome sequencing
WGS	whole genome sequencing

Amino Acids Abbreviations

Amino Acid	3 Letters code
Alanine	Ala
Arginine	Arg
Asparagine	Asn
Aspartic acid	Asp
Cysteine	Cys
Glutamic acid	Glu
Glutamine	Gln
Glycine	Gly
Histidine	His
Isoleucine	Ile
Leucine	Leu
Lysine	Lys
Methionine	Met
Phenylalanine	Phe
Proline	Pro
Serine	Ser
Threonine	Thr
Tryptophan	Trp
Tyrosine	Tyr
Valine	Val

Chapter 1 : Literature Review

1.1 Introduction to blood groups

The concept that led to the discovery of blood groups began in 1901 when Karl Landsteiner observed that the isolated plasma of some individuals clumped 'agglutinated' the RBCs of others. Accordingly, three groups (A, B, and O) have been identified, which were found to be antigens of the first blood group system, the ABO blood group system. In 1901 and 1902, a fourth antigen (AB) was discovered by Decastello and Sturli, respectively (Landsteiner, 1961). According to the International Society of Blood Transfusion (ISBT), 36 blood group systems have been identified, which comprise more than 300 antigens (International Society of Blood Transfusion, 2015).

The blood group antigen can be defined as an inherited marker outside the surface of the red blood cells (RBCs), which can be detected via a unique alloantibody. The antigens can be either proteins or carbohydrates. The carbohydrate antigens are attached to proteins in the form of glycoprotein or to lipids to be glycolipids (Daniels, 2013a; Reid et al., 2012). The antigens of the blood group are encoded by a single gene, for example the ABO blood group or by a cluster of two or three closely linked genes, such as the Rh blood group system and MNS blood group system, respectively (Avent and Reid, 2000; Reid, 2009). The next section will describe the nomenclature of the ISBT.

1.2 Nomenclature of the blood groups

In 1980, the ISBT set up the Working Party on Red Cell Immunogenetics and Blood Group Terminology to recognise the blood group antigens. Recently, the Working Party has become a Committee. The antigens must be identified by both computer software and eye. **Table 1.1** lists the table of the blood group systems, including the system

name, the ISBT symbol, the ISBT number, the number of antigens per system, the gene(s) encoding the system and its chromosomal location. Every authenticated antigen is classified into four categories which are systems, collections, 700 series and 901 series (Daniels et al., 2004).

Every system contains one or more antigens and is controlled by a single gene or two or three homologous genes. Every antigen that belongs to a blood group system is distinguished with a six-digit number. For instance, the Kell blood group system represents the first three digits of the system (006) and the second three digits recognise the antigen (006003), which is the Kp^a antigen in this case. Otherwise, the system symbol is used instead of the system number and is followed by the antigen number. For example, KEL003 or KEL3 without sinistral zeros. Phenotypes are observed by the ISBT symbol name then a colon followed by a list of all the antigens represented and commas between them. Alleles are denoted by the system symbol followed by an asterisk and then the antigen number in which everything needs to be italicised, such as (*KEL*3)*. Genotypes are written as the ISBT system symbol then an asterisk followed by alleles or haplotypes with a slash separating them, in which everything is written in *italics*. An example for the genotyping is *(KEL*2,3/2,4)*. Null allele 'Amorph' is designated by zero. For instance, *(KEL2,3/0)* (Daniels et al., 2004).

The collections are the identified antigens either serologically, biochemically or genetically which do not comply with the criteria of the blood group system. This is normally because the gene identity was not recognised. There are six sets of antigen collections and Vel was one of them until recently when it became a blood group system due to being genetically investigated (Cvejic et al., 2013; Storry et al., 2013). 700 series is the set of antigens with low prevalence, which occurs in less than 1% of the population. 901 series (previously denoted as 900 series) is a set of antigens

expressed in more than 90% of the population. Both 700 and 901 series do not belong to any blood group systems.

Table 1.1 Blood group systems.

System name	ISBT symbol	ISBT number	Number of antigens	Gene names	Chromosome
ABO	ABO	001	4	*ABO*	9
MNS	MNS	002	46	*GYPA, GYPB, GYPE*	4
P1PK	P1PK	003	3	*A4GALT*	22
Rh	RH	004	54	*RHD, RHCE*	1
Lutheran	LU	005	20	*BCAM*	19
Kell	KEL	006	35	*KEL*	7
Lewis	LE	007	6	*FUT3*	19
Duffy	FY	008	5	*DARC*	1
Kidd	JK	009	3	*SLC14A1*	18
Diego	DI	010	22	*SLC4A1*	17
Yt	YT	011	2	*ACHE*	7
Xg	XG	012	2	*XG, CD99*	X/Y
Scianna	SC	013	7	*ERMAP*	1
Dombrock	DO	014	8	*ART4*	12
Colton	CO	015	4	*AQP1*	7
Landsteiner-Wiener	LW	016	3	*ICAM4*	19
Chido/Rodgers	CH/RG	017	9	*C4A, C4B*	6
H	H	018	1	*FUT1*	19
Kx	XK	019	1	*XK*	X
Gerbich	GE	020	11	*GYPC*	2
Cromer	CROM	021	18	*CD55*	1
Knops	KN	022	9	*CR1*	1
Indian	IN	023	4	*CD44*	11
Ok	OK	024	3	*BSG*	19
Raph	RAPHJ	025	1	*CD151*	11
John Milton Hagen	JMH	026	6	*SEMA7A*	15
I	I	027	1	*GCNT2*	6
Globoside	GLOB	028	1	*B3GALT3*	3
Gill	GIL	029	1	*AQP3*	9
Rh-associated glycoprotein	RHAG	030	4	*RHAG*	6
Forssman	FROS	031	1	*GBGT1*	9
JR	JR	032	1	*ABCG2*	4
LAN	LAN	033	1	*ABCB6*	2
Vel	VEL	034	1	*SMIM1*	1
CD59	CD59	035	1	*CD59*	11
Augustine	AUG	036	1	*SLC29A1*	6

Adapted from (International Society of Blood Transfusion, 2015).

1.3 The molecular basis for blood group polymorphism

Basically, the presence of two or more variants in the population leads to the polymorphisms of the blood group antigens. The cloning of the blood group genes paves the way for a better understanding of the molecular background of the blood group variants. In 1986, the first gene of the MNS blood group system, *GYPA* was cloned (Siebert and Fukuda, 1986). Consequently, more genes have started to be cloned such as ABO and Rh genes were cloned in 1990 and 1992, respectively (Yamamoto et al., 1990; Le van Kim et al., 1992; Avent et al., 1990; Cherif-Zahar et al., 1990).

Most of the blood group variants give rise to amino acid substitutions as the consequence of single nucleotide polymorphisms [SNPs] (Daniels, 2005). Other genetic mechanisms are involved such as deletion, for example, the deletion of the entire *RHD* gene leads to RhD-negative individuals in Caucasians (Colin et al., 1991). Moreover, the Vel-negative antigen is a result of the deletion of 17 base pairs (bp) in the *SMIM1* gene (Cvejic et al., 2013). Furthermore, a single nucleotide deletion that leads to a reading frameshift occurs in phenotypes, such as *O* and A^2 alleles of the ABO blood group system (Yamamoto et al., 1992; O'Keefe and Dobrovic, 1996). Other mechanisms are involved such as sequence duplication in association with missense and nonsense mutations as in RhD-negative individuals of African descent, which results in an inactive *RHD* gene, pseudogene [*RHDΨ*] (Singleton et al., 2000). In addition, genetic mechanisms include an intergenic recombination between highly homologous genes that appear in Rh and MNS blood group systems. An example of a hybrid gene is *RH (D-CE-D)* of the D^{VI} antigen type I, which occurs by substituting exons 4 and 5 of the *RHD* gene, with *RHcE* allele encoding proline at position 226 (Avent, 1997). Null phenotypes can be expressed, such as K_{null} phenotype, in which no expression of the antigens of the Kell blood group system can be observed. This makes the transfusion

become more complicated in order to obtain a blood unit that matches this rare phenotype (Storry and Olsson, 2004).

1.4 Blood group immunogenicity

Naturally occurring antibodies to either A antigen, B antigen or both are expressed in adults regarding the antigen they lack. These antibodies are predominantly in the form of immunoglobulin M (IgM) antibodies, while the immune antibodies are predominantly in the form of immunoglobulin G (IgG) antibodies. Other blood group antibodies arise due to immunisation as a result of transfusion or transplantation. Moreover, they could arise from pregnancy which may lead to a condition denoted as haemolytic disease of the foetus and newborn [HDFN] (Daniels and Bromilow, 2013). The next subsections will describe these incidences briefly.

1.4.1 Haemolytic transfusion reactions

The transfusion of mismatched blood units will put the recipient's life at risk as a consequence of haemolytic transfusion reactions (HTR). Intravascular HTR is characterised by rapid haemolysis, in which most of the RBCs are damaged within 10 minutes. It is caused by IgM antibody-mediated haemolysis in the circulation, which activates the classical complement pathway followed by the formation of a membrane attack complex. Accordingly, this punctures the RBC membrane and releases the haemoglobin into the plasma. The antibodies of the ABO blood group system are predominantly involved with intravascular HTR. The clinical manifestations include shock, chills, hypotension, haemoglobinaemia and haemoglobinuria with further complications of disseminated intravascular coagulation and renal failure (Poole and Daniels, 2007).

Extravascular HTR can be immediate which occurs within a few hours following the transfusion or delayed which occurs within 5 to 7 days. The IgG is an antibody of this type, which is not associated with the complement activation and not involved in the ABO system. Basically, RBCs coated with IgG bind to Fc receptors on macrophages and trigger a process called 'phagocytosis' in the liver and spleen to eliminate the damaged cells. The symptoms of extravascular HTR are the same as intravascular HTR but less severe (Poole and Daniels, 2007).

1.4.2 Haemolytic disease of the foetus and newborn

HDFN occurs when the foetus' red cells with positive antigens of paternal origin leak into the maternal circulation of a mother who is antigen-negative, which is designated as foetomaternal haemorrhage (Urbaniak and Greiss, 2000). Consequently, the mother's immune system recognises the positive foetal antigens as 'foreign' and starts producing IgM followed by IgG antibodies (mostly IgG1 and IgG3) against foetal RBCs. In subsequent pregnancies, those antibodies are capable of crossing the placenta, entering the foetal circulation and destroying foetal RBCs by splenic macrophages (Kumpel and Elson, 2001; Kumpel, 2008). Anti-D is the most common antibody to cause HDFN (Bowman, 1998)

The HDFN can be managed by giving the immunised mother an injection of anti-D prophylaxis. Hence, this allows for rapid elimination of foetal RBCs via the maternal spleen by macrophages (Scott, 2001). It was thought that HDFN is exclusive to anti-D. However, many antibodies were reported including anti-c, anti-E, anti-K and anti-Fy^a (Babinszki and Berkowitz, 1999). The HDFN by anti-K is caused by suppressing the erythropoiesis, rather than haemolysis (Weiner and Widness, 1996).

1.5 Rh blood group system

The Rh blood group system (004) is extremely polymorphic and is considered to be the most complex blood group system. It contains 54 antigens numbered from RH1 to RH61, while seven antigens have become obsolete (International Society of Blood Transfusion, 2015). There are five chief antigens, which are D, C, c, E and e. Other antigens can be represented in composite specificities, for instance, ce or f antigen which is an epitope on the Rhce protein and is not observed with RhCe or RhcE. Many antigens are correlated to the ethnic group of a certain population. For example, V and VS antigens, which are an altered e antigen, are mainly found in Blacks (Chou and Westhoff, 2010).

The antigens of the Rh blood group system can be involved in HTR and HDFN. Anti-D is the most important antibody to cause severe HDFN which might lead to foetal mortality (Levine et al., 1941). Anti-D IgG prophylaxis has resolved the immunisation by anti-D antibodies (Freda et al., 1966). However, it has been reported that other Rh antibodies, such as anti-c and anti-E are involved in causing HDFN as mentioned above (Koelewijn et al., 2008). Furthermore, they are involved in HTR in patients who require a blood transfusion such as patients with sickle cell disease (SCD) when incompatibility of the blood has occurred and caused alloimmunisation (Aygun et al., 2002).

1.5.1 Rh proteins

RhD and RhCE polypeptides have 12 transmembrane domains expressed with a cytosolic NH_2 and COOH termini, forming six extracellular loops in which the immune responses are directed (Avent et al., 1990). The RhD/RhCE protein is associated with two Rh-associated glycoprotein (RhAG) proteins to form a trimer (Conroy et al., 2005). RhAG shares 36% identity of the RhD and RhCE proteins and is glycosylated on its

first loop, while the RhD and RhCE proteins are not (Eyers et al., 1994; Marini et al., 1997). The RhD protein is different from the RhCE protein by 30–35 amino acid substitutions. This is according to which *RHCE* allele is inherited [*RHce, cE, Ce, CE*] (Mouro et al., 1993). **Figure 1.1** shows the structures of the RhD, RhCE, and RhAG proteins.

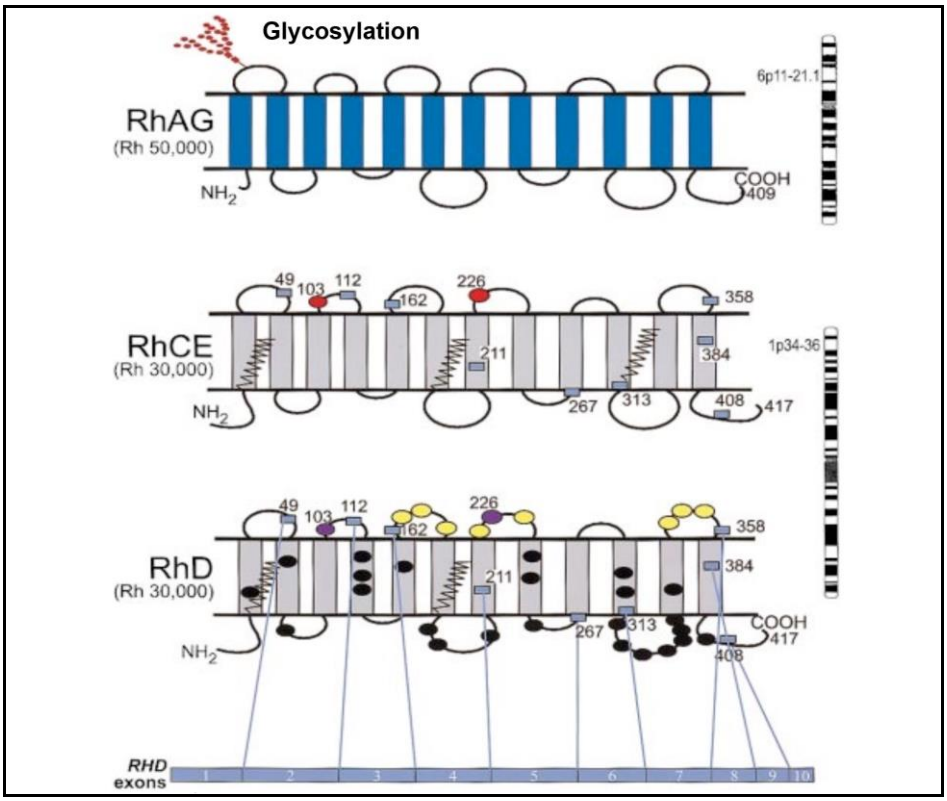


Figure 1.1 The structure of the Rh proteins.

12 transmembrane domains are expressed with cytosolic NH2 and COOH forming six extracellular loops. Glycosylation is taking place in the first extracellular loop of the RhAG. Two amino acid substitutions (Ser103Pro and Pro226Ala) are located in the second and fourth extracellular tubes which are mainly responsible for the antigenic polymorphism C/c and E/e of the RhCE protein. Arg229 has been identified as a crucial residue for E/e epitope existence. Although Pro226 is an essential residue; however, it is not adequate for the full existence of the E antigen. Adapted from (Avent and Reid, 2000).

1.5.2 RhD polymorphism

The Rh locus is located on chromosome 1p34-1p36. It consists of two highly homologous genes that share 92% sequence identity, *RHD* and *RHCE* genes (Cartron, 1994). A tandem arrangement of both genes is in an opposite orientation [*RHD* (5'- > 3')-(3' - < 5') *RHCE*]. They are separated by a third gene called *transmembrane protein 50A (TMEM50A)* previously denoted as *SMP1,* the role of which remains vague. Two rhesus boxes (approximately 9000 bp in length) flank the *RHD* gene and exhibit 98.6% homology to each other (Wagner and Flegel, 2000).

The RhD phenotype of an individual is determined by the presence or absence of the entire *RHD* gene being RhD-positive or RhD-negative **[Figure 1.2-a]**. Many polymorphisms exist in RhD-negative individuals. In Caucasians, for example, the entire *RHD* gene is deleted (Colin et al., 1991). Accordingly, this could result in an unequal crossover in the rhesus boxes which leads to the existence of a hybrid rhesus box, as shown in **Figure 1.2-b** (Wagner and Flegel, 2000). On the other hand, the RhD-negative phenotype in Africans could arise from stop codons as a result of point mutations or gene rearrangements. 66% of the RhD-negative Africans have the RhD-pseudogene (*RHDΨ*), resulting in an *RHD* inactive gene **[Figure 1.2-c]**. This is because the exhibition of a 37 bp nucleotide insertion in intron 3 and exon 4 causing disruption to the open reading frame by producing a premature stop codon (Tyr269stop). Moreover, there are three missense mutations in exon 5 of the *RHD* gene. This phenotype is usually associated as being *in cis* with *RHCE*ce*. In addition, 15% of Africans have hybrid genes (*RHD-CE-D*^{s}), which is associated with the production of c, VS, weak C and e antigens. 18% of RhD-negative Africans observe the deletion of the entire *RHD* gene (Singleton et al., 2000).

The RhD-negative phenotype is considered to be a rare phenotype in the Asian population. However, different mechanisms could be implicated for RhD-negative observation. These include the deletion of the entire *RHD* gene and D-elute (DEL) phenotypes **[section 1.5.2.4]** (Shao et al., 2002; Sun et al., 1998). Moreover, hybrid genes were determined in RhD-negative Asian individuals including *RHD-CE(2-9)-D* and *RHD-CE(3-9)-D* (Peng et al., 2003; Xu et al., 2005).

It has been noticed that some RhD-positive individuals form anti-D when they face RhD-positive RBCs resulting from transfusion or pregnancy. Therefore, the D antigen was considered to be a mosaic which comprises multiple epitopes with a lack of variants of D categories (Scott et al., 1996). The RhesusBase website lists regularly updated information for all Rh-variant alleles, including partial D alleles, weak D alleles, RhD-negative alleles, and RhCE allelic variants (The RhesusBase, 2015).

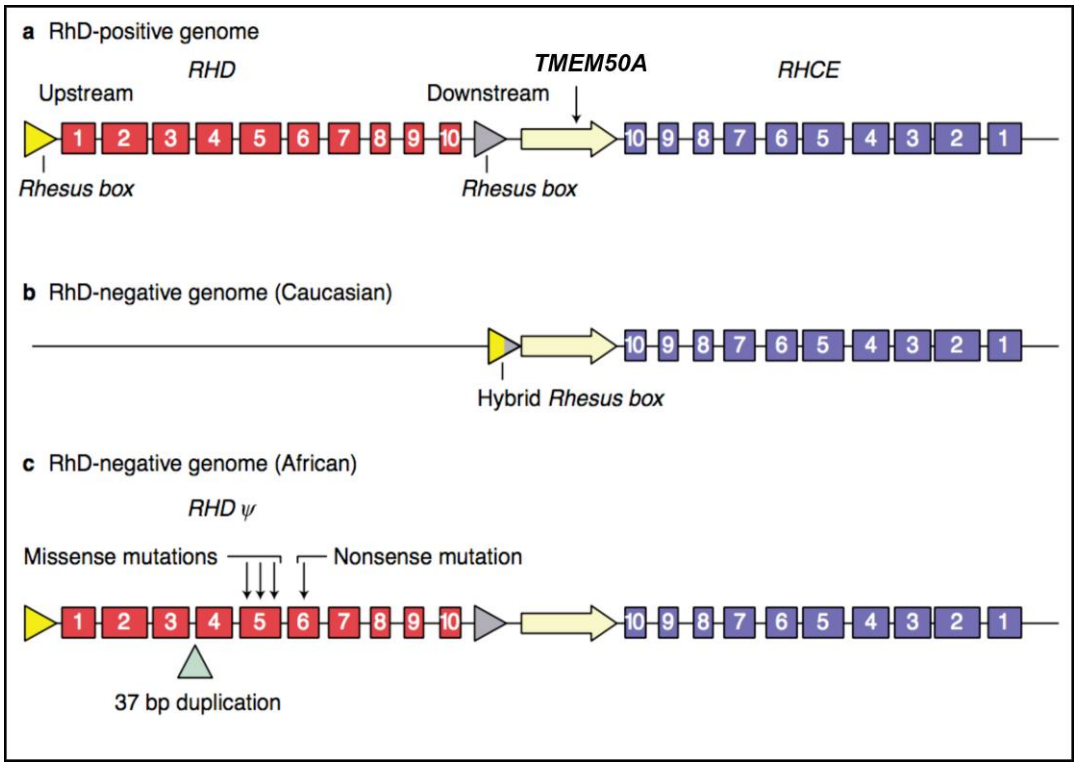


Figure 1.2 The genomic structure of the human Rh locus.

(a) The RhD-positive individuals possess the *RHD* gene. (b) The RhD-negative phenotype occurs in Caucasians as a result of the deletion of the entire *RHD* gene. (c) RhD-negative Africans have a 37 bp duplication in intron 3 and exon 4, associations of missense mutations in exon 5, and nonsense mutation within exon 6. Adapted from (Avent et al., 2006).

1.5.2.2 Partial D

RhD-positive individuals carry grossly conventional D antigen, which possesses more than 30 epitopes. Individuals with partial D or qualitative D variants carry modified *RHD* genes, which lack some of the D epitopes that make its identification by a monoclonal anti-D difficult. These individuals can produce alloanti-D if they are receiving a blood transfusion with intact D antigens because they have all the epitopes (Scott et al., 2000; Jones et al., 1995). This could arise from a gene conversion between *RHD* and *RHCE* genes, which results in the formation of a hybrid gene (*RHD-CE-D)* and the replacement of the *RHD* part by the corresponding *RHCE* part. Furthermore, partial D may occur due to the presence of missense mutations in the extracellular region of the RhD protein or multiple point mutations such as DAR alleles (Flegel and Wagner, 2002). **Figure 1.3** illustrates some types of partial D.

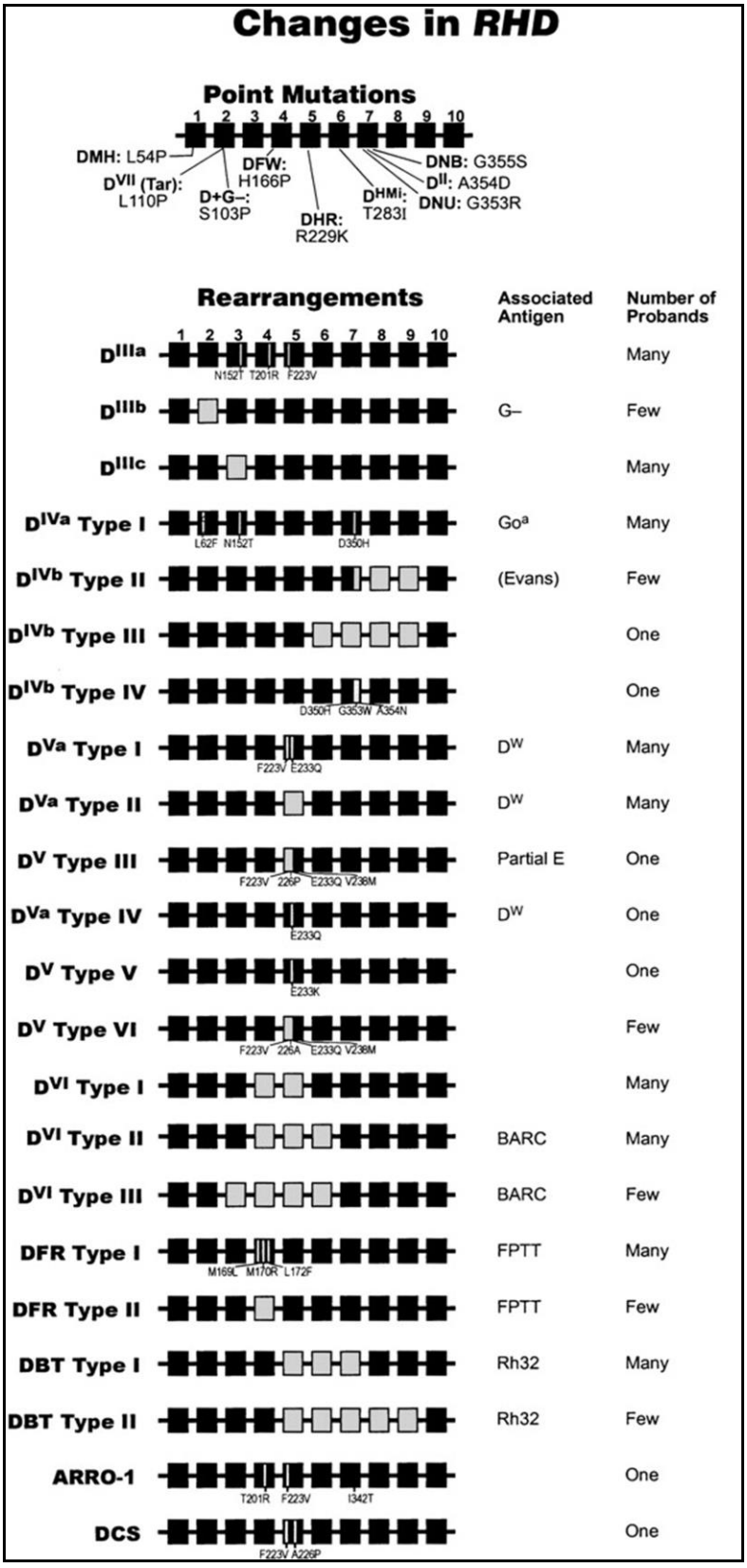


Figure 1.3 The molecular background of partial D phenotypes.

The black boxes depict the exons of *RHD*, while the grey boxes are the exons of the *RHCE* gene. Adapted from (Avent and Reid, 2000).

1.5.2.3 Weak D

Basically, the normal RhD-positive RBCs normally possess 10,000 to 30,000 antigen sites per red cell. Weak D or quantitative D variants (formerly denoted as D^u) express all the D epitopes but at a weak level. The weak D RBCs express a lower quantity of antigen sites per red cell, which ranges from 60 to 5200 depending on the weak D type. There currently are 73 types of weak D known with a further single subdivision for type 1 and type 2 (Reid et al., 2012).

Wagner and co-workers (1999) revealed the mystery through sequencing the *RHD* gene from weak D individuals. Nearly all weak D phenotypes are featured by missense mutations existing within the intracellular and transmembrane domains of the RhD protein. Therefore, it may be expected to restrict with the subunit assembly of the RhD protein complex, resulting in a decrease in the density of the D antigen site (Wagner et al., 1999).

In contrast to partial D, individuals with weak D do not normally generate an alloanti-D. However, in rare situations weak D individuals can produce an alloantibody (Wagner et al., 2000; McGann and Wenk, 2010). Therefore, the current term "weak partial D" has been designated for such phenotypes (Reid et al., 2012; The RhesusBase, 2015). Daniels (2013) suggested using a collective term "D variant" for both partial D and weak D. This is in order to prevent ambiguous terminology (Daniels, 2013b). The extreme expression of the D antigen is discussed in the next section.

1.5.2.4 Very weak D antigen D-elute (DEL)

The extreme weakness of expression of the D antigen can occur. This phenotype is designated as DEL and could possibly be mistyped as RhD-negative in blood banks. The individuals of DEL express a very weak form of the D antigen, which is difficult to

detect via serological typing such as the indirect antiglobulin test (IAT) and can only be investigated by adsorption and elution. Different genetic mechanisms are involved to encode DEL phenotypes. They are most commonly reported in Asians and are often caused by the deletion of 1013 bp in the region between intron 8 and intron 9 in the *RHD* gene (Chang et al., 1998). Moreover, it can be caused by a silent mutation 1227G>A which causes interference in the splicing of exon 9 (Shao et al., 2002). Furthermore, it was investigated that the European population shows DEL phenotypes (Körmöczi et al., 2005). According to many studies of high-throughput *RHD* genotyping, a significant number of DEL red cell units were identified in an RhD-negative donation pool (Flegel et al., 2009). According to the different genetic backgrounds of the D variants and because some of them are located in intronic regions of the *RHD* gene, the long-range polymerase chain reaction (LR-PCR) approach was designed followed by next-generation sequencing (NGS) in order to assist in the determination of any intronic mutations (see **Chapter 6**).

1.5.2.5 RHCE polymorphism

The principal antigens (C, c, E, e) are the products encoded by the *RHCE* gene. The amino acid substitution Ser103Pro in exon 2 is the main factor for C/c antigen expression and it is the only one outside the RBC membrane in the second extracellular loop. Moreover, there are another three amino acid substitutions, which lead to C/c expression, Cys16Trp in exon 1 and (Ile60Leu and Ser68Asn) in exon 2 (Mouro et al., 1993). Cys16 is not an essential amino acid in the expression of the C antigen (Westhoff et al., 2001). The presence of E/e antigens depends on amino acid substitution Pro226Ala in exon 5. This is located on the fourth extracellular loop of the RhCE protein (Simsek et al., 1994). Arg229 has been identified as a critical residue for E/e

epitope existence. Pro226 is an essential residue; however, it is not sufficient for the full existence of the E antigen (Chen et al., 2004).

The variants of the *RHCE* gene are observed as a typical *RHD* gene along with a modified *RHCE* gene. Point mutations within the *RHCE* gene give rise to modified phenotypes such as C^w and C^x (Mouro et al., 1995). A rare E^w antigen can be formed by missense mutation resulting in Met167Lys in exon 4 (Strobel et al., 2004). This antigen forms an antibody to the wild type of E antigen, which can be implicated in HDFN (Grobel and Cardy, 1971).

Weak expression of the e antigen can be found in Black ethnicity such as V and VS antigens. The V antigen results from amino acid substitution Leu245Val in association with a hybrid gene in exon 5. The VS antigen is a modified e antigen and is encoded by Leu245Val in the *RHCE*ce* allele (Faas et al., 1997). Some cases of the VS antigen observed hybrid alleles (Daniels et al., 1998). The same scenario for partial D, hybrid *CE-D-CE* alleles, has been observed which leads to weakened expressions and partial C, E, and e antigens may be generated, where some or all C/c and/or E/e antigens are missed due to the insertion of D epitopes (Noizat-Pirenne et al., 2002; Wagner and Flegel, 2004).

1.5.2.6 Rh_{null} phenotype

Rh_{null} disease or Rh deficiency syndrome is extremely rare which is characterised by a unique abnormality of a red cell morphological shape (stomato-spherocytosis). These individuals suffer from chronic haemolytic anaemia. This is due to the lack of an Rh complex (RhD, RhCE, RhAG), as well as the lack of intracellular adhesion molecule 4 (ICAM-4). Furthermore, expressing a reduction or lack of glycophorin B is considered to be the Rh_{null} phenotype. Furthermore, it was investigated that cells with an Rh

deficiency demonstrate a knockdown of CD47. Moreover, it possesses biochemical abnormalities including phospholipid asymmetry, abnormal water content, and cation co-transport. The genetic alterations to the *RHAG* gene are the chief cause of the Rh_{null} phenotype (Huang et al., 1998; Cherif-Zahar et al., 1998). The entire absence of the Rh complex in the red cell membrane is due to ablation or crucial alterations in the Rh multimer assembly (Avent et al., 2006).

1.6 Kell blood group system

The Kell blood group system is the sixth blood group system (006) according to the ISBT. It is considered as one of the most clinically relevant blood groups following the ABO and Rh systems. The first antibody of the Kell blood group system (anti-K) was discovered in 1946 following the introduction of the antiglobulin test. The system's name was derived from the first producer of that antibody, Mrs. Kelleher (Coombs et al., 1946). The system is highly polymorphic which comprises 35 antigens that are all expressed on the surface of the Kell glycoprotein (Reid et al., 2012).

1.6.1 Kell glycoprotein

The Kell glycoprotein is a type II single-pass membrane with a size of 93 kilodalton (kDa). It is expressed on the surface of erythrocytes at a level of 3500–17000 copies per red cell (Masouredis et al., 1980). Furthermore, this glycoprotein consists of 732 amino acids which are distributed as follows: 47 amino acids forming the short N-terminal intracellular domain, 20 amino acids within the single transmembrane domain and 665 amino acids forming the large extracellular domain where all the Kell antigens are expressed (Lee et al., 1991). **Figure 1.4** illustrates the structure of the Kell glycoprotein.

The extracellular region of the Kell glycoprotein has five N-glycosylation sites and contains 15 cysteine residues, which are believed to form intramolecular disulphide

bonds that give rise to the Kell glycoprotein's folded structure (Westhoff and Reid, 2004). The amino acid residues (550–732) in the extracellular domain seem to be conserved. The Kell glycoprotein was found to share 33–36% of the amino acid sequence homology with M13 or the neprilysin family of zinc endopeptidases (Russo et al., 1998). They appear to cleave the large inactive polypeptides into shorter proteolysis bioactive ones (Lee et al., 2003a). In the red cells, the Kell glycoprotein may cleave the big endothelin-3 into small bioactive endothelin-3, which acts as a potent vasoconstrictor and stanch haemorrhage (Lee et al., 1999). All of the Kell antigens, except for two antigens (Js^a and Js^b), are expressed prior to residue 550 of the extracellular domain (Westhoff and Reid, 2004).

The Kell glycoprotein is covalently attached to another protein in the red cell membrane, known as the XK protein, **(Figure 1.4)**, via a disulphide bond between cysteine 72 of the Kell glycoprotein to cysteine 347 of the XK protein (Khamlichi et al., 1995). The XK protein is a 37 kDa protein with 444 amino acids spanning the red cell membrane 10 times and forming five extracellular loops and is encoded by an X-linked gene (Cartron et al., 1998). The role of this protein remains mysterious, although it has been suggested that it may play a role as a membrane transport protein due its structure (Ho et al., 1994). A homologous protein, CED-8, is involved in apoptosis in *Caenorhabditis elegans* (Stanfield and Horvitz, 2000).

The existence of the XK protein on the red cells is highly crucial for the expression of the Kell glycoprotein on the red cell membrane, and its absence leads to McLeod syndrome, which results from a mutation in the *XK* gene that causes the absence of the Kx blood group antigen [XK1 or 019001] (Jung et al., 2007). This syndrome is characterised by reduced expression of the Kell antigens, *in vivo* reduction of RBCs survival, and acanthocytosis. In addition, males infected with this syndrome suffer from

neurological and muscular defects. On the other hand, the lack of the Kell glycoprotein expresses normal erythrocytes (Danek et al., 2001).

It was thought that the expression of the Kell glycoprotein is restricted to erythrocytes, but the Kell glycoprotein has also been found in bone marrow, the foetal liver, the tonsils, the spleen, and skeletal muscles (Russo et al., 2000). Furthermore, the transcription factors that are needed to activate Kell gene expression have been found in numerous tissues, with high levels in the brain, lymphoid tissues, and testes (Camara-Clayette et al., 2001).

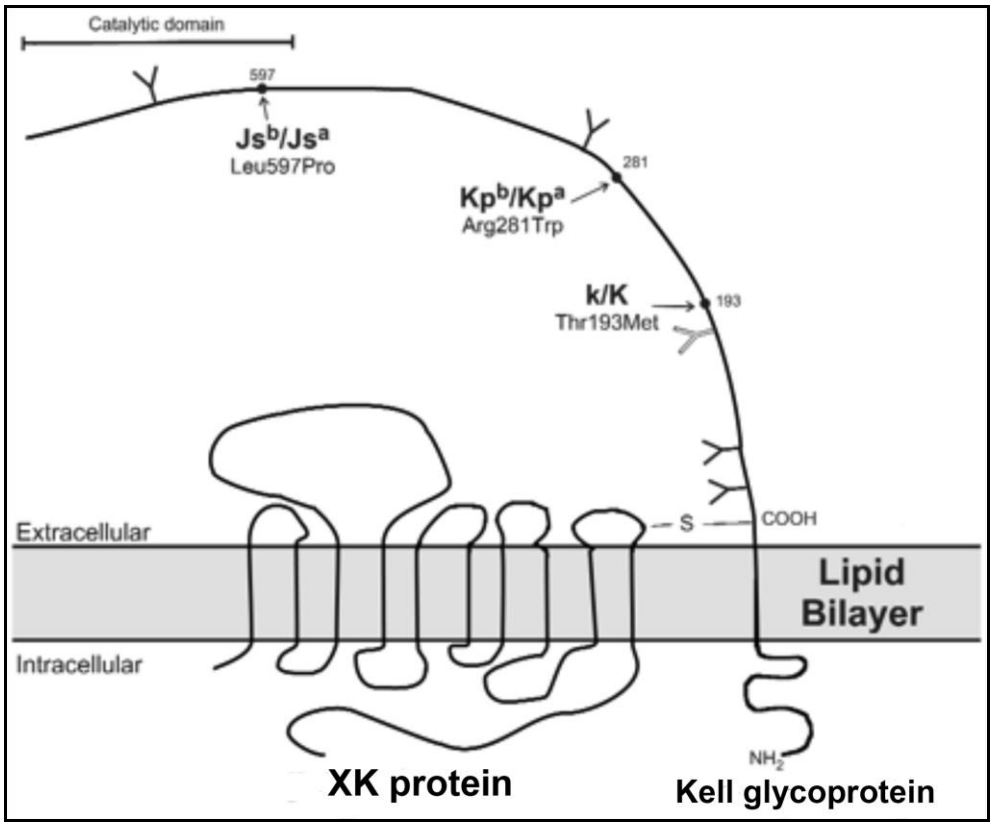


Figure 1.4 The structure of the Kell glycoprotein.

The Kell glycoprotein is covalently linked to another glycoprotein, which is called the XK protein. The linkage is a disulphide bond between cysteine 72 of the Kell glycoprotein and cysteine 347 of the XK protein. The Kell glycoprotein is a single-pass membrane type II glycoprotein of 93 kDa. The XK protein is a 37 kDa protein, which contains 444 amino acids, spans the red cell membrane 10 times and forms five extracellular loops. Adapted from (Westhoff and Reid, 2004).

1.6.2 Molecular basis of Kell system

A single gene (*KEL*), which is located on chromosome 7q33, encodes the Kell glycoprotein with 19 exons spanning about 21.3 Kilobases (Kb) of the human genomic DNA (Lee et al., 1993). All the Kell antigens are accompanied with a SNP except KEL20 (Lee, 1997). The KEL20 antigen requires the existence of the XK protein (Daniels, 2013a). Several genetic mechanisms may lead to the absence of all Kell antigens, which is known as K_{null} (K_0). These mechanisms include amino acid substitutions, premature stop codons, and a 5' splicing-site mutation in the coding and non-coding areas (Lee et al., 2001). Different missense mutations lead to the weak expression of Kell antigens, K_{mod}, which is a rarely inherited RBC phenotype that features weak expression of the high incidence antigens of the Kell system. However, in this case K_{mod} can be detected (Lee et al., 2003b).

1.6.2.1 Kell antigens

The antigens of the Kell blood group system are categorised in seven sets of allelic relevance which are (K, k), (Kp^a, Kp^b, Kp^c), (Js^a, Js^b), (KEL11, KEL17), (KEL14, KEL24), (KEL25, KEL28), (KEL31, KEL38). In addition, there are 17 high prevalence and 3 low prevalence antigens. **Table 1.2** lists the antigens of the Kell blood group system. In this thesis, the generic name is used for the common antigens including K, k, Kp^a, Kp^b, Kp^c, Ku, Js^a, Js^b, Ul^a and Km. However, the numerical antigen annotation proposed by the ISBT is used for the other antigens.

K antigen is the first antigen to be detected among the Kell blood group system. The prevalence of the K antigen among the English population is 9.02% (Daniels, 2013a). It varies from its antithetical antigen, k antigen, by an SNP. A nucleotide substitution 578C>T (previously reported as 698) on exon 6 gives rise to the K antigen, which encodes a different amino acid; Methionine rather than Threonine (Lee et al., 1995; Reid et al., 2012). Anti-K is implicated in HTR and severe HDFN. Disruption to the N-glycosylation occurs due to this amino acid substitution. This leads to the suggestion that the possible absence of N-oligosaccharides in this location may expose the K antigen's domain, which a sugar moiety normally covers. Consequently, this leads to the production of an antigenic site (Lee et al., 1995).

Kp^a and KEL17 antigens are mostly found in Caucasians. Js^a antigen is nearly exclusive in Blacks, with a frequency of 20% in African Americans. The Finns and the Japanese express the Ul^a antigen (Reid et al., 2012).

Table 1.2 List of the antigens of the Kell blood group system.

Number	ISBT Symbol	Generic Name	Obsolete name	Mutation	Exon	Amino acid	Prevalence	Antithetic
001	KEL1	K	Kell, K1	576C>T	6	Thr193Met	Low	1
002	KEL2	k	Cellano, K2	576T>C	6	Met193Thr	High	1
003	KEL3	Kp^a	Penny, K3	841C>T	8	Arg281Trp	Low	2
004	KEL4	Kp^b	Rautenberg, K4	841T>C, G at 842	8	Trp281Arg	High	2
005	KEL5	Ku	Peltz, K5	Complex, associated with K_{null}	-	-	High	-
006	KEL6	Js^a	Sutter, K6	1790T>C	17	Leu597Pro	Low	1
007	KEL7	Js^b	Matthews, K7	1790C>T	17	Pro597Leu	High	1
010	KEL10	Ul^a	Karhula, K10	1481A>T	13	Glu494Val	Low	-
011	KEL11	K11	Côte	905C>T	8	Ala302Val	High	1
012	KEL12	K12	Boc (Bockman, Spears	1643 G>A	15	Arg548His	High	-
013	KEL13	K13	SGRO	986C>T	9	Pro329Leu	High	-
014	KEL14	K14	San, Santini, Dp	539C>G	6	Pro180Arg	High	1
016	KEL16	K16	Weak k, k-like	Not defined	Not defined	Not defined	High	-
017	KEL17	K17 (Wk^a)	Weeks	905T>C	8	Val302Ala	Low	1
018	KEL18	K18	V.M., Marshall	388T>C	4	Trp130Arg	High	-
				389A>G		Gln130Arg		
019	KEL19	K19	Sub, Sublett	1475A>G	13	Gln492Arg	High	-
020	KEL20	Km	K20	Not defined	Not defined	Not defined	High	-
021	KEL21	Kp^c	Levay, K21	842G>A	8	Arg281Gln	Low	2
022	KEL22	K22	N.I., Ikar	965T>C	9	Val322Ala	High	-
023	KEL23	K23	Centauro	1145A>G	10	Gln382Arg	Low	-
024	KEL24	K24	CL, Callais, Cls	539 >C	6	Arg180Pro	Low	1
025	KEL25	VLAN	-	743G>A	8	Arg248Gln	Low	1
026	KEL26	TOU	-	1217A>G	11	Gln406Arg	High	-
027	KEL27	RAZ	-	745A>G	8	Lys249Glu	High	-
028	KEL28	VONG	-	742C>T	8	Arg248Trp	Low	1
029	KEL29	KALT	-	1868A>G	17	Lys632Arg	High	-
030	KEL30	KTIM	-	913A>G	8	Asn305Asp	High	-
031	KEL31	KYO		875G>A	8	Arg292Gln	Low	-
032	KEL32	KUCI	-	1271T>C	11	Val424Ala	High	-
033	KEL33	KANT	-	1283T>G	11	Leu428Arg	High	-
034	KEL34	KASH	-	758G>A	8	Cys253Tyr	High	-
035	KEL35	KELP	-	780T>G	8	Phe260Leu	High	-
				2024A>G	18	Gln675Arg		
036	KEL36	KETI	-	1391T>C	12	Ile464Thr	High	-
037	KEL37	KHUL	-	877C>T	8	Arg at 293	High	-
038	KEL38	KYOR		G at 875	8	Arg at 292	High	

The ISBT symbol, generic and obsolete names, mutation, and prevalence. Adapted from (Daniels, 2013a; International Society of Blood Transfusion, 2015).

1.7 Other blood groups

The antigens of other blood group systems were genotyped in this project **(Chapter 4)**. Most of the antigenic polymorphisms arise from SNPs. On the other hand, some antigens observe the deletion of a number of nucleotides such as the Vel blood group antigen. The nucleotides changing and amino acid substitutions are given in **Table 1.3.**

Table 1.3 Polymorphisms in other blood group systems.

System	Gene	Polymorphism	Nucleotide change	Amino acid
ABO	*ABO*	A/B	526C>G, 703G>A, 796C>A, 803G>C	Arg176Gly, Gly235Ser, Leu266Met, Gly268Ala
MNS	*GYPA*	M/N	59C>T, 71G>A, 72T>G	Ser1Leu, Gly5Glu
	GYPB	S/s	143C>T	Thr29Met
Duffy	*DARC*	Fy^a/Fy^b	125G>A	Gly48Asp
Kidd	*SLC14A1*	Jk^a/Jk^b	838G>A	Asp280Asp
Diego	*SLC4A1*	Di^a/Di^b	2561C > T	Pro854Leu
Dombrock	*ART4*	Do^a/Do^b	793G>A	Asp265Asn
Colton	*AQP1*	Co^a/Co^b	134C>T	Ala45Val
Yt	*ACHE*	Yt^a/Yt^b	1057C > A	His353Asp
Vel	*SMIM1*	Vel+/Vel−	Deletion of 17 nucleotides in exon 3	Frameshift deletion

1.8 Human platelet antigens (HPAs)

Platelets are blood cells that help to cease bleeding and form a blood clot. In order to avoid confusion with the previous names given to platelet antigens, the ISBT platelet working party adopted a numeric nomenclature that was performed in 1990 by ISBT and the International Committee for Standardisation in Haematology [ICSH] (von dem Borne and Décary, 1990). Following that period, the number of HPAs has been increased to 24 with more understanding of the molecular basis. Therefore, the nomenclature of HPAs was revised again in 2003 (Metcalfe et al., 2003).

To date, 33 HPAs have been defined, which are located on six functional glycoprotein complexes (GPIIb, GPIIIa, GPIbα, GPIbβ, GPIa and CD109). Twenty of these antigens are expressed on GPIIb and GPIIIa complex. Twelve antigens of which are grouped in a cluster of six biallelic groups. These include HPA-1, HPA-2, HPA-3, HPA-4, HPA-5 and HPA-15 (HPA Sequence Database, 2015). The HPAs are numbered, in which the high frequency platelet antigen is designated as "a" and the lower frequency form of the antigen is nominated as "b". All the HPAs are encoded by a SNP, apart from HPA-14, and these SNPs lead to amino acid substitutions (Lucas and Metcalfe, 2000). HPAs were marked as "w" for (workshop) when the antibody to one of the two antigens has been detected (Curtis and McFarland, 2014).

The antibodies to the HPAs can be involved in alloimmune platelet disorders, mainly foetal and neonatal alloimmune thrombocytopaenia [FNAIT] (Mueller-Eckhardt et al., 1989). Furthermore, it can cause multitransfusion platelet refractoriness (MPR) and posttransfusion purpura [PTP] (Novotny, 1999; Gonzalez and Pengetze, 2005). In the Caucasian population, the most common antibody to cause MPR, PTP and NAIT is anti-HPA-1a. This antigen is expressed in 98% of the Caucasian population (Newman et al., 1989). Moreover, this antibody was the reason for 85% of the incidences of FNAIT

cases, which was confirmed by serology (Curtis and McFarland, 2014). The second antibody to cause immune platelet disorders, which has more frequency next to the HPA-1a antibody, is the anti-HPA-5b antibody (Kiefel et al., 1989; Curtis and McFarland, 2014).

1.9 Typing of blood groups

1.9.1 Serological methods

The RBCs are not prone to agglutinate each other because they possess a negative charge due to the presence of sialic acid. The adjoining RBCs can become close to each other, causing a direct agglutination by bridging the gap by pentamer IgM. This is represented in compatibility typing for ABO, while IgG cannot accomplish this (Pamphilon and Scott, 2007). A direct antiglobulin test (DAT) is performed to investigate immune haemolytic anaemia and to observe whether RBCs are sensitised by antibodies. Antihuman globulin is incubated with the washed RBCs of the patients, which leads to an observable agglutination. This is also designated as the Coombs test due to its discovery by Robin Coombs (Coombs and Roberts, 1959).

Next, the IAT test is carried out in order to investigate a patient's serum with the donor's RBCs to provide a compatible blood unit. Following several washes, the addition of the antiglobulin reagent takes place, and the RBCs are examined to determine whether they have become sensitised. The antiglobulin test has been implemented in many applications, including haemolysis by alloimmunisation including HTR and HDFN as well as autoimmune haemolysis (Coombs and Roberts, 1959).

The typing of the blood groups by serology remains the gold standard approach for the compatibility procedure in order to make the blood transfusion feasible. It is a simple and rapid procedure and possesses an appropriate sensitivity to major transfusion

practices. However, the disadvantages of the serotyping include that it is labour-intensive and the high cost of the reagents, which lessen its feasibility to accomplish extended screening of blood groups for blood donors (Reid, 2009). Moreover, the cost of serotyping is higher in some cases such as multiply transfused patients. This is due to the extended screening of further blood groups in order to provide safer compatible blood units and reduce the risk of alloimmunisation (Mazonson et al., 2014). Furthermore, some of the serological reagents are limited, are weakly reactive or are not available for certain blood groups (Reid, 2009).

In fact, the RBCs of patients who have received a recent transfusion or mothers during pregnancy can be coated by IgG antibodies. In addition, the presence of a donor's RBCs in a patient's blood hampers the serotyping and prevents the identification of the right phenotype, in particular in cases of receiving a massive transfusion. Consequently, the existence of rare or weak expression of the antigens can complicate the serological typing. It may be difficult to differentiate between alloantibodies and autoantibodies in antigen-positive individuals (Reid et al., 2000; Reid, 2009). All the previous issues can be resolved by applying the blood group genotyping (BGG).

1.9.2 Genotyping of blood groups

Following the cloning of the blood group genes and the advent of the PCR, deoxyribonucleic acid (DNA)-based testing became more feasible (Mullis et al., 1987). Therefore, there are multiple approaches for BGG that are based on PCR. For example, the restriction fragment length polymorphism-PCR (RFLP-PCR) (Nishihara et al., 1994; Fukumori et al., 1995), sequence specific primers-PCR (SSP-PCR) and real-time PCR [RT-PCR] (Polin et al., 2008) capture many DNA targets using multiplex PCR (Tax et al., 2002; Beiboer et al., 2005) and microarray technology (Karpasitou et al., 2008).

The microarray technology possesses a high multiplexing capacity using a miniaturised solid surface. The solid surface can be either glass-arrays or bead-arrays. These surfaces are coupled to probes, which have numerous specific blood group alleles attached to the surface. During the assay, the labelled amplified PCR products are hybridised to the corresponding probes followed by measuring the fluorescent intensity in order to determine the blood group genotype.

There are several commercial kits of glass-arrays; human erythrocyte antigens (HEA) BeadChip™ array provided by Bioarray Solutions identifies 24 antigens (Hashmi et al., 2005; Hashmi et al., 2007) and BLOODchip®, which detects 47 variants and is produced by Progenika Biopharma S.A. The BLOODchip® was the outcome products from the Bloodgen project (Avent et al., 2007). An example of the bead-array is the ID Core+ kit provided by Progenika Biopharma S.A., which is associated with the Luminex system. The ID Core+ kit was used for BGG **(Chapter 3)**.

1.9.2.1 Applications of BGG

There are numerous benefits of using BGG over conventional serology. Daniels (2013) stated three chief reasons for the utilisation of the molecular approach using DNA. These include when the source of the RBCs is no longer available, the provision of more or better knowledge than the serotyping and the efficiency and cost effectiveness of the BGG in contrast to the serology (Daniels, 2013a).

The BGG has been applied in the determination of the foetal blood group by testing the isolated foetal DNA from the maternal plasma to prevent the risk of HDFN (Avent, 1998). Furthermore, extensive genotyping of blood donors is required in alloimmunised patients to preclude the formation of further alloantibodies and those causing alloimmunisation. BGG can be used in the investigation of the blood group of a recently transfused patient. Moreover, rare blood phenotypes can be found by screening the blood donors. BGG can determine the frequency of blood group polymorphisms in a specific population. Furthermore, the *RHD* zygosity of a father to a foetus at risk of HDFN can be assessed. BGG can distinguish the D variants properly and classify them properly and can conserve the RhD-negative units in stocks. Patients with autoimmune haemolytic anaemia can be typed easily using BGG (Anstee, 2009; Westhoff, 2006). The next section will explain the importance of genotyping for HPAs.

1.9.2.2 Applications of the genotyping of HPAs

The requirement of serological typing to have adequate numbers of fresh platelets, the scarcity of the serological reagents and the labour-intensive technique itself all hamper the typing by serology. In addition, it is only restricted to specific antigens, thus not all the HPAs can be tested (Curtis, 2008).

PCR paves the way for genotyping the blood groups and HPAs. Ribonucleic acid (RNA) was isolated with difficulty from platelets and reverse transcriptase PCR amplification then took place in order to determine the HPA-1a/1b allele. Following that complex procedure, Newman and colleagues accomplished the genotyping of the HPA-1a/1b allele using RFLP-PCR (Newman et al., 1989). After that, the SSP-PCR became the most widespread technique used for HPAs genotyping as it is a simple procedure (Curtis and McFarland, 2014). However, the drawback of this technique includes the post-analysis using gel documentation for genotyping (Hurd et al., 2002). Moreover, many alleles have been discovered and an automated system is required to facilitate the genotyping of different alleles of HPAs. Many assays have arisen such as melting curve analysis and RT-PCR (Curtis, 2008). High-throughput systems become available commercially in the form of array-based platforms including glass and bead arrays (Beiboer et al., 2005; Denomme and Van Oene, 2005).

The benefits of these platforms include: fast procedure that is less laborious, interpretation of the results is better using sophisticated computer software and fewer clerical errors can happen in data handling and during storage. The disadvantages of the high-throughput systems are the requirement of costly specific instruments with special reagents, as well as the expertise needed for operation.

The disadvantage of all the genotyping platforms mentioned above is that they cannot identify new alleles. Therefore, they require improving and updating to the current assay when new alleles have been discovered. Therefore, investigators of BGG move to use NGS to obtain the high-throughput sequencing information as well as the ability to identify the unprecedented alleles (Avent et al., 2015).

Sanger sequencing has frequently been used to resolve the complexity of the blood group variants, in particular the new ones. Nowadays, NGS has been used recently in

BGG. For example, Stabentheiner et al. (2011) used NGS to genotype the 10 exons of the *RHD* gene. Moreover, a panel was designed to capture the exons of the genes for 18 blood groups for sequencing (Fichou et al., 2014). Avent et al. (2015) utilised NGS through amplification of the entire genes of the blood groups by LR-PCR followed by NGS. Moreover, NGS was used in prenatal diagnosis and determination of the foetal genotyping from maternal plasma (Rieneck et al., 2013; Wienzek-Lischka et al., 2015). In order to explain the NGS principle, the knowledge regarding Sanger sequencing is discussed first and NGS is discussed in the next sections.

1.10 Conventional sequencing 'Sanger sequencing'

Sanger sequencing is also called the chain-termination method. The sequencing template must be denatured and obtained in a single stranded form. Furthermore, the primer must be complementary to the 3' strand of the template, thus the DNA polymerase can extend the primer and generate the sequencing by synthesising a nascent strand of the template.

Equal amounts of this reaction mixture are placed into four tubes, and four dideoxynucleotides triphosphates (ddNTPs) are added to each tube. The replication is initiated by adding deoxynucleotide triphosphates (dNTPs), however, the addition of ddNTPs will terminate the sequence synthesis (Sanger et al., 1977).

The contents of the reaction tubes are then subjected to four lanes of polyacrylamide electrophoresis gel and the oligonucleotide sequences are separated according to the size and type of nucleotides. The shortest sequences are moved further down to the bottom of the gel and the reading can be achieved from the bottom to the top. **Figure 1.5** illustrates the principle of Sanger sequencing while running on polyacrylamide electrophoresis gel.

Following that period, the method was evolved to include fluorescent labelling (Prober et al., 1987). Moreover, the automation of 96 or 384 independent capillaries was designed to increase the throughput and facilitate sequencing of more samples (Shendure and Ji, 2008). Furthermore, the read length of sequencing can be achieved by Sanger sequencing up to around 1000 bp and a high level of accuracy of 99.999% can be obtained. Sanger sequencing is still the gold standard of DNA sequencing although the cost of it has been valued at $0.50 per Kb (Shendure and Ji, 2008). Interestingly, Sanger sequencing was used in sequencing the entire human genome in the Human Genome Project [HGP] (Bentley, 2000). The sequencing of five individuals, around 14.8 billion bp, took over nine months (Venter et al., 2001). Therefore, the requirement of new platforms to provide rapid sequencing at an affordable price was needed in order to obtain the benefits of the sequencing.

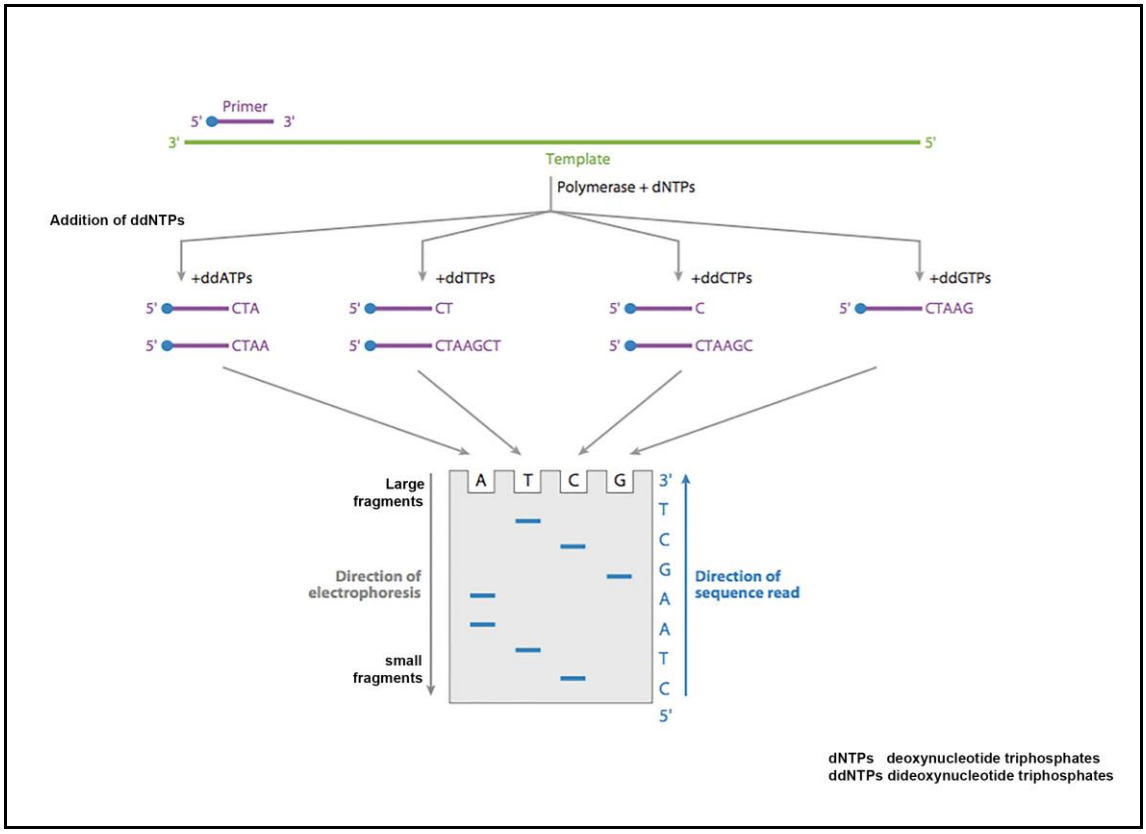


Figure 1.5 Sanger sequencing.

A complementary primer anneals to a 3' single-stranded template to start the extension of the template by adding dNTPs using polymerase enzyme. Equal amounts of the reaction mixture are placed into four tubes, and four ddNTPs are added to each tube, which terminate the sequence synthesis. The contents of the reaction tubes are then subjected to four lanes of polyacrylamide electrophoresis gel and the oligonucleotide sequences are separated according to the size and type of nucleotides. The smallest sequences are moved further down to the bottom of the gel, while the biggest sequences are remained at the top of the gel. The reading can be carried out from the bottom to the top as depicted with the blue arrow. Adapted from (Mardis, 2013).

1.11 Next generation DNA sequencing

Following the accomplishment of the HGP, the requirements of the rapid production of DNA sequencing at low cost were to provide a high-throughput which increases the number of assessed samples. These led to the advent of the technology of NGS, which can be acquired by personal investigators including small laboratories (Shendure and Ji, 2008). NGS is also denoted as massively parallel sequencing, deep sequencing or total genome sequencing (Hert et al., 2008).

In contrast to Sanger sequencing, NGS uses a different principle in its mode of action. In all NGS platforms, following the extraction of the genetic materials, a sequencing library is prepared. Initially, the sequencing library is constructed by shearing the targeted DNA into small fragments, either by using enzymes or physical sonication. A ligation process then takes place by attaching both ends of the fragmented DNA to adaptors, which are pairs of signal sequence oligonucleotides. The fragmented DNA along with the adaptors forms the sequencing library (paired end library). The sequencing library is then fixed to bead particles or solid surfaces, depending on the platform used, in order to prepare the sequencing template for clonal amplification. The clonal amplification either uses emulsion PCR or solid phase amplification depending on the platform used, which is performed in order to enrich and obtain a high number of sequencing templates to ascertain appropriate signal production in the sequencing reaction that is sufficient for a sensitive detection.

The amplification procedure involves high consideration because a low fidelity polymerase may generate errors by incorporating mismatched nucleotides. Accordingly, high rates of sequencing reads may be excluded and successively lead to a low depth of coverage especially in homozygous samples. A depth of coverage can be defined as numbers of times that the sequencing reads are sequenced repeatedly. Moreover, allelic

dropouts can be occurred due to PCR primer optimisation or possibility of presence of any new alleles in the primer-binding site. In addition, low quality data can be obtained and need to be trimmed if do not meet the quality standards (Gabriel et al., 2011). See [section 2.5.5] for further information regarding the data analysis of NGS.

However, some sequencers are known as third generation sequencing and they do not need any amplification involved in the process of constructing the sequencing library. These are also designated as single molecule sequencing which includes platforms such as Pacific Biosystems and Oxford Nanopore (Mardis, 2008).

1.11.1 Platforms of NGS

There are a number of platforms offered nowadays, and each one possesses a different capacity to generate read length, coverage, speed, and accuracy. **Table 1.4** demonstrates different types of NGS platforms. The first commercially available platform of the NGS was the Roche/454 FLX. This platform is also referred to as pyrosequencing, specifically due its performance of using pyrophosphate-based sequencing (Margulies et al., 2005). The sequencing libraries are constructed by fragmenting the targeted template and ligating adaptors at both ends. Then the clonal amplification takes place, in which a single sequencing library is attached to a bead within droplets of PCR reaction mixture in oil emulsion. This results in each bead being covered by ten million copies of the targeted template. Following the clonal amplification procedure, the emulsion is broken and denaturation of the DNA template occurs in order to obtain a single stranded template. The beads that carry single stranded DNA are deposited into the wells of a fibre-optic slide. The sequencing is based on synthesising a new strand in which the pyrophosphate is released and a photon is generated (Margulies et al., 2005). Basically, most NGS platforms generate a nascent DNA strand during the sequencing. This is

actually called sequencing by synthesis. The next section will discuss the NGS platform used in this PhD.

Table 1.4 NGS platforms.

NGS Platform	Clonal amplification	Sequencing chemistry	Average read length
454	Emulsion PCR	Pyrosequencing (sequencing by synthesis)	Up to 1000 bp
Illumina	Bridge amplification	Reverse dye terminator (sequencing by synthesis)	50-300 bp
Solid	Emulsion PCR	Oligonucleotide 8-mer chained ligation (sequencing by ligation)	75 bp
Ion Torrent	Emulsion PCR	Proton detection (sequencing by synthesis)	100-400 bp
PacBio	N/A	Phospholinked Fluorescent nucleotides (sequencing by synthesis)	8500 bp

Adapted from (Hodkinson and Grice, 2015).

1.11.1.1 Ion Torrent Personal Genome Machine™ (Ion PGM™)

Ion PGM™ was developed by Ion Torrent (Guildford, CT, United States), and is currently owned by Thermo Fisher Scientific (Carlsbad, CA, United States). It is a benchtop semiconductor sequencer. Ion Torrent entered the market in 2010 as a simple sequencer that was affordable and fast due to its running time from two hours to seven hours. It utilises a combination of semi-conductor chip technology and simple chemistry. The sequencing technology of the Ion PGM™ is based on sequencing by synthesis. In other words, every time the DNA polymerase incorporates a nucleotide to generate a new strand of DNA, a hydrogen ion (H^+) is released. Therefore, it does not depend on intermediary light in comparison with other platforms. This hydrogen ion alters the pH of the surrounding solution (0.02 pH units) for every single base that has been incorporated. The detection then takes place using a sensor underneath each well, working as a solid-state pH meter. Off-chip electronics finally digitise the voltage to digital information and the base calling. The detection procedure takes around 4 seconds per base. Following a flow of every nucleotide, a wash step is utilised to ensure the well is free from any nucleotides (Rothberg et al., 2011). **Figure 1.6** demonstrates the sequencing reaction of the Ion Torrent™ platforms.

Three sizes of chip are provided for Ion PGM™, (Ion 314™ Chip, Ion 316™ and Ion 318™ Chip) in order to give each laboratory the scalability according to the assay they are performing. Four read lengths are provided for the Ion PGM™ platform (100 bp, 200 bp, 300 bp and 400 bp). The output of this sequencer ranges from 30 Mb to 2 Gb depending on the type of chip used as well as the reading length considered for the sequencing (Thermo Fisher Scientific, 2015a). The total reads can be produced by the Ion PGM™ which ranges from 400 thousand to 5.5 million reads. Ion PGM™ is ideal for

small projects and small laboratories in which it is able to sequence genes, panels of genes, small genomes, and profile gene expressions.

In January 2012, a new platform (the Ion Proton™) was launched which is even more powerful in producing sequencing data than the Ion PGM™. It has the capability to carry out whole genome sequencing (WGS) and whole exome sequencing [WES] (exons of the human genome) within a range of two to four hours. The read length available to this platform is only 100 bp and 200 bp. The Ion Proton™ platform provides an extremely high output which reaches up to 10 Gb and 60 to 80 million reads can be generated by such a platform (Thermo Fisher Scientific, 2014a).

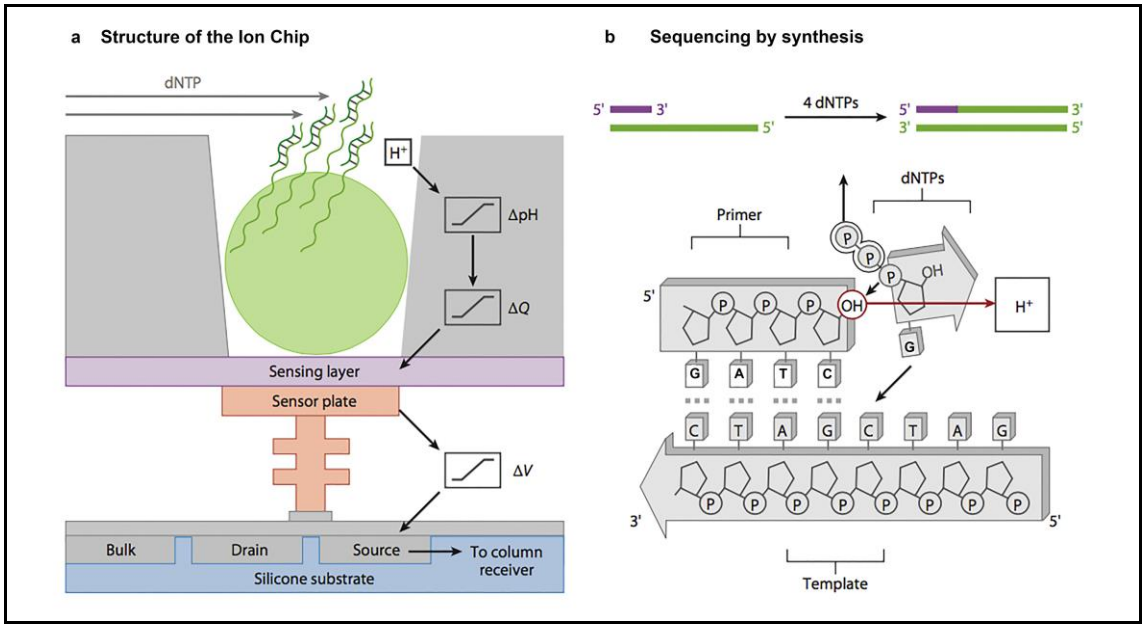


Figure 1.6 The sequencing reaction on the Ion PGM™.

(a) A schematic structure of the Ion Chip demonstrates the addition of nucleotide causing release of the hydrogen ion H+. This leads to changing of pH of the surrounding solution and the detection then occurs using a sensor underneath each well, working as a solid-state pH meter. Off-chip electronics finally digitise the voltage to digital information and the base calling. (b) The sequencing by synthesis is based on pH sensing, in which as every nucleotide is incorporated, a hydrogen ion is released and the base is called. Adapted from (Mardis, 2013).

1.12 Thesis aims and objectives

BGG provides more comprehensive blood typing in comparison to the conventional serology. Indeed, it shows a high impact in reducing the risk of alloimmunisation to multiply transfused patients, such as patients with SCD. Microarray technology (bead-arrays) was utilised in order to genotype blood group alleles **(Chapter 3)**. This was as training for different approaches of genotyping assays.

NGS has been utilised in many research aspects for genotyping. The aim of this PhD project was to assess the genotyping by sequencing based on NGS using the Ion PGM™ platform. In this study, two approaches were designed using NGS for BGG.

The first approach targeted and selected amplicons of interest using the Ion Ampliseq™ Custom Panel, which is designated as the Human Erythrocyte Antigens and Human Platelet Antigens Panel (HEA and HPA Panel). The panel assay includes 11 blood group systems and 16 human HPAs **(Chapter 4)**. This project comprises the following objectives:

- To develop a comprehensive assay in order to genotype both the alleles encoding the blood group antigens and HPAs in a single sequencing run.
- To establish a rapid protocol for genotyping with a lower cost in contrast to the previous BGG platforms.
- To identify novel alleles which have not been reported previously.

The second approach was a LR-PCR followed by the NGS for the Kell **(Chapter 5)** and Rh blood group systems **(Chapter 6)**. This type of sequencing develops a better understanding about the various genetic mechanisms that may be involved in the blood group genes because of sequencing the entire gene of interest including the coding and non-coding areas.

This project contains the following objectives:

Kell blood group

- To develop a high resolution assay for genotyping the *KEL* gene.
- To predict the main antigens of the Kell blood group system, in particular the high prevalence ones.
- To identify the *KEL* alleles encoding the Kell antigens that cannot be identified by the standard serology.
- To utilise the Kell blood group system as a model to establish the same protocol for the Rh blood group system.

Rh blood group

- To develop a high resolution assay for genotyping the *RHD* and *RHCE* genes.
- To distinguish between RhD-positive, weak D, and partial D samples.
- To investigate all the SNPs which are related to a particular allele.
- To determine the specific allele from the intronic SNPs and predict its associated phenotype.
- To identify rare alleles that may be common in different ethnicity.
- To differentiate if an allele is wrongly genotyped to a different gene, by other BGG platforms, in case of non-specific primers due to the high homology between *RHD* and *RHCE* genes.

Chapter 2 : Materials and Methods

2.1 Sourcing of blood samples

Blood samples were provided by the National Health Service Blood and Transplant (NHSBT) Filton from anonymous volunteer blood donors with informed ethical donor consent. The random blood samples were received in ethylenediaminetetraacetate (EDTA) anti-coagulated tubes and serological testing was accomplished for all samples by the same institute. The phenotyping was for the following blood groups: ABO, Rh, MNS, P1, Lutheren, Kell, Lewis, Duffy, Kidd and Sickle cell status.

Twenty-eight samples were chosen randomly and genotyped by microarray using the ID Core+ assay [Cohort A, Chapter 3]. Another 28 samples, Cohort B, were analysed for red cell antigens and HPAs using HEA and HPA Panel based on NGS [Chapter 4]. Regarding Cohort C, 20 samples were chosen known serology and used for Kell LR-PCR followed by NGS, 18 kk and 2 Kk [Chapter 5]. Regarding cohort D, 10 samples were chosen with known serology for Rh LR-PCR followed by NGS, in which five RhD-positive and five weak D samples. A further investigation was carried out for a single sample of the HEA and HPA Panel samples by Rh LR-PCR followed by NGS. This sample observed a novel allele in the *RHCE* gene detected by the HEA and HPA Panel [Chapter 4]. Figure 2.1 shows a diagram for the numbers of blood samples were used for different projects.

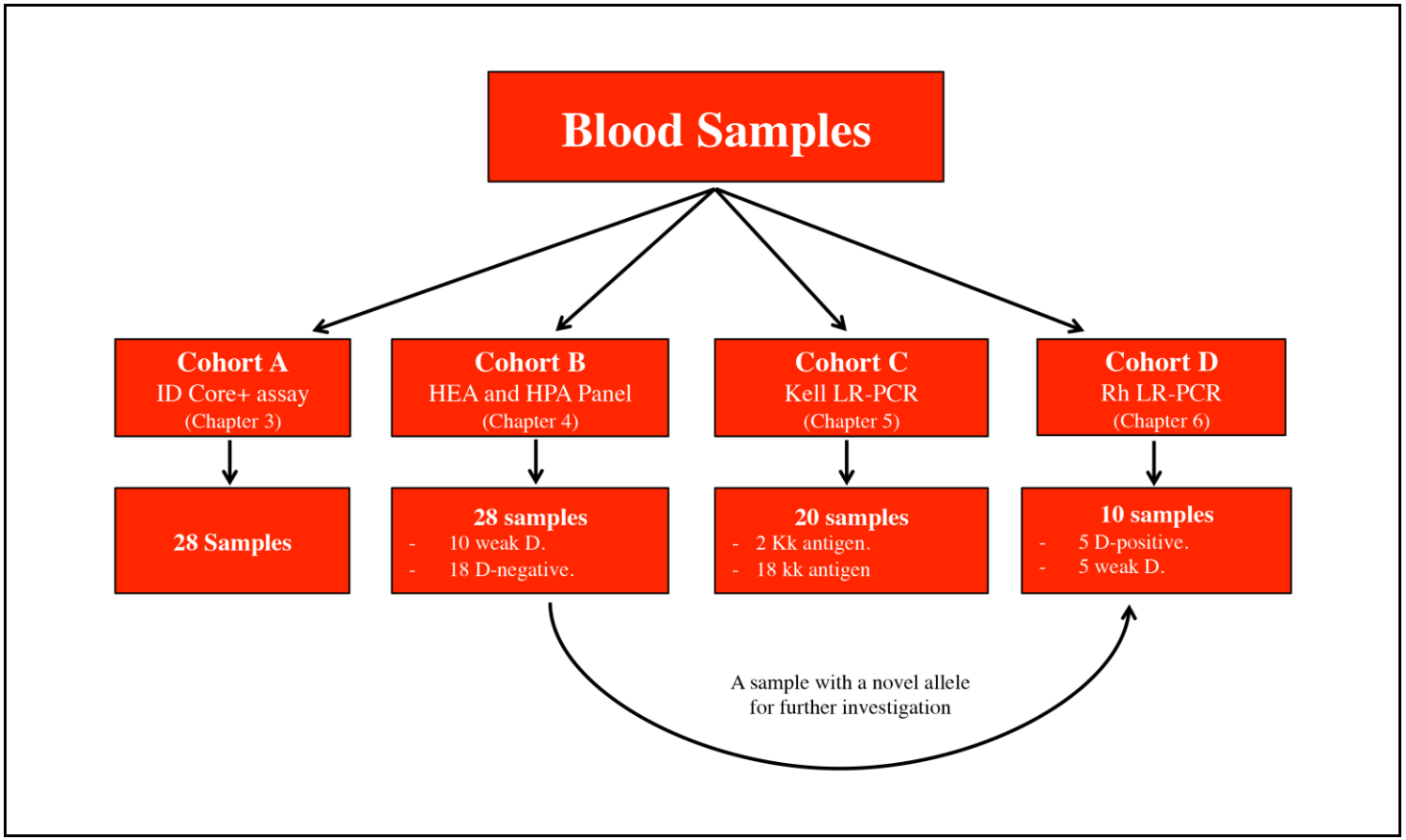


Figure 2.1 The different blood samples were used in this PhD project.

2.2 Genomic DNA extraction

2.2.1 Routine genomic DNA extraction

To extract the genomic DNA, buffy coats need to be obtained from the sample of the whole blood. In order to obtain buffy coats, the blood samples were spun at 2500 × g for 10 minutes in order to separate the components of the whole blood into three parts, which are plasma, a layer of buffy coat, and concentrated red cells. The buffy coat layer was collected and the other two components were discarded. Genomic DNA was purified using the QIAamp DNA Blood Mini Kit (Qiagen Ltd, West Sussex, United Kingdom). A volume of 20 µl of Qiagen Proteinase K was added to 200 µl of the extracted buffy coat in a clean 1.5 mL Eppendorf tube. Then, a volume of 200 µl of AL buffer was added to the mixture. The mixture was mixed vigorously and incubated at 56 °C for 10 minutes. Next, a volume of 200 µl of absolute ethanol (Fisher Scientific UK, Leicestershire, United Kingdom) was added to the mixture and was mixed for 15 seconds. After that, the mixture was applied to the QIAamp mini spin column, which had a collection tube, and the tube was spun at 6000 × g for 1 minute. The QIAamp mini spin column tube was placed into a new clean collection tube and the previous collection tube was discarded which contained the filtrate. Next, a volume of 500 µl of AW2 was added to the QIAamp mini spin column tube and was spun at 20,000 × g for 3 minutes. Then, the QIAamp mini spin column was placed in a new 1.5 mL Eppendorf tube and the tube which contained the filtrate was discarded. A volume of 200 µl of AE buffer was added to the QIAamp mini spin column and the tube was incubated at room temperature for 1 minute. Finally, the tube was spun at 6000 × g for 1 minute and the purified DNA was obtained and stored at −20°C for long-term use.

2.2.2 Genomic DNA extraction for long amplicons

Genomic DNA was extracted from blood samples and purified for very long genomic DNA in order to amplify a PCR product bigger than 20,000 bp. Gentra® Puregene® Blood kit from Qiagen Ltd (West Sussex, United Kingdom) was utilised in this experiment. The buffy coat layer was obtained as described above [section 2.2.1]. Because the buffy coat contained red cells, 3 volumes of RBCs Lysis Solution (750 µl) were added to a 15 mL centrifuge tube with 250 µl of buffy coat. The tube was inverted in order to mix it properly and was incubated at room temperature for 10 minutes. After the incubation period, the tube was spun at 2000 × *g* for 5 minutes. Then, the supernatant was discarded by pouring and leaving around 200 µl of the solution with the pellet. After that, the tube was mixed vigorously in order to resuspend the pellet in the remaining solution. A volume of 3 mL of Cell Lysis Solution was added to the tube and was mixed thoroughly to lyse the cells. Next, a volume of 15 µl of RNase A Solution was added to the tube and the solution was mixed by inverting the tubes 25 times. Following the proper mixing, the tube was incubated at 37°C for 15 minutes. After that, the tube was incubated on ice for 3 minutes for fast cooling of the sample. Then, a volume of 1 mL Protein Precipitation Solution was added to the solution and was mixed at high speed for 20 seconds followed by spinning at 2000 × *g* for 5 minutes. A volume of 3 mL of isopropanol was dispersed into a new 15 mL centrifuge tube and the supernatant was poured cautiously from the previous step. The tube was mixed thoroughly by inverting the tube 50 times followed by spinning at 2000 × *g* for 3 minutes. The supernatant was then discarded and the tube was drained by inverting it on absorbent paper. It must be ensured that the pellet is still in the tube. After that, a volume of 3 mL of 70% ethanol was added and was mixed properly by inverting the tube many times in order to wash the DNA pellet. Then the tube was spun at 2000 × *g* for 1 minute. After washing the pellet with ethanol, the tube was drained again by

inverting the tube on absorbent paper for up to 10 minutes. After that, a volume of 300 µl of DNA Hydration Solution was added to the tube and mixed vigorously for 5 seconds followed by incubation at 65°C for 1 hour in order to dissolve the DNA. Next, the tube was incubated at room temperature overnight with gentle shaking in order to mix the genomic DNA with the DNA hydration Solution. Finally, the pure genomic DNA was transferred into a new 1.5 mL Eppendorf tube and kept at −20°C for long-term storage.

2.2 Assessment of DNA quality and quantity

DNA concentration was obtained using Qubit® dsDNA BR Assay kit (Thermo Fisher Scientific, Paisley, United Kingdom). A Qubit™ Working solution was prepared by diluting the Qubit™ reagent in a ratio of 1:200 with Qubit™ buffer. Two different standards (standard 1 and standard 2) were prepared by adding 190 µl of the Qubit™ Working solution with 10 µl of standard 1 and standard 2, respectively. For each sample, a volume of 199 µl of Qubit™ Working solution was mixed with 1 µl of the DNA sample to obtain a total volume of 200 µl. Next, the assay tubes were mixed vigorously for 2 seconds followed by incubation at room temperature for 2 minutes. Finally, the samples were read on the Qubit™ Fluorometer instrument and readings of ng/µl were obtained. Samples ranged in concentrations between 90-250 ng/µl (Thermo Fisher Scientific, Paisley, United Kingdom).

DNA purity was assessed by utilising a NanoVue Plus Spectrophotometer (GE Healthcare, Little Chalfont, United Kingdom). The optical density (OD) A260/A280 ratio was used to assess DNA purity, which was in the range of 1.7 to 1.95. The DNA samples were stored at −20°C for long-term storage. The results of the DNA concentrations and purity are demonstrated in Appendix B.

2.3 Genotyping using microarray beads

2.3.1 Multiplex PCR

The ID Core+ kit was utilised which is based on microarray suspension beads. All the reagents of the ID Core+ kit were provided by Progenika Biopharma S.A. (Derio, Spain; distributed by Grifols). Two reactions of multiplex PCR were prepared. The first reaction (A) could identify the antigens for RhCE, Kell, Kidd, Duffy, and MNS blood group systems. The second reaction (B) was used for the antigens of Diego, Dombrock, Colton, and Yt blood group systems. The two tubes comprised IDCore A/IDCore B PCR mixes with Master ID PCR Mix, Biotin deoxycytidine triphosphate (dCTP), IDCore Primer Set A, and IDCore Primer Set B, respectively. Both mixes (IDCore A PCR Mix and IDCore B PCR Mix) were gently mixed and spun down. A volume of 10 µl of these mixtures was added into each well of a 96-well PCR plate for column A and column B. A volume of 2.5 µl of genomic DNA was added with 50 ng/µl, and the plate was then sealed with a sealing film. Next, the plate was placed in a centrifuge plate to be spun until it reached 900 rpm. After that, the 96-well PCR plate was covered by a MicroAmp compression pad before starting the run on a Veriti thermocycler. The following programme was conducted: the polymerase was activated at 95°C for 15 minutes, followed by 40 cycles at 95°C for 30 seconds for denaturation, 60°C for 30 seconds for annealing, and 72°C for 80 seconds for extension. Finally, the final extension was 72°C for 7 minutes, and was held at 4°C.

2.3.2 Hybridisation

Hybridisation mixtures were prepared according to the hybridisation step for both column A and column B, respectively. The mixtures were made by using IDCore Beads A/B, a hybridisation ID buffer, and a control ID buffer. The mixtures were spun for 30 seconds prior to use. A volume of 95 µl of the IDCore+A/B hybridisation mix was added to a 96-Thermowell Polycarbonate PCR Microplate (Corning, Flintshire, United

Kingdom). Then, a volume of 5 µl of the PCR product was added to the hybridisation mixture. The mixture then needed to be mixed thoroughly 20 times by pipetting up and down. After that, the plate was then sealed with the appropriate film (Microseal 'A' Film, BioRad, United Kingdom). After placing the plate on the thermocycler, the plate was covered by two silicone pads. The hybridisation programme was set to 95°C for 5 minutes for denaturation and to 52°C for 15 minutes for hybridisation, and was finally held at 52°C.

2.3.3 Labelling

A Millipore filter plate (MultiScreen$_{HTS}$ Filter Plates Millipore, Merck Millipore, United Kingdom) was pre-moistened with 50 µl of a conjugate of a streptavidin and phycoerythrin (SAPE) dilution buffer. Vacuum pressure was applied at −10 kPa 3 times at 30 second intervals. A labelling mix was prepared with SAPE and SAPE dilution buffer and kept in the dark. A volume of 95 µl of the hybridisation mix was transferred to the Millipore filter plate. The dispensing occurred against the wall to prevent any contact that may happen between the pipette tips and the membrane filter. After that, vacuum pressure was applied again to the Millipore filter plate with the same conditions as previously described. A volume of 75 µl of the labelling mix was added to each well. Finally, the Millipore filter plate was wrapped with aluminium foil to protect the reaction from the light and placed into a plate shaker at 1000 rpm for 15 minutes.

2.3.4 Quantification

The Millipore plate was instantly analysed with the Luminex® 100/200™ platform for quantification. The results were analysed using Luminex xPonent® LX100/LX200 Version 3.1. This software worked in association with BLOODchip® ID Software (BIDS) [Workstation Progenika Version 2.0 Beat 6] to produce a genotyping report of every sample and to predict the phenotypes.

2.5 NGS libraries construction

2.5.1 NGS libraries construction using a panel of short amplicons

The HEA and HPA Panel were designed using Ion Ampliseq™ Designer (Ion Ampliseq Designer, 2014). The design included 11 blood group systems (ABO, Rh, Kell, Kidd, Duffy, MNS, Diego, Dombrock, Yt, Colton and Vel) and targeted HPAs 1-16. Two pools of ultra-high multiplex primers were provided for all the targets. The starting material of DNA was diluted to 10 ng per pool and was assessed using a Qubit® 2.0 Fluorometer using the Qubit® dsDNA BR Assay kit. Therefore, 20 ng of each DNA sample was used to amplify all the targets. The total number of amplicons was 167, and the first pool comprised 85 amplicons, while the second one contained 82 amplicons. The panel size was 31.18 Kb and the coverage was 99.33%. Table 2.1 demonstrates the coverage summary of the design including the target name, chromosome location, start position, end position, number of amplicons used to amplify the required target, total bases, covered bases by the design and the missed bases.

Table 2.1 Coverage summary of the HEA and HPA Panel.

The panel was designed by Ion Ampliseq™ Designer (Ion Ampliseq Designer, 2014). The table shows the target name, chromosome location, start and end position, number of amplicons used to amplify the whole target, total bases, covered bases and missed bases in the design. The fields regarding the starting and end positions for the genes are blank due to the design covers all the exons within these genes.

Target name	Chromosome	Starting position	End position	Number of amplicons	Total bases	Covered bases	Missed bases
ABO	chr9	.	.	11	1064	1034	30
GYPA	chr4	.	.	7	453	453	0
GYPB	chr4	.	.	5	276	276	0
GYPE	chr4	.	.	3	237	237	0
RHD	chr1	.	.	11	1254	1254	0
RHCE	chr1	.	.	11	1300	1300	0
KEL	chr7	.	.	21	2199	2136	63
DARC	chr1	.	.	6	2028	2028	0
SLC14A1	chr18	.	.	9	1489	1489	0
SLC4A1	chr17	.	.	24	2736	2691	45
ART4	chr12	.	.	6	945	945	0
ACHE	chr7	.	.	15	1976	1976	0
AQP1	chr7	.	.	10	1215	1215	0
SMIM1 exon 1	chr1	3689340	3689430	1	90	86	4
SMIM1 exon 2	chr1	3689610	3689770	1	160	160	0
SMIM1 exon 3	chr1	3691880	3692100	2	220	220	0
SMIM1 exon 4	chr1	3692330	3692570	2	240	240	0
HPA-1	chr17	45360589	45360773	1	184	184	0
HPA-2	chr17	4836321	4836433	1	112	112	0
HPA-3	chr17	42452988	42453128	0	140	140	0
HPA-4	chr17	45361875	45361995	0	120	120	0
HPA-5	chr5	52358655	52358844	2	189	189	0
HPA-6	chr17	45369669	45369851	0	182	182	0
HPA-7	chr17	45369500	45369700	2	200	200	0
HPA-8	chr17	45377730	45377950	2	220	220	0
HPA-9	chr17	42452930	42453120	2	190	190	0
HPA-10	chr17	45360780	45361020	2	240	240	0
HPA-11	chr17	45376680	45376950	2	270	270	0
HPA-12	chr22	19711170	19711530	2	360	360	0
HPA-13	chr5	52368960	52369240	2	280	280	0
HPA-14	chr17	45376680	45376950	2	270	270	0
HPA-15	chr6	74493400	74493600	2	200	200	0
HPA-16	chr17	45361700	45361990	2	290	290	0

2.5.1.1 Amplification of the HEA and HPA Panel

Each sample was amplified in two different pools using ultra-high multiplex PCR using the Ion Ampliseq™ Library Kit (Thermo Fisher Scientific, Paisley, United Kingdom). The reaction mix contained 5X Ion AmpliSeq™ HiFi Mix, 2X Ion AmpliSeq™ Primer Pool, genomic DNA and nuclease-free water. The thermocycler was set as follows: 99°C for 2 minutes to activate the enzyme, 19 cycles of denaturation at 99°C for 15 seconds, annealing and extension at 60°C for 4 minutes and the reaction was then held at 10°C. The primers were then partially digested by adding 2 µL of FuPa reagent and the thermocycler was set to the following programme: 50°C for 10 min, 55°C for 10 min, 60°C for 20 min and then held at 10°C for up to 1 hour.

2.5.1.2 Adapters ligation of HEA and HPA Panel

For every barcoded adapter, a final dilution of (1:4) was prepared by mixing Ion P1 adapters and Ion Xpress™ barcode (X). Then, 4 µL of Switch solution was mixed with 2 µL of the diluted adapters to obtain a total volume of 28 µL. After that, the thermocycler was set to the following programme: 22°C for 30 minutes followed by 72°C for 10 minutes and held at 10°C for up to 1 hour.

2.5.1.3 Purification of HEA and HPA Panel

A purification step then took place using the Agencourt® AMPure® XP reagent by adding 45 µL of the beads to each sample (1.5X of the original sample). The mixture was incubated for 5 minutes at room temperature and then placed in a magnetic rack for 2 minutes. The required sequencing libraries were in the pellet, therefore the supernatant was discarded without touching the pellet. Two ethanol washes were performed by adding 150 µL of 70% ethanol and then the pellet was left to air-dry for up to 5 minutes and finally the sample was eluted using 50 µL of low Tris-EDTA. Following this procedure, quantitation of the Ampliseq™ sequencing libraries were

quantified by Agilent® 2100 Bioanalyzer (Agilent Technologies LDA UK Limited, Stockport, United Kingdom) as explained in the next section.

2.5.1.4 Quantification of the sequencing libraries using qPCR

This experiment was performed by Dr. Michele Kiernan at the Systems Biology Centre at Plymouth University. By using the Ion Library Quantitation Kit, a 100-fold dilution was carried out by mixing 5 µL of the sample with 495 µL of nuclease-free water. Each standard, negative control and the samples were analysed in duplicate of 20 µL reaction. Typically, the unamplified libraries have yields between 100 and 500 pM. 10-fold serial dilutions of the *E. coli* DH10B Ion Control Library were prepared, which is around ~68 pM from the Ion Library Quantitation Kit, at 6.8 pM, 0.68 pM and 0.068 pM.

Three 10-fold serial dilutions were prepared of the *E. coli* DH10B Ion Control Library (~68 pM; from the Ion Library Quantitation Kit) at 6.8 pM, 0.68 pM, and 0.068 pM. Reaction mixtures were prepared for each sample, control, and standard by combining 20 µL of 2X TaqMan® MasterMix with 2 µL of 20X Ion TaqMan® Assay and we then dispensed 11-µL aliquots into the wells of a PCR plate. 9 µL of the diluted (1:100) Ion AmpliSeq™ library or 9 µL of each control was diluted in each well as duplicates, for a total reaction volume of 20 µL.

The PCR condition was set up at 50 °C for 2 minutes, 95 °C for 20 seconds followed by 40 cycles of 95 °C for one second and 60 °C for 20 seconds. Following the PCR, the average concentration of the diluted libraries was multiplied by the dilution factor (100). Following quantitation the sequencing libraries were diluted to 100 pM as recommended in the protocol to ensure they were suitable for template preparation.

2.5.2 Next-generation sequencing for long amplicons

2.5.2.1 Long-range PCR for Kell blood group system

Two long PCR products, each approximately 12 Kb in size, were both amplified to cover the entire *KEL* gene. Table 2.2 lists the primer sequences, which were designed by Dr. Narendra Kaushik using National Centre for Biotechnology Information (NCBI) Primer Blast (National Centre for Biotechnology Information Primer Blast, 2015) and received in a high performance liquid chromatography (HPLC) form from Eurofins MWG Operon (London, United Kingdom). The reaction contained a 1X master mix (LongAmp Hot Start Taq 2X Master Mix, New England Biolabs, United Kingdom), with 200 ng of DNA template and 0.4 µM for the forward and reverse primers. The temperature profile was started with 94°C for 3 minutes for the initial denaturation, followed by 30 cycles of denaturation at 94°C for 20 seconds, annealing at 62°C for 30 seconds and extension at 65°C for 11 minutes. The final extension was 65°C for 10 minutes, and it was held at 4°C.

2.5.2.2 Agarose gel electrophoresis for the KEL gene

The PCR samples were loaded on 0.7% agarose gel, 1X TAE [40 mM Tris-acetate and 1 mM EDTA pH 8.0] (Sigma-Aldrich Company Ltd, Dorset, United Kingdom), which was mixed with fluorescent nucleic acid dye with a 1:10000 dilution of GelRed (Biotium Inc, Hayward, United States). A volume of 10 µl PCR sample was mixed with 2 µl of the DNA loading buffer [10 mM Tris-hydrochloric acid (pH 7.6), 0.03% bromophenol blue, 0.03% xylene cyanol FF, 60% glycerol, and 60 mM EDTA]. The Quick-Load® 1 Kb Extend DNA Ladder (New England Biolabs, United Kingdom) was used to assess the amplicon size. The samples were subjected to electrophoresis (Powerpac Basic Biorad, Hercules, United States) by applying 80V constant voltage. The samples were run until they reached the bottom of the gel, typically around 1 hour and 40 minutes. Then, the gel was placed into an EC3 imaging system (Ultra Violet

Products Ltd, Cambridge, United Kingdom). The software Launch Vision WorksLS was utilised to conduct the analysis of the PCR amplicons.

Table 2.2 Primer sequences used in LR-PCR to amplify the entire *KEL* gene.

Primer	Sequence 5'-3'	Exons	GC (%)	T_m(°C)	Size (bp)
Kel-121	agcagcatgttcatacccacgacct	1-9	48	67	12,243
Kel-122	ggaggctgtaagagtgctaatgttg		48	62.6	
Kel-123	gggaactccaagaatatctgctgtc	10-19	48	63.5	12,226
Kel-124	ccagccaatgataagcagccaacta		48	67.2	

T_m = melting temperature.

2.5.2.3 Long-range PCR for Rh blood group system

A total of seven LR-PCR products were designed to cover the entire genes of *RHD* and *RHCE;* three products for *RHD* and four products for the *RHCE* gene. The primer sequences for *RHD* and *RHCE* are shown in Table 2.3 and Table 2.4, respectively. The primers that exceed 75°C in their melting temperature were used in the Bloodgen project (Avent et al., 2007) and other primers were designed using NCBI Primer Blast (National Centre for Biotechnology Information Primer Blast, 2015). This primer sequence (Ds2-s) was adapted from (Legler et al., 2001). The primers were ordered in HPLC form from Eurofins MWG Operon (London, United Kingdom).

The reaction contained a 1X final concentration of 5X PrimeSTAR GXL Buffer, Takara, Japan), 200 µM of dNTP mixture, 0.2 µM of each primer and 1.25 unit of PrimeSTAR GXL Polymerase per 50 µl and the DNA concentration was 500 ng per reaction. A two-step protocol was performed as 25 cycles of 98°C for 10 seconds and 68°C for 24 minutes and finally it was held at 4°C. The amplicons were purified on 0.5% agarose gel with 80V applied for 1 hour and 40 minutes.

Table 2.3 Primer sequences used in LR-PCR to amplify the whole *RHD* gene.

Primer set	Primer name	Sequence 5'-3'	T_m (°C)	Size (bp)	Exons covered
RHD 1st amplicon	RHD F1 S	gattgggtccgtgattggcatt	60.3	22,829	1, 2 and 3
	397R*	ggccgcgggaattcgattgttgtctttatttttcaaaaccct	>75		
RHD 2nd amplicon	Ds2-s§	gccgcgaattcactagtgtgacgagtgaaactctatctcgat	>75	23,610	2 to 7
	798R*	ggccgcgggaattcgattgaggctgagaaaggttaagcca	>75		
RHD 3rd amplicon	701F*	gccgcgaattcactagtgacaaactccccgatgatgtgagtg	>75	22,731	7 to 10
	1097R*	ggccgcgggaattcgattgtggtacatggctgtattttattg	>75		

§= This primer sequence was adapted from (Legler et al., 2001).
*= These primer sequences were used in the Bloodgen project (Avent and Madgett, unpublished data).

Table 2.4 Primer sequences used in LR-PCR to amplify the whole *RHCE* gene.

Primer set	Primer name	Sequence 5'-3'	T_m (°C)	Size (bp)	Exons covered
RHCE 1st amplicon	RHCE R1 X	agtatccactttccaccttccactt	61.3	19,426	1 to 3
	ce ex3 right*	ggccgcgggaattcgatttttttcaaaaccccggaag	>75		
RHCE 2nd amplicon	RHCE SUB 2	ctactatcaagctcactgcccgattt	63.2	11,215	3 & 4
	RHCE R2	atcctggctctccttctca	56.7		
RHCE 3rd amplicon	RHCE F2	caagtccatgtgcagtgac	56.7	17,856	4 to 7
	RHCE_NR3	acagccagcatcttctttcagtcag	58.9		
RHCE 4th amplicon	RHCEex7-left*	ggccgcgggaattcgattcacatctccgtcatgcactc	>75	24,387	7 to 10
	RHCE F3	gggcagagacttgacactcc	61.4		

*= These primer sequences were used the in Bloodgen project (Bloodgen Consortium, unpublished data).

2.5.2.4 Purification of LR-PCR Products

1.8X of Agencourt® AMPure® XP reagent (~63 µl) was added to the sample (~35 µl). The mixture needed to be mixed thoroughly and incubated at room temperature for 5 minutes. The tubes were placed in a magnet for 3 minutes in order to separate the small fragments, which contain the primer-dimer and polymerase. The supernatant was discarded and the bead pellet was washed twice with 300 µl freshly prepared 70% ethanol (Fisher Scientific UK, Leicestershire, United Kingdom). The samples were air dried for up to 4 minutes and ultimately the samples were eluted with 50 µl of nuclease-free water.

2.5.2.5 Equimolar of PCR products

The starting concentration of the sequencing libraries for the long amplicons is 100 ng. Therefore, a 100 ng concentration was divided by the number of amplicons and the outcome was divided by each amplicon to obtain the volume for each one to pool the amplicons in equimolar of 100 ng. For instance, there were two products of each sample for the *KEL* gene, therefore, 50 ng of each amplicon was pooled together to obtain a final concentration of 100 ng. Regarding the products of both genes, *RHD* and *RHCE*, each gene product was pooled separately from the other gene.

Finally, the equimolar products were measured using the Qubit® 2.0 Fluorometer using the Qubit® dsDNA BR Assay kit [see section 2.2]. The products were pooled together equally in 1.5 mL Eppendorf tubes.

2.5.2.6 Fragmentation

The procedure of NGS library construction has four consecutive steps which include fragmentation, ligation to barcoded adaptors, size selection and quantification of the library using the Agilent® 2100 Bioanalyzer assay [section 2.5.2.7] prior to preparing the template for sequencing. The following protocol was used except for the size selection: 'Preparing Long Amplicon (>400 bp) Libraries Using the Ion Xpress™ Plus Fragment Library Kit (Thermo Fisher Scientific, 2012). The equimolar amplicons of 100 ng were mixed with 5 µl of Ion Shear™ Plus 10X Reaction Buffer and nuclease-free water to bring the volume to 40 µl in a 1.5 mL Eppendorf tube. Then, 10 µl of Ion Shear™ Enzyme Mix II was added to the mixture for a total volume of 50 µl. The mixture was mixed thoroughly without generating any bubbles. The tube was then incubated at 37°C for 15 minutes to obtain a 200 base-read library. 5 µl of Ion Shear™ Stop buffer was added to the mixture, which was mixed vigorously for at least 5 seconds to terminate the fragmentation reaction. The fragmented DNA was purified by Agencourt® AMPure® XP reagent by adding 1.8X of the total volume of the fragmented sample as shown in section 2.3 except the elution, which was 25 µl of low Tris-EDTA pH 8.0. The size of the fragmented DNA was assessed using the high sensitivity DNA Kit which was run on the Agilent® 2100 Bioanalyzer as discussed in the next section.

2.5.2.7 NGS libraries quantification by Agilent® 2100 Bioanalyzer system

A gel-dye mix needed to be prepared to begin the experiment. A volume of 15 µl of High Sensitivity DNA dye concentrate (blue dye) was added to a High Sensitivity DNA gel matrix with a red vial. The mixture was then mixed vigorously and transferred to the spin filter. Next, the tube was spun at 2240 × *g* for 10 minutes. Finally, the filter was discarded and the solution was kept in the dark at 4°C.

Prior to starting the assay, the gel-dye mix needed to be equilibrated at room temperature for 30 minutes. A new High Sensitivity DNA chip was placed in the chip

priming station. A volume of 9 µl was added to the first indicated well as shown in Figure 2.2 (a). The plunger was positioned in 1 mL in the chip priming station and was then closed. Then, the plunger was pressed and held by the chip for 60 seconds. Next, the plunger was slowly pulled back to the 1 mL position. The chip priming station was opened and a volume of 9 µl of gel-dye mixture was added to the three indicated wells as shown in Figure 2.2 (b). Then, a volume of 5 µl of marker was added to the twelve indicated wells as demonstrated in Figure 2.2 (c). After that, a volume of 1 µl of High Sensitivity DNA ladder was added to the indicated well as shown in Figure 2.2 (d). Finally, the samples were loaded in each well separately with 1 µl as indicated in Figure 2.2 (e).

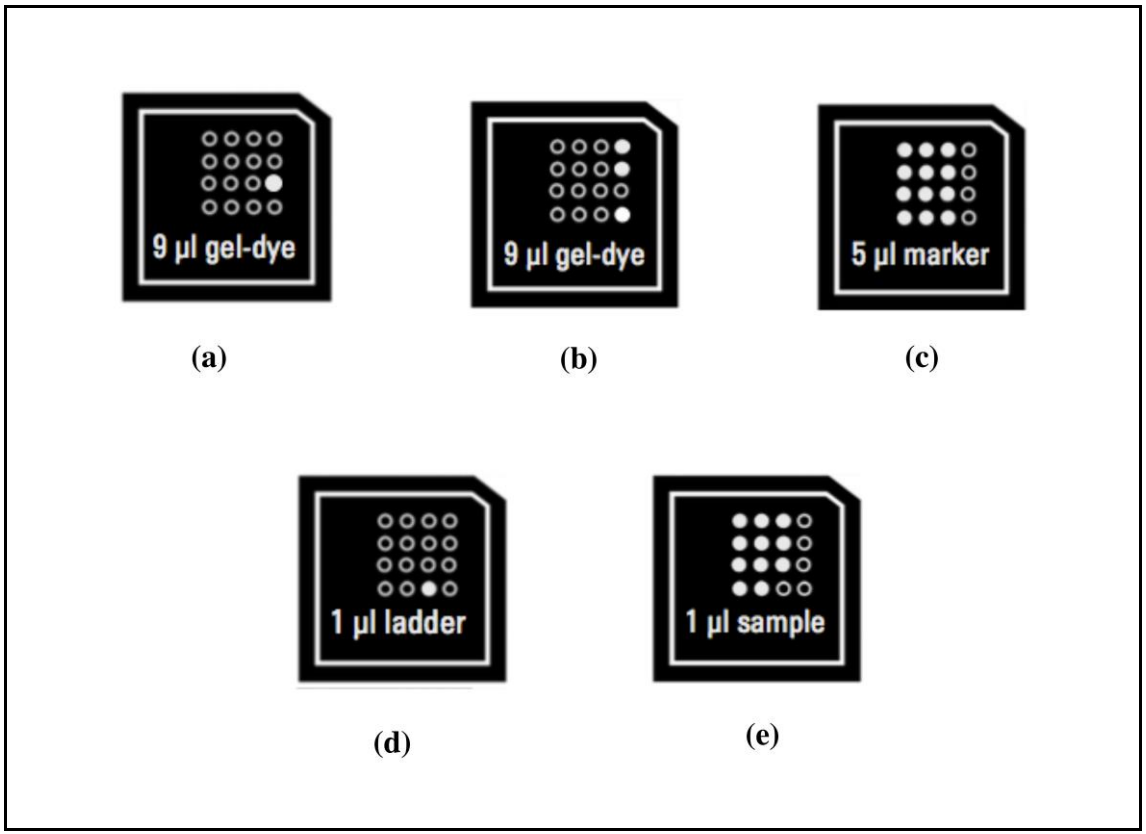


Figure 2.2 An illustration of the indicated wells of the High Sensitivity DNA chip regarding the loading positions for the quantification assay.

The assay was carried out using the Agilent® 2100 Bioanalyzer system. Adapted from (Agilent Technologies, 2013). The method is discussed in [section 2.5.2.7].

2.5.2.8 Ligation

The purified DNA fragments were then ligated to adaptors and nick-repaired by mixing 25 µl of the purified DNA with 10 µl of 10X ligase buffer, 2 µl of Ion P1 adaptor, 2 µl of Ion Xpress™ Barcode X (because multiple samples were used), 2 µl dNTP mix, 49 µl of nuclease-free water, 2 µl DNA ligase, and 8 µl nick repair polymerase in a 0.2 mL PCR tube. Then, the tube was placed in a thermocycler at 25°C for 15 minutes, then at 72°C for 5 minutes and the samples were finally held at 4°C. Following the run on the thermocycler, the ligation reactions were purified with 1.4X of Agencourt® AMPure® XP reagent for 200 base-read sequencing. The reactions were purified as discussed above in section 2.5.2.4.

2.5.2.8 Size selection of DNA Library

Initially, the purified ligated libraries were brought up to a total volume of 100 µl. The ratio used for the size selection was (0.8X-0.7X). Various optimisation ratios were performed are demonstrated in Appendix C. For such a starting material of 100 µl of the sequencing library, 0.8X of SPRIselect® (80 µl) was added and then incubated at room temperature for 5 minutes. The reaction was then placed on a magnet in order to remove the large DNA fragments, which were larger than 200 bp. Then, the supernatant was transferred, which was around 180 µl (i.e. 100 µl of DNA + 80 µl of SPRIselect® beads) into a new 1.5 mL tube. After that, 0.7X of SPRIselect® (70 µl) was added and then incubated at room temperature for 5 minutes and placed on a magnet. The desired fragments were bound to the beads, while the small fragments lower than 100 bp were not. The supernatant was discarded and the beads were then washed twice using 300 µl of 80% freshly prepared ethanol. The pellet was left to air-dry for up to 4 minutes. Finally, the beads were eluted using 50 µl of nuclease-free water and transferred to a new tube. Prior to preparing the sequencing template, the size-selected sequencing

libraries were quantified using the Agilent® 2100 Bioanalyzer instrument as discussed in section **2.5.2.7.**

2.5.3 Preparation of sequencing template

2.5.3.1 Assessment of the sequencing libraries and preparation of sequencing template

In order to prepare the template for sequencing on the Ion PGM™, the sequencing libraries needed to be quantified in order to determine the template dilution factor. The size distribution and the concentrations of the sequencing libraries were assessed using the Agilent® 2100 Bioanalyzer with the High Sensitivity DNA Kit as described in [section 2.5.2.7] and Ion Library Quantitation Kit [section 2.5.1.4]. The dilution factor for the sequencing libraries was determined using the following formula:

Dilution Factor = (Library concentration in pM/100 pM)

For instance, the library concentration was 363 pM. Therefore, the dilution factor would be as follows:

Dilution factor = 363 pM/100 pM= 3.63

Thus, 1 µl of DNA library was mixed with 2.63 µl of nuclease-free water to obtain a yield of approximately 100 pM.

The sequencing template was then immobilised to ion sphere particles (ISPs) that were clonally amplified using emulsion PCR, followed by emulsion breaking and enrichment for positive sphere particles using the Ion OneTouch™ system. The following sections will discuss this procedure in detail.

2.5.3.2 Monoclonal Amplification by Ion OneTouch™ 2

The template preparation and the sequencing protocol were performed by Dr. Michele Kiernan at the Systems Biology Centre at Plymouth University. This was according to

the following: protocol Ion PGM™ Template OT2 200 Kit for use with the Ion OneTouch™ 2 (Thermo Fisher Scientific, 2014b). The sequencing libraries of all the samples were diluted with nuclease-free water to give a total volume of 25 µl. The samples were then mixed vigorously for 5 seconds, spun down for 2 seconds and placed on ice. After that, an amplification solution was prepared by mixing a volume of 25 µl of the diluted sequencing libraries with the following reagent: 500 µl of Ion PGM™ Template OT2 200 Reagent Mix, 300 µl of Ion PGM™ Template OT2 200 PCR Reagent B, 50 µl of Ion PGM™ Template OT2 200 Enzyme Mix and 25 µl of nuclease-free water. The mixture was then mixed vigorously for 5 seconds and spun down for 2 seconds.

100 µl of ISP was mixed vigorously and then added to the mixture. The amplification mixture was mixed vigorously for 5 seconds. The prepared amplification solution was added to the sample port of the Ion PGM™ OneTouch Plus Reaction Filter Assembly [Figure 2.3]. Then, 1000 µl of Ion OneTouch™ Reaction Oil was added through the sample port. After that, a further 500 µl was added to the sample port. Then the filled Ion PGM™ OneTouch Plus Reaction Filter Assembly was inverted and installed into the three holes on the top stage of the Ion OneTouch™ 2 Instrument. Finally, the clonal amplification was performed by running the emulsion mixture on the Ion OneTouch™ 2 Instrument.

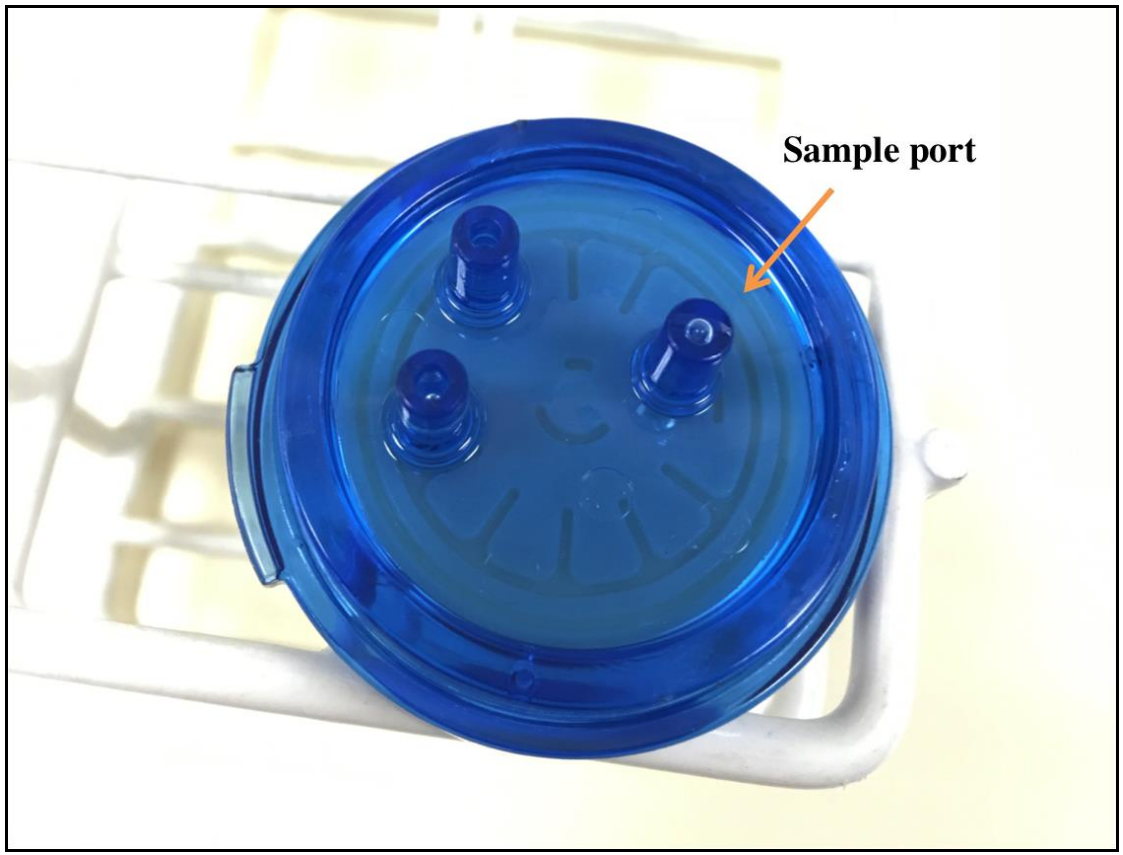


Figure 2.3 The sample port that was used to prepare the sequencing template for the clonal amplification.

The clonal amplification was performed using the Ion OneTouch™ 2 System.

2.5.3.3 Recovery of the positive ISPs

At the end of the run, which took about 4 hours, the reaction was spun in the centrifuge of the same instrument, i.e. the Ion OneTouch™ 2 Instrument. The Recovery Router was discarded immediately following the centrifugation step. All but 50 µL of the Ion PGM™ OT2 Recovery Solution was removed from each Recovery Tube. The ISPs were resuspended in the remaining Ion PGM™ OT2 Recovery Solution. The pellet was mixed properly until it dissolved in the solution. After that, the suspensions from both tubes were transferred to a new 1.5 mL Eppendorf tube for a total volume of 100 µL followed by adding 1 mL of Ion OneTouch™ Wash Solution. The ISPs were mixed by pipetting up and down and were then transferred into Well 1 of an 8-well strip [Figure 2.4] for a total volume of 100 µL of suspended ISPs. The quality of the unenriched template and positive ISPs was assessed using the Qubit® 2.0 Fluorometer. Ultimately, enrichment was accomplished to obtain many copies of the positive ISPs.

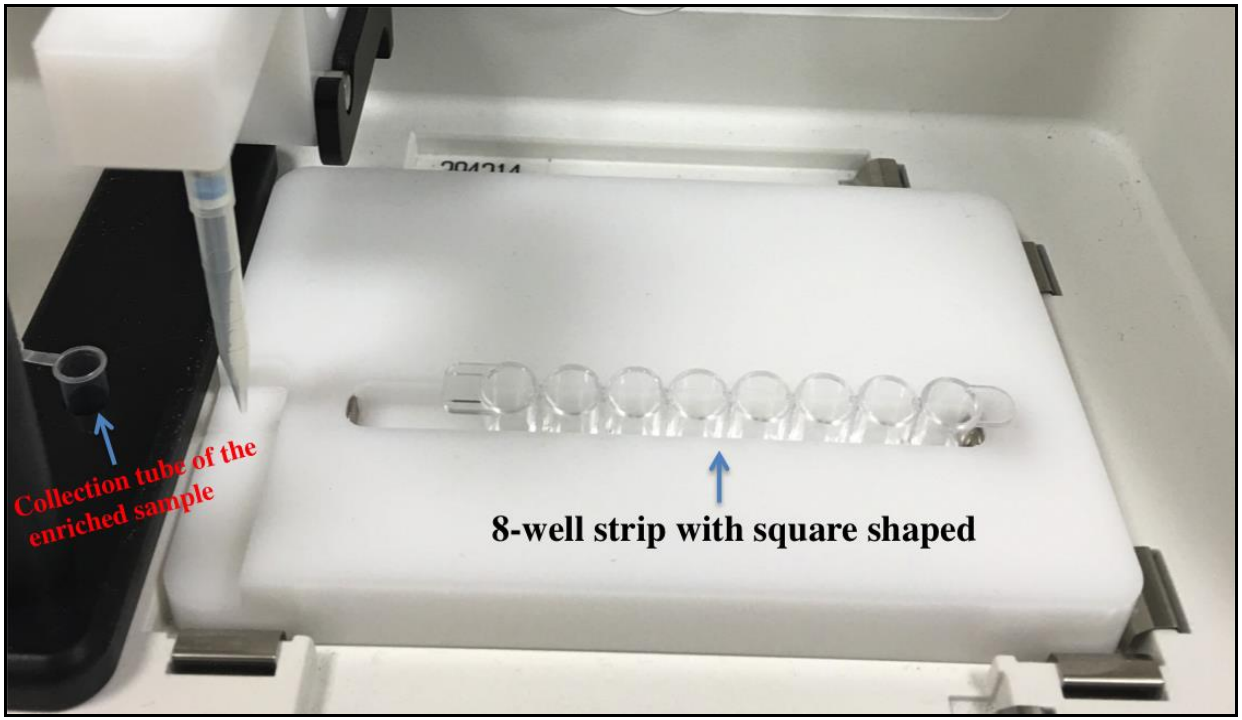


Figure 2.4 The eight wells for the enrichment of the template using the Ion OneTouch Enrichment System.

2.5.3.4 Enrichment the template positive ISPs

Dynabeads® MyOne™ Streptavidin C1 Beads were created by mixing the tube for 30 seconds to resuspend the beads followed by spinning for 2 seconds. The beads were then dispersed and mixed properly using a new pipettor tip and immediately 13 µL of Dynabeads® MyOne™ Streptavidin C1 Beads was transferred to a new 1.5-mL Eppendorf LoBind® Tube. After that, the tube was placed on a magnet for 2 minutes and the supernatant was carefully removed and discarded without disturbing the pellet. 130 µL of MyOne™ Beads Wash Solution was added to the beads. The tube was taken off the magnet and mixed for 30 seconds and spun for 2 seconds. The total amount of 130 µL was added to Well 2 of the 8-well strip [Figure 2.3]. Then Wells 3, 4 and 5 were filled with 300 µL of Ion OneTouch™ Wash Solution (W). Wells 6 and 8 were empty [Figure 2.3]. Well 7 was filled with freshly prepared melt-off solution that was prepared by mixing 280 µL of Tween solution with 40 µL of 1 M NaOH.

2.5.3.5 The Run of the Enrichment System

The Enrichment System was prepared by loading a new tip in the Tip Arm as shown in [Figure 2.3]. An opened 0.2 mL tube was inserted into the hole in the base of the Tip Loader and was filled with 10 µL of neutralisation solution. Prior to starting the run, the mixture in well 2 was mixed properly to resuspend the beads. The run time was up to 35 minutes. After finishing the run, the 0.2 mL tube was closed and removed from the Enrichment System, which possessed the positive ISPs. The mixture of the 0.2 mL tube was mixed by gently inverting the tube 5 times. Then, the templates were ready to be sequenced on the Ion PGM™. The quality of the enrichment efficiency for positive ISPs was assessed using the Qubit® 2.0 Fluorometer.

2.5.4 Sequencing on the Ion Torrent PGM™

The Ion PGM™ must be cleaned and initialised prior to use for sequencing according to the manufacturer's instructions. In a 0.2 mL PCR tube, the enriched template positive ISPs were prepared by adding 5 µL of Control Ion Sphere™ Particles to the enriched template-positive ISPs. The solution was mixed thoroughly and was spun at 15,500 × g for 2 minutes. The supernatant was then removed but 15 µL was left in the tube without disturbing the pellet. A total volume of 12 µL of sequencing primer was added to obtain a total volume of 27 µL. The whole solution was mixed carefully without disturbing the pellet. The reaction was run on a thermocycler on the following programme: 95°C for 2 minutes followed by 37°C for 2 minutes with utilising the heated lid option. This step was performed to anneal the sequencing primer to the ISPs. After the run was finished, 3 µL of Ion PGM™ Sequencing 200 v2 Polymerase was added to the ISPs to obtain a total volume of 30 µL. The sequencing reaction was then mixed properly and incubated at room temperature for 5 minutes.

The Ion 316™ chip was assessed for sequencing using the Ion PGM™ according to the manufacturer's instructions. In order to load the sequencing reaction to the chip, the liquid inside the chip needed to be removed. The chip was tilted to 45 degrees and the liquid was removed by inserting the pipette tip firmly into the loading port, which is the lower port on the chip. The chip was then placed in an upside-down position in the centrifuge adapter bucket and the bucket was transferred to the MiniFuge with the chip tab pointing in, i.e. toward the centre of the centrifuge. Then the chip was spun for 5 seconds to entirely empty the chip. Next, The chip was placed in the bucket on a flat and steady surface and the loading was performed slowly in the loading port at a rate of 1 to 2 µL per second. The chip and bucket were spun in the MiniFuge for 30 seconds with the chip tab pointing in. After setting the pipette to 30 µL, the sequencing reaction was mixed by tilting the chip to 45 degrees by pipetting the sample in and out of the

chip three times. The chip was spun again for 30 seconds in the MiniFuge but with the chip pointing out. The sample was mixed again thoroughly by pipetting the solution in and out three times. The liquid on the chip's port was removed from the loading port by tilting the chip to 45 degrees. The chip was finally positioned in the Ion PGM™ sequencer and the sequencing reaction was starting by running the sequencing programme as described by the manufacturer's instructions. The sequencing on the Ion PGM™ is based on sequencing by synthesis in which a polymerase incorporates a nucleotide. Accordingly, a hydrogen ion is released and the base is called.

2.5.5 Data analysis for NGS

Ion PGM™ sequencer transferred the raw sequencing data, which were the digital representation of the voltage results from pH change in the wells, to the Torrent Server. Consequently, the Torrent Server converted the raw sequencing data to base calls. Figure 2.5 illustrates the workflow data analysis of NGS. Torrent Suite™ Software Version 4.4 was used in order to generate a summary sequencing report indicating the number of reads generated by the sequencer and the percentage of chip loading. The Torrent Suite™ was used different filters to remove duplicate reads, low quality reads, polyclonal templates and primer dimer. Furthermore, it generated different files for the analysis including Fastq files, Binary Alignment Map (BAM) in a conjunction with Binary Alignment Index (BAI) and Variant Call Format (VCF) files.

The FastQC software was run to assess the quality control across the reads generated (Andrews, 2010). The sequencing samples were then aligned to reference the human genome (hg19) using the BAM/BAI files and were visualised using Integrative Genome Viewer (IGV) Version 2.3.46. Moreover, the IGV software was used to assess the depth of coverage of the sequencing reads, zygosity, quality of the sequencing reads and the mapping quality.

The samples were then annotated using the VCF files to obtain the SNPs and indels to analyse the genotype and predict the phenotype. The Ion Reporter™ Version 4.6 and SeattleSeq Annotation tool 141 site were performed to annotate the sequencing data of the HEA and HPA Panel and LR-PCR approach, respectively (SeattleSeq Annotation Tool 141, 2014). Antigens were determined by choosing the right transcript according to the Blood Group Antigen Factsbook (Reid et al., 2012). Each antigen was determined by its chromosomal location, the type of variant whether it was an SNP or indel, gene, the reference nucleotide, the changing nucleotide, depth of coverage, the zygosity, the transcript used in analysis based on the NCBI database, the location of the variant whether it was intronic, synonymous or exonic SNPs, codon, an exon number of that variant, an amino acid substitution and the position of the nucleotide change. According to Bentley et al. (2008), the depth of coverage was determined to 15X and 33X for homozygous and heterozygous SNPs, respectively.

2.5.5.1 Masking the Rh genes

Due to the high homology between the *RHD* and *RHCE* genes, when analysing the gene of interest masking the other gene was required. The unrequired gene was annotated by 'Ns' on their sequencing nucleotides. Figure 2.6 demonstrates the masking procedure for the Rh sequencing data. By using Browser Extensible Data (BED) files, the maskfasta utility from the bedtools website was used to mask the *RHCE* gene in order to analyse the *RHD* gene and vice versa (Bedtools, 2015). This type of analysis was performed by Dr. Xinzhong Li.

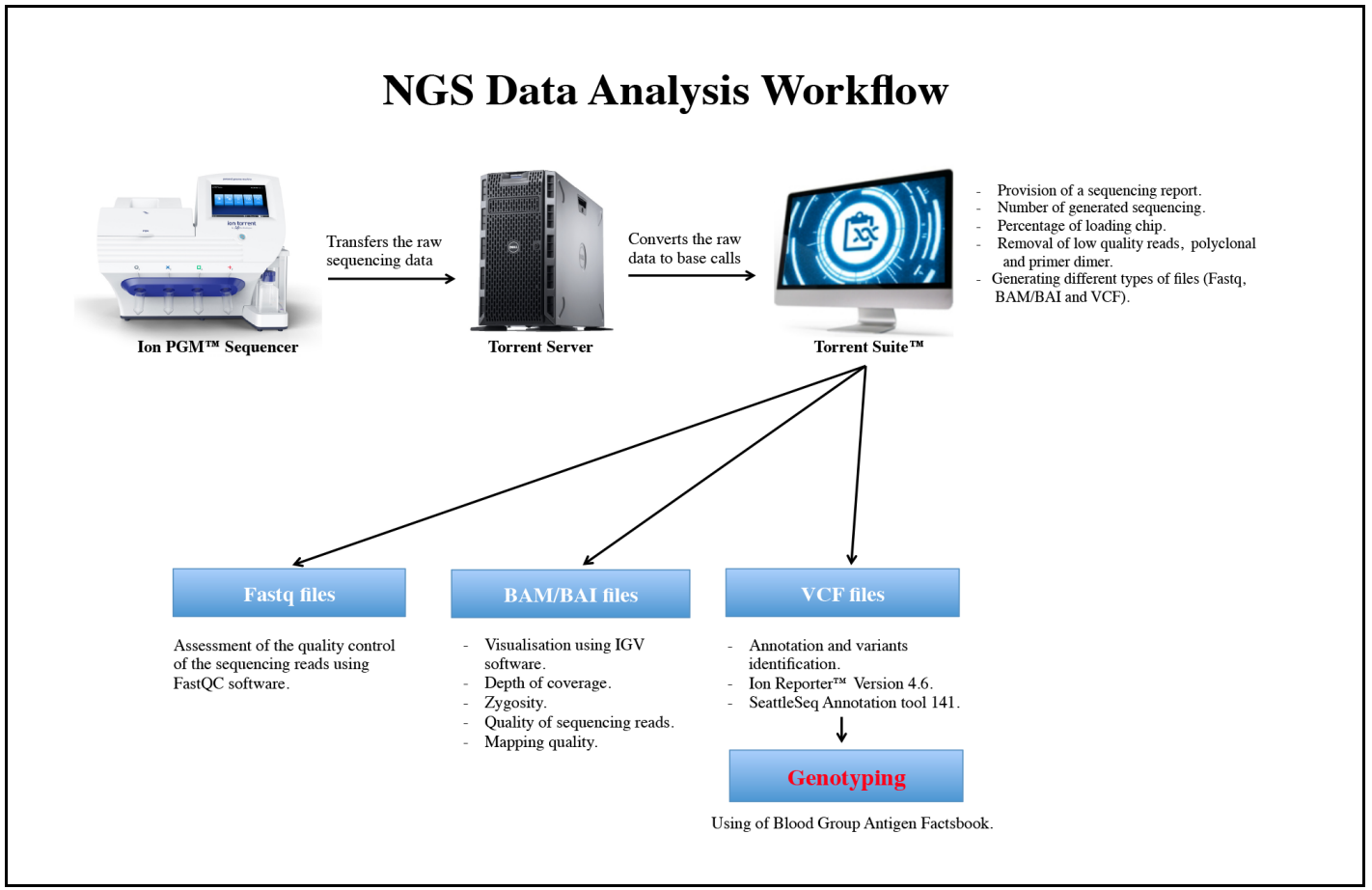


Figure 2.5 A brief summary of NGS data analysis workflow.

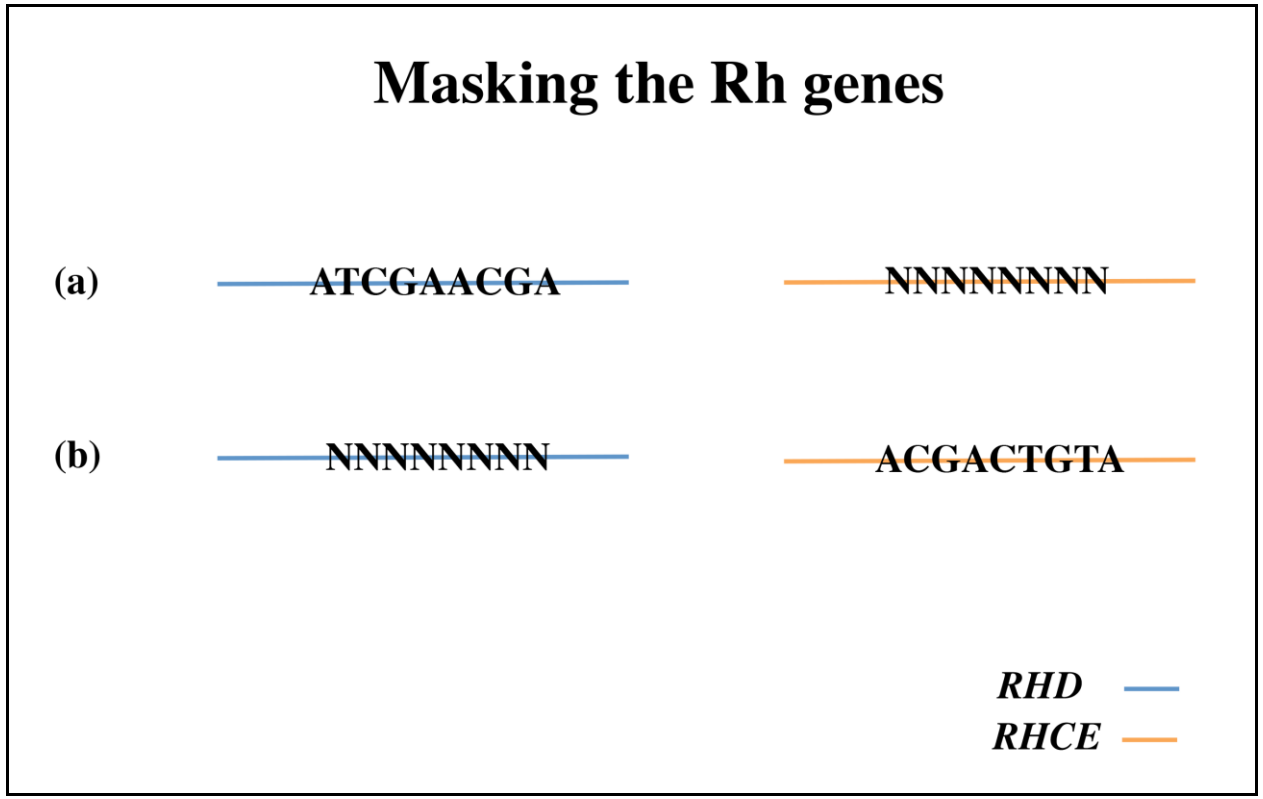


Figure 2.6 Masking the Rh genes.

(a) When the analysis carried out for the *RHD* gene, the *RHCE* gene was masked by 'Ns' to prevent any misalignement occurred.
(b) The *RHD* gene was masked by 'Ns' when the analysis was performed to the *RHCE* gene.

2.5.6 Validation of the NGS data by Sanger sequencing

Sanger sequencing was performed in order to validate some of the *KEL* results that were obtained by the LR-PCR approach. The purified amplicon was mixed with one of the primers and nuclease-free water and was then sent to the Eurofins service for sequencing. The files were received in (.ab1) format for the DNA sequence chromatogram and this was visualised and analysed using MacVector Software Version 12.7.0. Table 2.5 lists the primers used to validate SNPs found in the *RHCE* and *KEL* genes.

The forward primer for exon 2 of the *RHCE* was adopted from (Legler et al., 2001), while the reverse primer was adopted from the Bloodgen project (Avent et al., 2007).

The primers of exon 6 of the *KEL* gene were adopted from (Beiboer et al., 2005). The primers for exon 4, 17 and 19 of the *KEL* gene was designed using NCBI Primer Blast (National Centre for Biotechnology Information Primer Blast, 2015) and received in HPLC form from Eurofins MWG Operon (London, United Kingdom).

Table 2.5 Primer sequences used to validate the SNP encoding the K antigen from NGS data.

Primer	Gene	Sequence 5'-3'	Exon	GC (%)	T_m (°C)	Product size (bp)
Kell exon 4 F Kell exon 4 R	*KEL*	acttgcacagagcatcttcc atgttctttgcctagcccc	4	50% 52.6%	57.3 56.7	290
Kell exon 6 F Kell exon 6 R	*KEL*	gccgcgaattcactagtgagctgtgtaagagccgatcc ggccgcgggaattcgattaagggaaatggccatactga	6	55.3% 52.6%	>75 >75	207
Kell exon 17 F Kell exon 17 R	*KEL*	gccatgcaactgtacttg agaccacaagggaggcc	17	50% 64.7%	53.7 57.6	455
Kell exon 19 F Kell exon 19 R	*KEL*	ccagggaggtgttggtcg gtcaccgggtcagctttg	19	66.6% 61.1%	60.5 58.2	275
Ds2-s 296R	*RHCE*	gccgcgaattcactagtgtgacgagtgaaactctatctcgat ggccgcgggaattcgattagaagtgatccagccaccat	2	47.6% 55.2%	>75 >75	1315

Chapter 3 : Blood Group Genotyping by Microarrays

3.1 Introduction

By the 1990s all the major blood group systems had been identified at the molecular level. Most of the polymorphisms of these blood groups are caused by SNP, which change the conformation on the protein structures (Avent, 1997). The previous genotyping techniques such as RFLP-PCR and SSP-PCR were carried out to genotype some blood groups such as Lewis blood groups system and the Kell blood group system, respectively (Avent and Martin, 1996; Kudo et al., 1996). These techniques have disadvantages, including time consuming, labour intensive and analysis of the post-PCR gel. Furthermore, incomplete digestion of the PCR products can be occurred for the RFLP-PCR (Veldhuisen et al., 2009). In addition, they are not suitable for high-throughput genotyping of the blood groups. Consequently, a high-throughput platform, to genotype the blood group antigens, are required as well as the ease of the automation of the process to be utilised in the blood banks [see section 1.9.2] (Veldhuisen et al., 2009).

The Bloodgen project (2003-2006) was conducted by a consortium of universities, blood centres and Progenika Biopharma SA. The aims of Bloodgen project were to design a DNA array-based assay in order to obtain extended genotyping of the blood group antigens for routine blood donors and for the patients. This was aim to help patients especially those suffering from alloimmunisation due to receiving multiple blood units. Moreover, another aim was to have long-term health benefits by extensive genotyping at the birth of each individual (Avent et al., 2009). The assay is more accurate in comparison to the standard typing by serology and might replace it in the future.

There are currently many platforms using microarray for BGG. Two approaches of have been manufactured by Progenika Biopharma SA. These were: glass array and

suspension beads array. BLOODchip® is a product depends on glass array to genotype genomic polymorphisms that give rise to the blood group antigens and HPAs.

3.2 Aims

In this chapter, the ID Core+ kit is used to genotype the blood group alleles. The method of this test is discussed in [section 2.3]. The ID Core+ kit is a test targeting 31 SNPs from the following blood group system (RhCE, Kell, Kidd, Duffy, MNS, Diego, Dombrock, Colton and Yt). The first step in this procedure was an amplification of the targeted region of genomic DNA using multiplex PCR with biotinylated dCTP. Denaturation then took place to hybridise the amplified products to oligonucleotide probes, which are attached to the beads. After that, the hybridised DNA was labelled with a fluorescence SAPE conjugate. The existence of specific SNP in the sample can be determined by the correlation of the fluorescence signal, which is intrinsic to each microsphere with the existence or absence of a corresponding phycoerthrin signal. The detected signals were then quantified by a Luminex® 100/200™ flow cytometer. Finally, Progenika's Work Station Software was used to analyse the data and produce a report with the predicted phenotypes. The aim of this method was to provide training on blood group genotyping using the microarray technology as well as assessment of the accuracy of the ID Core+ kit in comparison to the standard typing by serology.

3.3 Results

Twenty-eight samples (Cohort A) were chosen randomly and examined in total, of which 24 samples obtained full results, while four samples had an issue of non-valid results. Table 3.1 lists the serological results provided by NHSBT Filton. Table 3.2 illustrates the predicted phenotypes of the 24 samples that resulted when using the ID Core+ kit. All of the 24 sample results from the ID Core+ test were similar to those previously typed by serology, which were done at the NHSBT Filton, Bristol. Table 3.2 demonstrates the results of the predicted phenotypes of four samples that had an issue of a non-valid assay test using the ID Core+ kit. In another sample all the first pool (RhCE, Kell, Kidd, Duffy and MNS) was non-valid and in two samples the second pool (Diego, Dombrock, Colton and Yt) was non-valid. Sample 227P shows two non-valid results for *KEL*02.06* and *KEL*02.07* alleles encoding the Js^a and Js^b antigens, respectively.

Four samples were genotyped as heterozygous *RHCE*C/c* and three as homozygous *RHCE*C/C*, while 17 samples were found to be homozygous for the *RHCE*c* allele. Regarding the E antigen, three samples were found to be homozygous *RHCE*E/E* and four were heterozygous *RHCE*E/e*. Seventeen samples were found to be homozygous for the *RHCE*e* allele.

The K antigen was observed as heterozygous *KEL*01/02* in a single sample out of total 24 samples. The *KEL*02.03* allele encoding the Kp^a antigen was found in two samples to be heterozygous with the *KEL*02.04* allele encoding the Kp^b antigen. The *KEL*02.07* allele, encoding the Js^b antigen, was found to be homozygous in all the samples. For the Kidd blood group system, nine samples were found to be homozygous for the *JK*A* allele, four homozygous for *JK*B* allele and 11 heterozygous for both alleles, *JK*A* and *JK*B*. There were nine samples homozygous for the *FY*A* allele, six

homozygous samples for the *FY*B* allele and nine heterozygous samples for both alleles.

For the MNS blood group system, 12 samples were found to be heterozygous for both *GYPA*M* and *GYPA*N* alleles, nine homozygous for the *GYPA*M* allele and three homozygous for the *GYPA*N* allele. Three samples of *GYPB*S* allele, 10 samples of homozygous *GYPB*s* allele and 11 samples of heterozygous *GYPB*S* and *GYPB*s* alleles were observed. The U antigen was predicted in all the samples. Regarding the Diego blood group system, all the samples observed the *DI*B* allele.

12 samples were found to be heterozygous for the *DO*A* and *DO*B* alleles of the Dombrock blood group. The *DO*A* allele was found to be homozygous in five cases, while *DO*B* allele was found to be homozygous in seven cases. The *DO*02.04* and *DO*02.05* alleles, which encode the Hy and Jo^a antigens, respectively were observed in all the samples. Regarding the Colton blood group system, the *CO*A* allele was found to be heterozygous with *CO*B* in two samples. However, the remaining 22 samples were found to be homozygous for the *CO*A* allele. *YT*B* allele was found to be heterozygous in one sample, while the remaining 23 samples were found to be homozygous for the *YT*A* allele.

Table 3.1 The serological results provided by NHSBT Filton for 28 samples that were used for microarray.

	ABO	Rh	D	C	E	c	e	C^w	M	N	S	s	P1	Lu^a	Lu^b	K	k	Kp^a	Kp^b	Le^a	Le^b	Fy^a	Fy^b	Jk^a	Jk^b	A1
057D	O+	R_0r	+	−	−	+	+		+		+	+		−	+	−		−	+			−	+	+	+	
103E	A-	rr	−	−	−	+	+	−	+		+	−		−	+	−		−	+			+	−	−	+	
105U	B−	rr	−	−	−	+	+	−	+	−	+	+	+			−				−	+	+	+	+	−	
303N	A+	R_2r	+	−	+	+	+									−										
310P	A−	rr	−	−	−	+	+		+		+	+		−		−		−				−	+	+	+	
311N	A+	R_2r	+	−	+	+	+									−										+
314B	A+	R_2R_2	+	−	+	+	−	−	+		−	+		−	+	−		−	+			+	−	+	−	
352*	O−	rr	−	−	−	+	+	−								−										
363*	O−	rr	−	−	−	+	+									−										
427H	A−	rr	−	−	−	+	+	−	−	+	−	+		−	+	−			+	−	+	−	+	+	−	
524J	O−	rr	−	−	−	+	+	−	+	+	+	+				−				+	−	+	−	+	−	
526K	O+	R_1r	+	+	−	+	+									−										
552G	A+	R_1R_1	+	+	−	−	+		+		+	−		−		−		−				+	+	+	+	+
631E	A+	R_1R_1	+	+	−	−	+	−	+			+		−		−		−				+	−	+	−	
728Q	AB−	rr	−	−	−	+	+	−	+	−	−	+	+	−	+	−	+	−	+	+	−	+	−	−	+	
729V	B+	R_1R_2	+	+	+	+	+	−								−										
731H	B+	R_1R_1	+	+	−	−	+									−										
739K	O−	rr	−	−	−	+	+	−	+	−	−	+				−		−				−	+	+	+	
744Q	O+	R_2r	+	−	+	+	+		+		−	+		−		−		−				+	+	−	+	
753T	B+	R_2R_2	+	−	+	+	−	−	+	−	−	+		+		−		−				+	−	+	−	
846D	O+	R_1r	+	+	−	+	+	−	+	+	+	+		−	+	−		−	+	−	+	+	+	+	+	
910H	O−	rr	−	−	−	+	+									+										
967L	B+	R_1r	+	+	−	+	+	−	+		−	+				−						+	+	+	+	
0657	A+	R_2R_2	+	−	+	+	−	−								−										
227P	B+	R_1R_2	+	+	+	+	+									+										
140Y	O+	R_2r	+	−	+	+	+	−	+			+	+			−				−		−		+	+	
5015	A−	rr	−	−	−	+	+	−	+		−					−						−		+	+	
389A	A+	R_1r	+	+	−	+	+		+							−				−				+	+	

Table 3.2 The results of the predicted phenotypes of 24 samples using the ID Core + kit.

The kit investigates the alleles of the following blood group systems: *RHCE* gene from Rh system, Kell, Kidd, Duffy, MNS, Diego, Dombrock, Colton, and Cartwright. DNA samples were extracted from the blood donors, amplified using multiplex PCR, hybridised to allele-specific oligonucleotide probes containing the SNP, labelled to fluorescent dyes, and finally quantified using the Luminex® system. The genotypes and phenotypes of the blood groups were detected via software analysis using BIDS (Workstation Progenika Version 2.0 Beat 6, which worked in combination with Luminex xPonent® LX100/LX200 Version 3.1). The blank fields are considered as negative results for clarity.

	RhCE							Kell						Kidd		Duffy		MNS					Diego		Dombrock				Colton		Yt	
Sample ID	**C**	**c**	**E**	**e**	**C^{x}**	**C^{w}**	**VS**	**K**	**k**	**Kp^{a}**	**Kp^{b}**	**Js^{a}**	**Js^{b}**	**Jk^{a}**	**Jk^{b}**	**Fy^{a}**	**Fy^{b}**	**M**	**N**	**S**	**s**	**U**	**Di^{a}**	**Di^{b}**	**Do^{a}**	**Do^{b}**	**Hy**	**Jo^{a}**	**Co^{a}**	**Co^{b}**	**Yt^{a}**	**Yt^{b}**
057D		+		+					+		+		+	+	+		+	+	+	+	+	+		+	+	+	+	+	+	+	+	
103E		+		+					+		+		+		+	+		+		+		+		+	+	+	+	+	+		+	
105U		+		+					+		+		+	+		+	+	+		+	+	+		+	+	+	+	+	+		+	
303N		+	+	+					+		+		+	+	+	+		+	+	+	+	+		+		+	+	+	+		+	+
310P		+		+					+		+		+	+	+		+	+	+	+	+	+		+	+	+	+	+	+		+	
311N		+	+	+					+		+		+	+	+	+			+	+	+	+		+	+		+	+	+		+	
314B		+	+						+		+		+	+		+		+			+	+		+		+	+	+	+		+	
352*		+		+					+		+		+	+	+	+			+		+	+		+	+	+	+	+	+		+	
363*		+		+					+		+		+	+	+		+	+	+		+	+		+	+	+	+	+	+		+	
427H		+		+					+		+		+	+			+		+		+	+		+	+	+	+	+	+		+	
524J		+		+					+		+		+	+		+		+	+	+	+	+		+		+	+	+	+	+	+	
526K	+	+		+					+		+		+	+	+		+	+	+	+	+	+		+		+	+	+	+		+	
552G	+			+					+		+		+	+	+	+		+		+		+		+	+	+	+	+	+		+	
631E	+			+					+		+		+	+		+		+	+		+	+		+	+	+	+	+	+		+	
728Q		+		+					+		+		+		+	+		+			+	+		+	+		+	+	+		+	
729V	+	+	+	+					+		+		+		+	+	+	+	+	+		+		+	+	+	+	+	+		+	
731H	+			+					+		+		+	+		+	+	+		+	+	+		+	+	+	+	+	+		+	
739K		+		+					+		+		+	+	+		+	+			+	+		+	+		+	+	+		+	
744Q		+	+	+					+		+		+		+	+	+	+	+		+	+		+		+	+	+	+		+	
753T		+	+						+		+		+	+		+		+			+	+		+	+		+	+	+		+	
846D	+	+		+					+		+		+	+	+	+	+	+	+	+	+	+		+	+	+	+	+	+		+	
910H		+		+				+	+		+		+	+		+	+	+		+	+	+		+		+	+	+	+		+	
967L	+	+		+					+	+	+		+	+	+	+	+	+	+		+	+		+		+	+	+	+		+	
0657		+	+						+	+	+		+	+		+	+	+	+	+	+	+		+	+		+	+	+		+	

Table 3.3 The predicted phenotypes of four samples that had some non-valid results using the ID Core+ kit.

One of the samples (227P) only had two non-valid (NV) SNPs, in another sample all the first pool (RhCE, Kell, Kidd, Duffy and MNS) was non-valid and in two samples the second pool (Diego, Dombrock, Colton and Yt) was non-valid. Sample 227P shows two non-valid results for the SNPs encoding Js^a and Js^b antigens. The blank fields are considered as negative results for clarity.

	RhCE							Kell						Kidd		Duffy		MNS					Diego		Dombrock				Colton		Yt	
Sample ID	**C**	**c**	**E**	**e**	**C^x**	**C^w**	**VS**	**K**	**k**	**Kp^a**	**Kp^b**	**Js^a**	**Js^b**	**Jk^a**	**Jk^b**	**Fy^a**	**Fy^b**	**M**	**N**	**S**	**s**	**U**	**Di^a**	**Di^b**	**Do^a**	**Do^b**	**Hy**	**Jo^a**	**Co^a**	**Co^b**	**Yt^a**	**Yt^b**
227P	+	+	+	+				+	+		+	NV	NV	+		+	+		+	+	+	+		+	+	+	+	+			+	
140Y	NV	NV	NV	NV	NV	NV	NV	NV	NV	NV	NV	NV	NV	NV	NV	NV	NV	NV	NV	NV	NV	NV		+		+	+	+	+		+	
5015		+		+					+		+		+	+	+		+	+	+		+	+	NV	NV	NV	NV	NV	NV	NV	NV	NV	NV
389A	+	+		+					+				+	+	+	+		+		+		+	NV	NV	NV	NV	NV	NV	NV	NV	NV	NV

3.4 Discussion

The importance of BGG has become extremely obvious, especially in multiply transfused patients. Consequently, the requirement of a high-throughput platform for both donors and patients is needed (Avent et al., 2007). The ID Core+ test is a genotyping blood test for most clinically significant blood groups (RhCE, Kell, Duffy, Kidd, MNS, Colton, Dombrock, Diego and Yt). This kit depends on the microsphere-based suspension array using the (xMAP technology) Luminex® system for the detection of SNPs in blood samples. Up to 48 samples can be examined in a single run to increase the high-throughput screening of the donors.

The results obtained from this experiment were matched to the serology testing provided and gave 100% accuracy. Some issues were reported in which the samples were not valid for all the SNPs in one of the two pools [Table 3.3]. Although the reason one pool works and the other does not may be related to the hybridisation that has not taken place and not related to the purity of the DNA samples, as some results were obtained. There may be a technical problem in the hybridisation, which requires mixing the beads properly with the PCR products 20 times by pipetting up and down. Therefore, this might be the actual reason for the non-valid results. In one sample, there were two non-valid SNPs that were associated with the alleles *KEL*02.06* and *KEL*02.07* encoding the Js^a and Js^b antigens although the remaining SNPs in both pools were working properly and results were obtained. It could be that the amplification of the responsible PCR amplicons did not take place in the first instance. This may be due to a mutation in primer-binding sites or a novel allele, often this is referred to as allelic dropout.

It should be noted that the genotyping of blood groups with such a kit gives an extensive prediction of the phenotypes of the blood group systems such as Diego, Dombrock, Colton and Yt. For example, the low prevalence alleles can be examined easily by the ID Core+ test. The predicted antigens include Yt^b and Co^b, the first of which is found as 8% in Europeans and the latter as 10% in all populations (Reid et al., 2012).

The drawback of ID Core+ kit is that it could not detect the ABO and RhD antigens, which are the most clinically significant systems. This is according to the fact that the D antigen is encoded by the *RHD* gene, which has many polymorphisms due to various genetic events. These include the deletion of the entire *RHD* gene, as appeared in RhD-negative individuals, and gene rearrangement in some of the partial D phenotypes (Avent et al., 2006). On the other hand, BLOODchip® depends on a glass microarray that allows the genotyping of extended blood group antigens, which are detected by ID Core+, as well as the D antigen with its variants (D+, D−, Del, partial D, and weak D), the ABO blood group system (A, B, AB, O, and weak A), and HPAs. The BLOODchip® genotyping takes longer to run, which is about 9 hours in comparison to the ID Core+ test, but it is a more comprehensive test. It might be used as a front-line method, due to its higher resolution compared to serological typing, as well as due to its discrimination between partial D and weak D (Avent et al., 2007; Avent et al., 2009).

BLOODchip® v 1.0 is the product of the Bloodgen project that includes five steps: DNA isolation, multiplex PCR amplification, fragmentation, labelling, hybridisation and data analysis. The test can detect 116 SNPs of the blood groups alleles, of which 73 SNPs were designed for the *RHD*. 3000 samples were investigated by BLOODchip® at eight centres across Europe to obtain CE marking. There was a discrepancy caused by serology but BLOODchip® correctly genotyped the samples. Some of these samples

were typed by serology as RhD-positive, whilst they were genotyped as partial D and weak D by BLOODchip®. Moreover, three samples were typed as weak D but they were in fact partial D. Furthermore, BLOODchip® was able to predict the weak Fy(b+^{w}) phenotype. In addition to the higher resolution of the BLOODchip® in comparison to the standard serology, BLOODchip® provides extensive genotyping for the blood group alleles that do not have serological reagents such as the Dombrock blood group (López Marínez et al., 2009). Interestingly, all the previous reasons suggest that genotyping is more accurate when using BLOODchip® and that it provides a higher resolution for predicting the phenotypes in comparison to the serology. It should be noted that any new alleles would not be determined by BLOODchip® as the SNPs need to be added to the test panel beforehand.

Denomme and Van Oene (2005) designed a microarray platform to examine 12 SNPs defining the following alleles that include exons 4 and 9 of the *RHD* gene, *RHCE*C/c*, *RHCE*E/e*, *GYPB*S/s*, *KEL*01/02, KEL*02.03/02.04*, *FY*A/B, FY*02N.01*, *JK*A/B*, *DI*A/B* and *HPA*1a/1b*. Among the 372 samples tested, 345 were successfully genotyped. The undetermined genotypes include two samples of *RHD* exon 9, three samples of *RHCE*E*, one sample of *KEL*02.03*, two samples of *JK*A/B* and two samples of *FY*B.* It should be noted that four samples failed to detect all the SNPs (Denomme and Van Oene, 2005). Although this system is capable of detecting those alleles, it still needs to be improved to overcome the error rate of detection. This can be overcome by using a technology such as NGS.

Hashmi and co-workers (2005) designed a microarray platform, which is known as BLOOD-1 BeadChip™. The assay contains nine SNPs defining the following alleles that include *FY*A/B, FY*02N.01 DO*A/B, CO*A/B, LW*A/B, DI*A/B* and *SC*1/2*. Then the panel was extended to analyse 18 targeted SNPs including the following: *GYPA*M/N, GYPB*S/s, LU*A/B, KEL*01/02, FY*265, JK*A/B, DO*02.04* encoding

Hy antigen, *DO*01.05* encoding Jo^{a} antigen and Haemoglobin S. The platform was designated as HEA BeadChip™.

The HEA BeadChip™ procedure starts by amplifying these targets using multiplex PCR. The amplicons need to be purified to remove residual primers and dNTPs prior to carrying out the downstream analysis of the microarray. This gives an advantage to the ID Core+ kit and BLOODchip® over HEA BeadChip™ that does not require any purification following the amplification. Regarding the HEA BeadChip™ kit, a denaturation step takes place to generate a single stranded DNA. After that, the PCR amplicons are annealed and elongated to beads that are coupled to oligonucleotide probes and immobilised on synthetic microparticles. The fluorescence of each bead is finally analysed using the array imaging system to determine the allelic presence or absence. Special software calculates the adjusted intensity in every reaction to assign genotypes and report the predicted phenotypes. The duration of the assay is 5 hours which includes the hands-on time but not the DNA extraction process (Hashmi, 2007).

An analysis was performed of more than 400 blood donors and showed concordant results with serology, as well as the molecular techniques including SSP-PCR and RFLP-PCR. In addition, two alleles of the Dombrock blood group were identified (Hashmi et al., 2005). However, one sample Fy(a−b−) was collected on two occasions and typed as *FY*B/B* and *GATA* heterozygous *A/B* by both BeadChip™ and RFLP-PCR, as well as *FY*265* homozygous A by only the RFLP-PCR. It should be noted that the sequencing of all the exons and introns, including the regulatory regions, revealed that there was no reason for silencing the *FY*B* alleles (Hashmi et al., 2005). In 2007, HEA BeadChip™ was used to examine 2355 donors to genotype minor blood group antigens. 24 discrepancies were found, of which 16 were a consequence of clerical errors in data entry or when recording the serological results. Eight of these were discordant due to

the silent allele of *GYPB*S* (Hashmi et al., 2007). Sequence-based typing using NGS may resolve the issues of silencing alleles.

Karpasitou and co-workers (2008) designed an in-house panel that is based on microarray beads using the Luminex® system. The designed panel examined the following alleles using a single pool; *JK*A/B, FY*A/B, GYPB*S/s, KEL*01/02, KEL*02.03/02.04, KEL*02.06/02.07, CO*A/B* and *LU*A/B* (Karpasitou et al., 2008). A further test on 2020 samples was performed to genotype the same targeted SNPs. The consistency was 100% with the predefined serology apart from one Co(a−b−), 74 Fy(b+^{w}) and 56 Fy(a−b−) samples, which were identified only by serology (Drago et al., 2009; Drago et al., 2010). The outcome of this study stressed that the microarray beads may not be a suitable platform for detecting the weak or null alleles. Genotyping by NGS is able to overcome such an issue in the detection of these variants (Avent et al., 2015).

Paris and co-workers (2014) designed a 96-well flexible microarray platform that used two robotic workstations in order to facilitate the automation of the blood group genotyping process from the DNA extraction to data analysis. Eight SNPs were examined for the following alleles: *KEL*01/02, KEL*02.03/02.04, JK*A/ B, FY*A/B, FY*02M.01, FY*01N.01, GYPA*M/N* and *GYPB*S/s*. The advantages of this test were in the automation of the genotyping procedure and that the duration of the whole procedure was less than 8 hours. On the other hand, out of the 960 samples examined, three were discordant between serology and genotyping in the detection of a SNP in each of the Kell, Kidd and MNS blood group systems. The undetermined alleles include *KEL*02.03* encoding Kp^{a} antigen, *FY*02M.01* encoding Fy^{x} phenotype and *GYPB*s*. Furthermore, very limited SNPs were examined and the number of investigated SNPs needs to increase (Paris et al., 2014). It might be an issue to increase the number of

examined SNPs when increasing the number of pools for the multiplex PCR, as it may affect the number of samples and reduce the throughput and the price will be higher.

3.4.1 Cost of the microarray platforms

The cost of the ID Core+ test is around £50 per sample and around £1.51 per SNP. The price per SNP of the test provided by Karpasitou and co-workers is more expensive than the ID Core+, which was ~£2 per SNP (Karpasitou et al., 2008). Paris et al. (2014) designed a microarray platform using the 96-well format system for the test the price was £1.67 per SNP, including the genomic DNA extraction, genotyping, labour and equipment. This price per SNP might be a slightly higher than the ID Core+ test but this included the genomic DNA extraction and the rest of the procedure (Paris et al., 2014). Denomme and Van Oene (2005) stated that the price for the assay they designed is lower by 20-fold than the standard typing by serology. This may give the microarray system the ability to perform a high-throughput screening of the donor samples (Denomme and Van Oene, 2005).

Nevertheless, the approach that was provided by Jungbauer and co-workers (2012) is still more affordable in contrast to microarray platforms, which costs around £10.73 per sample and £0.31 per SNP. The authors stated that the genotyping cost could be reduced to 25% of the assay cost for every SNP that has been tested in comparison to the standard serology (Jungbauer et al., 2012).

The test includes the same phenotypes tested by the ID Core+ kit as well as *KEL*02.11/02.17*, *KEL*02.21*, *FY*02M.01*, *LU*A/B*, *LU*02.08/02.14* and *IN*A/B* alleles. Similarly to the ID Core+ kit, this test does not type the ABO and the Rh blood group antigens. The experiment used six pools of multiplex primers, followed by purification of the results using gel electrophoresis. The results of the genotyping of 6000 samples gave 99.83% accuracy, of which 350 SNPs were not detected in the first experiment and required the test to be repeated to overcome the situation. In addition to

that, ten cases of *FY*02M.01* allele could not be investigated (Jungbauer et al., 2012). However, the post analysis of such a technique which involves the running of agarose gel electrophoresis cannot be automated. Moreover, this technique may be prone to obtaining transcription errors (Jungbauer et al., 2012). This in fact gives the advantage in accuracy of detection to the microarray technology over the approach proposed by Jungbauer et al. (2012) in detecting the SNPs.

Generally, the suspension beads array provides fast reaction kinetics, rapid interpretation of the data analysis and report production. Furthermore, no purification of the PCR amplicons is required after the multiplex PCR amplification in most platforms. In addition, it has a high sensitivity in allele detection apart from in antigens with weak or null expression as mentioned above. Drago et al. (2009) stated that the suspension beads array is flexible, in which the beads can be formulated depending on the laboratory's needs. However, the disadvantages of this technique include the lack of automation, particularly in the washing steps and sample loading. Moreover, large numbers need to be tested to ensure the test is cost effective, therefore, the laboratories with a low workload will not obtain the benefits of such a procedure (Drago et al., 2009).

All of the microarray platforms require previous knowledge of the allele to be incorporated into the platforms. This issue can be overcome using NGS technology. Microarray platforms have an obstacle in identifying the weak and null alleles as discussed above. McBean et al. (2014) stated that the BLOODchip® platform was designed to test the European population and BeadChip™ was designed to detect the alleles of the American population, in particular the Africa-American population. However, in Australia, a large population from multi-ethnic backgrounds and a large population with Asian ancestry need a special test. These two platforms will not be a suitable choice for such populations (McBean et al., 2014). In fact, NGS could be a

suitable platform for all populations because it can define and discover new alleles because it depends on genotyping by sequencing. Moreover, it can detect weak and null alleles easily. Such a technique can obtain genotyping by sequencing at an affordable price and can provide automation as well as barcoding for the samples to facilitate dealing with the data analysis and their storage. In the next chapters, NGS will be discussed in further detail with two different approaches: amplicon-based target selection using the Ampliseq™ Custom Panel and LR-PCR approach to sequence the entire genes of the blood group systems.

Chapter 4 : Human Erythrocyte Antigens and Human Platelet Antigens Panel (HEA and HPA Panel)

4.1 Introduction

Haemagglutination still remains the gold standard method that makes blood transfusion feasible (Reid, 2009). It is extremely difficult to interpret the serological data in some cases, such as in patients who suffer from SCD, aplastic anaemia, autoimmune haemolytic anaemia and thalassemia. This is because of the presence of donor cells in the blood stream of the patients or they might be coated with immunoglobulin in a positive DAT (Reid et al., 2000).

Serological testing of a large cohort is highly expensive, time consuming and labour-intensive and some blood groups do not have a special reagent to type them. These issues have increased the demand of BGG to overcome the limited typing by conventional serology (Veldhuisen et al., 2009). Foetal genotyping has been used for routine services when the foetus is at risk of alloimmunisation to red cell antigens or platelets (Reid, 2003). It is highly recommended to genotype the patients who receive multiple units of blood through transfusion such as patients with SCD (Avent et al., 2007). Moreover, BGG can be performed for blood donors and patients to build a database for rare phenotypes. These include low prevalence antigens or any silencing or null mutations that cannot be identified by the standard serological typing or other BGG platforms, such as microarray platforms [see Chapter 3]. In addition, BGG can be feasible with high throughput, through which a large number of blood donors can be tested. NGS has provided high-throughput test in which the screening of a large cohort can take place (Avent et al., 2015).

Platelet transfusions are required in patients who suffer from thrombocytopenia. With the same scenario as blood group antigens, the HPAs need to be matched between the donors and patients. The antibodies to HPAs can be involved in alloimmune platelet

disorders [see section 1.8]. The detection of platelet antibodies by serology in association with genotyping the HPAs establishes a comprehensive diagnosis of alloimmune platelet disorders. Furthermore, the genotyping test makes the typing of the low prevalence platelet antigens feasible due to the scarcity of serologic reagents for these rare antigens (Curtis, 2008). Therefore, genotyping of the HPAs rather than the conventional serology is considered to be the gold standard for platelet typing (Curtis, 2008). Accordingly, NGS can be used to target the HPAs as well as the red cell antigens in a single assay due to its high capacity and throughput.

4.2 Aims

The aim of this project was to design a panel that is based on NGS and capable of genotyping the blood group systems and HPAs to establish it for a high-throughput assay to be used in routine work. The design was for 11 blood group systems including ABO, Rh, Kell, Kidd, Duffy, MNS, Diego, Dombrock, Yt, Colton and Vel and HPAs (1-16). The size of the panel was 31.18 Kb with two pools of primer pairs and the coverage was 99.33%.

The method was based on Ion Ampliseq™ Custom Panel provided by Thermo Fisher Scientific, Paisley, United Kingdom. Most of the blood group antigens arise through SNPs, although other genetic events such as gene deletion and gene rearrangement that give hybrid genes are present (Daniels, 2009). Therefore, the panel targeted the exons, the coding regions, of the blood group genes. Regarding HPAs, PCR products of each antigen were designed by setting the primers within upstream and downstream regions of the SNP of interest.

4.3 Results

Twenty-eight blood samples (Cohort B) were chosen according to the serology. The extracted DNA was initially amplified using two different pools of primers with ultra-high multiplex PCR. Then, the sequencing libraries were made by partial digestion of the primer sequences of each amplicon and ligated to barcoded adaptors [section 2.5.1]. Unlike other protocols of NGS, there was no fragmentation step or multiple purification steps which gives an advantage to the rapid method of BGG by using such a panel. Figure 4.1 illustrates a schematic diagram for the library construction using Ion Ampliseq™ Custom Panel.

In order to proceed with the template preparation and sequencing the samples on Ion PGM™, the sequencing libraries need to be quantified to 100 pM. The Ion Library Quantitation Kit was used for concentration normalisation. After that, the sequencing libraries were immobilised to beads in order to prepare the template for sequencing. These beads had a complementary strand to one of the adaptors and the monoclonal amplification took place followed by enrichment of the positive sequencing templates [section 2.5.3.2]. Finally, the sequencing templates were loaded onto Ion 316™ chip and sequenced on Ion PGM™ [section 2.5.4].

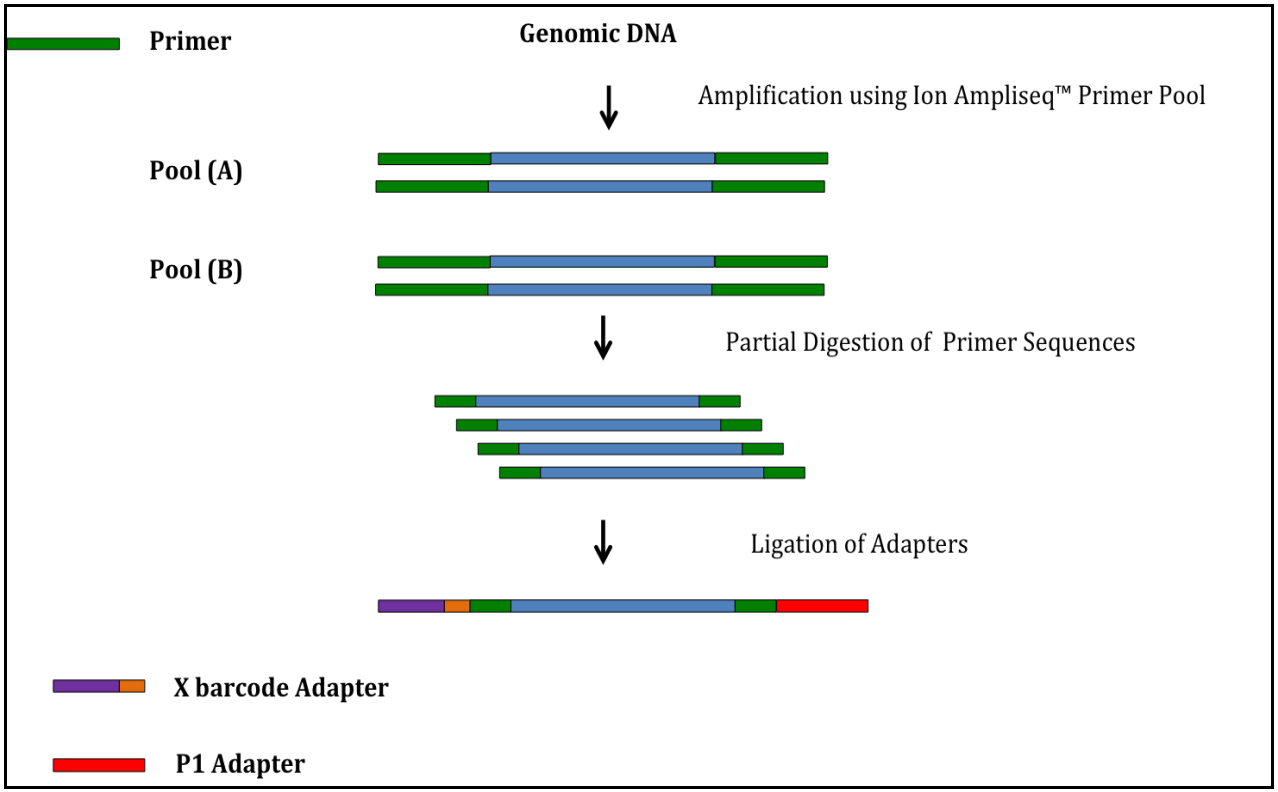


Figure 4.1 Schematic diagram for constructing a sequencing library using Ion Ampliseq™ Custom Panel.

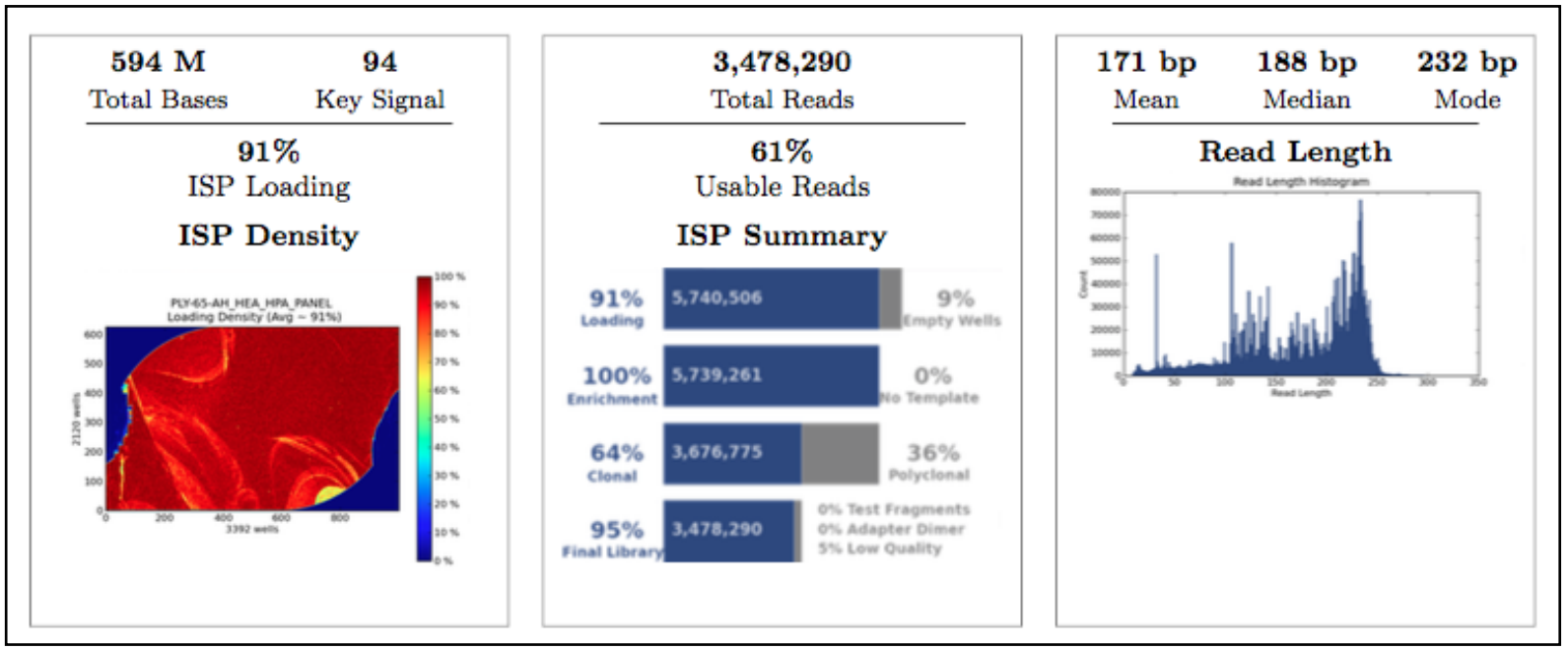
The amplification process of genomic DNA, using two different pools of primer sets (Pool A and Pool B), takes place by extending the annealing primers using a polymerase enzyme. A partial digestion is then carried out and finally the adaptors are ligated to both ends.

4.3.1 Sequencing report

Twenty-eight samples in total were sequenced using the HEA and HPA Panel on the Ion PGM™. A summary of the sequencing report run is demonstrated in **Figure 4.2**. The loading density of the chip for the sequencing run was 91%, in which this was the percentage that the ISPs addressed by the chip wells. The total reads were 3,478,290 and the read length mean was 171 bp [**Figure 4.2]**.

The enrichment of the clonal amplified templates successfully achieved a high percentage of 100%. The clonal amplification was 64%, whereas the polyclonal amplification was 36%. The percentage of the clonal templates shows the template that has a single sequencing library attached to the ISP after performing the emulsion PCR. This is in contrast to the polyclonal templates, where more than a single sequencing library is attached to the ISP [**Figure 4.2]**. The percentage of the usable reads was 61% and the final library was 95% in which 5% was removed due to poor quality.

Two sets of software were used as parts of the Ion PGM™ sequencing report run; the first was well classification and the other was responsible for filtering and trimming. The well classification considered the addressable wells in comparison to the empty wells and was reported as bead loading density. Regarding the filtering and trimming software, the filter removed the polyclonal reads with more than one template that was attached to the beads, reads with low quality or less than 4 bp and a primer dimer with less than 8 bp.



(a) (b) (c)

Figure 4.2 An overview of a report on the sequencing run of the HEA and HPA Panel.

(a) Loading density of ISP that was addressed by the chip wells, which was 91%.
(b) Total of 3,478,290 reads after filtering and trimming.
(c) A histogram of the read length of the sequencing libraries with a mean of 171 bp, on which the y-axis demonstrates the read count, while the x-axis shows the read length in the bp.

4.3.2 Quality Control

4.3.2.1 Per base sequence quality

FastQC software was utilised to assess the quality control of the generated sequences by HEA and HPA Panel. Phred score is software used to calculate the quality of sequencing reads with a special algorithm and it was developed for the HGP (Ewing et al., 1998). A Phred score of 20 gives an accuracy of sequencing reads of 99%, in which the probability of the base being called incorrect is 1 in 100. Generally, any high-throughput sequencing platforms start the run with the best quality reads and then the quality starts to drop when the run progresses.

Figure 4.3 displays the quality score across the position reads. The x-axis represents the position of the reads in the bp, while the y-axis represents the Phred score of quality. The background of the y-axis is divided into three colour areas: green for the best reads, orange for the reasonable reads and red for the low quality reads.

Box and Whisker plots were drawn on every position per bp. The inter-quartile range (25-75%) was represented by yellow boxes, which obviously stopped just following the highest position in bp. The upper and lower whiskers demonstrate 10% and 90% points. The blue line shows the mean of the Phred quality score across the sequencing reads in bp and the central red line demonstrates the median of the quality score.

Regarding the sequencing run of the HEA and HPA Panel, the mean of the quality score started with a high quality score around 32 and then dropped and remained steady at around 27 and at the end of the sequences it dropped again after position in reads 239 bp and remained steady at a score of 20. Accordingly, this quality score represented a high quality and achieved accuracy of higher than 99%.

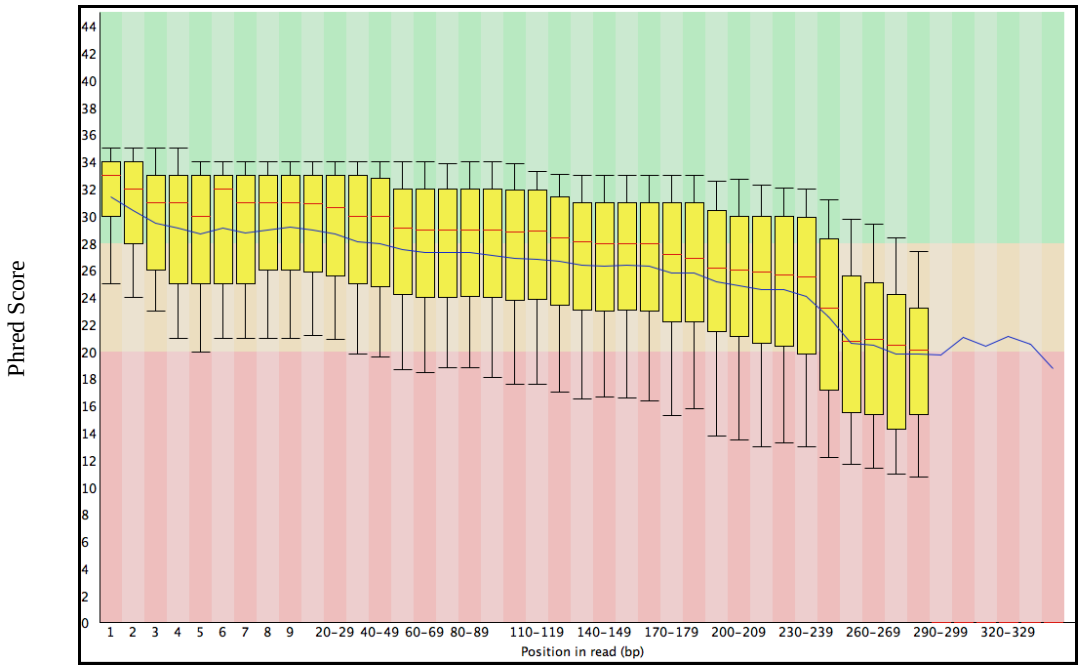


Figure 4.3 Phred quality scores across all the bases for a single sample of the HEA and HPA Panel.

The background of the y-axis demonstrates the Phred score, while the x-axis shows the position in reads in bp. The background of the graph divides the y-axis into three different areas with three different colours; green for the best quality sequencing reads, orange for the reasonable sequencing reads and red for the low quality ones. Box and Whisker plots were drawn on every position per bp. The inter-quartile range (25-75%) was represented by yellow boxes, which stops after the highest position of the reads. The upper and lower whiskers demonstrate 10% and 90% of points. The blue line demonstrates the mean of the quality according to the Phred score across the reads in position, while the central red line shows the median of the quality score. The mean started at a Phred score of 32 and then dropped to 20 at the end of the sequencing run. All the analysed samples had the same Phred score.

4.3.2.2 Per Sequence Quality Scores

FastQC software was also used to evaluate if a subset of one of the sequencing reads represented poor quality reads. Figure 4.4 demonstrates the mean of the quality sequence per reads across the count of the reads according to the Phred score. The x-axis shows the mean of sequencing quality, while the y-axis demonstrates the sequencing counts. The mean quality across the samples was around 27 which means the sequencing reads were of high quality and provided accuracy of over 99%.

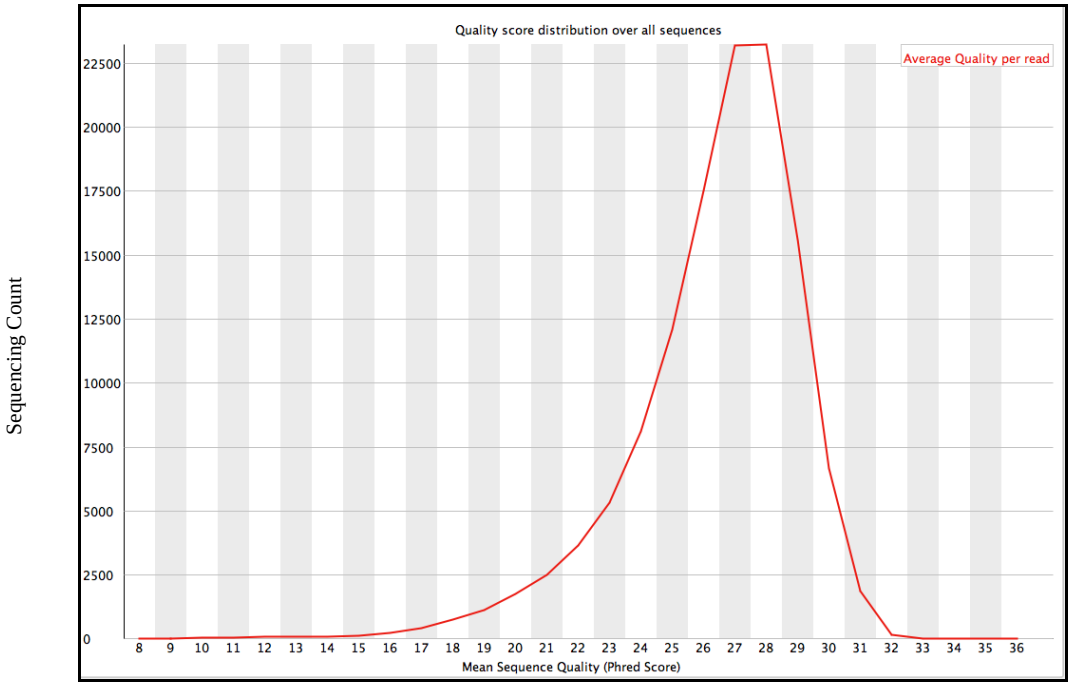


Figure 4.4 Quality scores per sequencing count for a single sample of the HEA and HPA Panel.

The mean sequence quality according to the Phred score is demonstrated across the read counts. FastQC software was used to assess if a subset of the sequencing reads are of poor quality. The results show the quality of most reads was 27 according to the Phred score, which indicates high quality reads with an accuracy of 99%. All the analysed samples had the same Phred score.

4.3.3 Sequencing visualisation

The sequencing data of HEA and HPA Panel were visualised using IGV software Version 2.3.46. The IGV software mapped the generated sequences by Ion PGM™ to the whole human genome (hg19) and then the navigation was performed by looking at the targets of interest in order to assess the genotyping analysis. This software assisted to evaluate the depth of coverage of each target as well as to determine the zygosity. Figure 4.5 shows a homozygous SNP, hemizygous SNP in this case for weak D, 809 T>G (Val270Gly) for a weak D type 1 in exon 6 of the *RHD* gene. The depth of coverage represented for that SNP was 346 reads, as the number of repeated sequences at specific targets. This can appear as multiple grey lines above the reference sequence. Figure 4.6 shows a heterozygous SNP 2108A>C (Tyr703Ser) for the HPA-15 a/15b genotype in *CD109* gene.

4.3.3.1 Alignment obstacles

Mapping quality in IGV software can be seen in colours. In other words, for the good quality reads, the mapping quality score was represented in grey colour as shown in Figure 4.7. On the other hand, the poor quality mapping reads appeared in white colour, as seen in Figure 4.8.

Some areas in the *ABO* and *RHD* genes were observed as a very low depth of coverage. Figure 4.9 shows a poor mapping quality score in exon 8 of the *RHD* gene, as observed in white colour, as well as a very low depth of coverage. In addition to that, there were only two sequencing reads in amino acid positions (377-379), which were missed in some samples within exon 8. Two missed regions in exon 7 of the *ABO* gene were observed in amino acid positions (202-225) and (347-354). Figure 4.10 shows the missed regions in exon 7 of the *ABO* gene.

An interesting finding was observed regarding to the exon 1 of the *RHCE* gene. There was an allelic imbalance with a ratio of 77% to the reference nucleotide in comparison

of 23% to the changing nucleotide. Figure 4.11 shows the allelic imbalance found in *RHCE* gene.

Furthermore, in the MNS blood group system, some samples observed a low depth of coverage and missed regions. Figure 4.12 illustrates a comparison between two samples in the *GYPA* gene of the MNS blood group system. There was no difference in the nucleotides between *GYPA*N* allele, which matches the reference gene, and the antithetical *GYPA*M* allele regarding the amino acid position 24. Figure 4.12 a and Figure 4.12 b show the amino acid position 24 for *GYPA*N* allele and *GYPA*M* allele, respectively. Issues related to this sample include low coverage depth, low quality mapping and a missed region. Figure 4.13 illustrates an output from the IGV software demonstrating two samples with A and B phenotypes observed a nucleotide insertion. The purple sign indicates the insertion of a single nucleotide (G) at amino acid number 88. This is because the reference allele used was *O* allele. In this allele, a deletion of G nucleotide at position 261 causes a frameshift in the reading frame leading to 118 stop codon (88fs118stop).

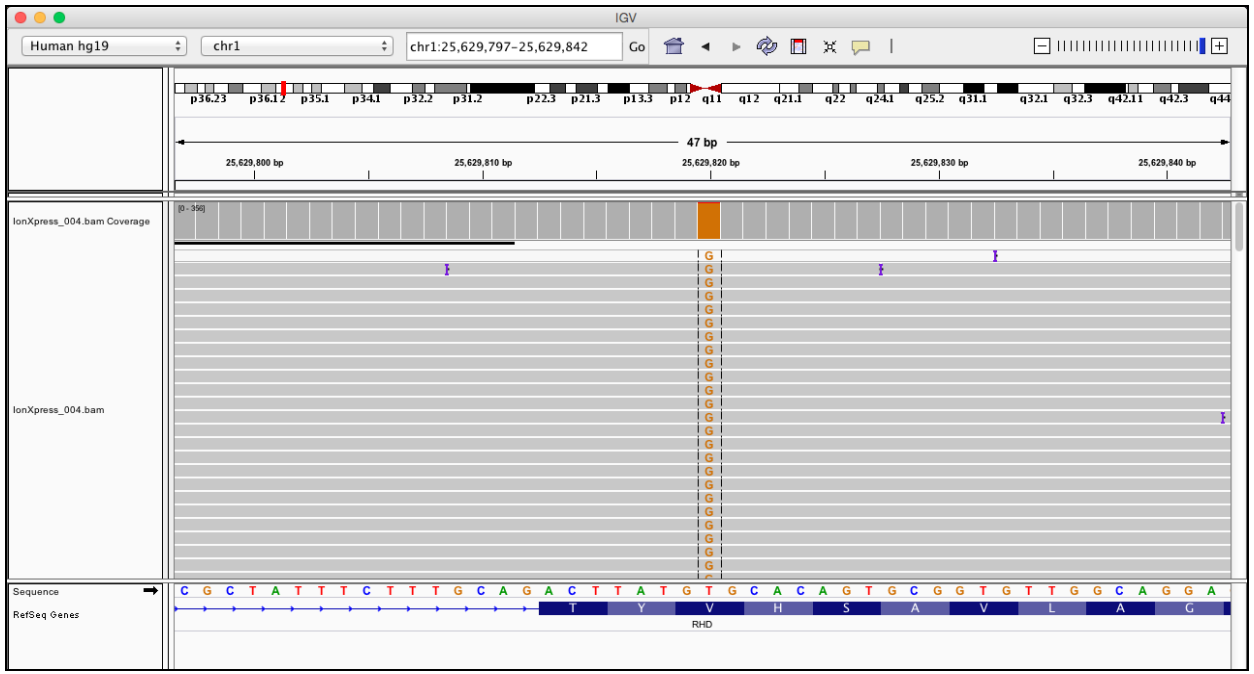


Figure 4.5 A homozygous SNP (hemizygous in this case for weak D sample).

The SNP is 809 T>G (Val270Gly) in exon 6 of the *RHD* gene. The depth of coverage of this SNP was 346 and it appeared as multiple grey lines above the reference gene (in blue) of *RHD*. The zygosity shows very clearly that T was replaced by G in all rows as indicated by a full box with a brown colour at the top. Output from IGV software demonstrates the visualisation of the sequencing data. Chromosomal locations are shown on the top and the reference gene in blue at the bottom.

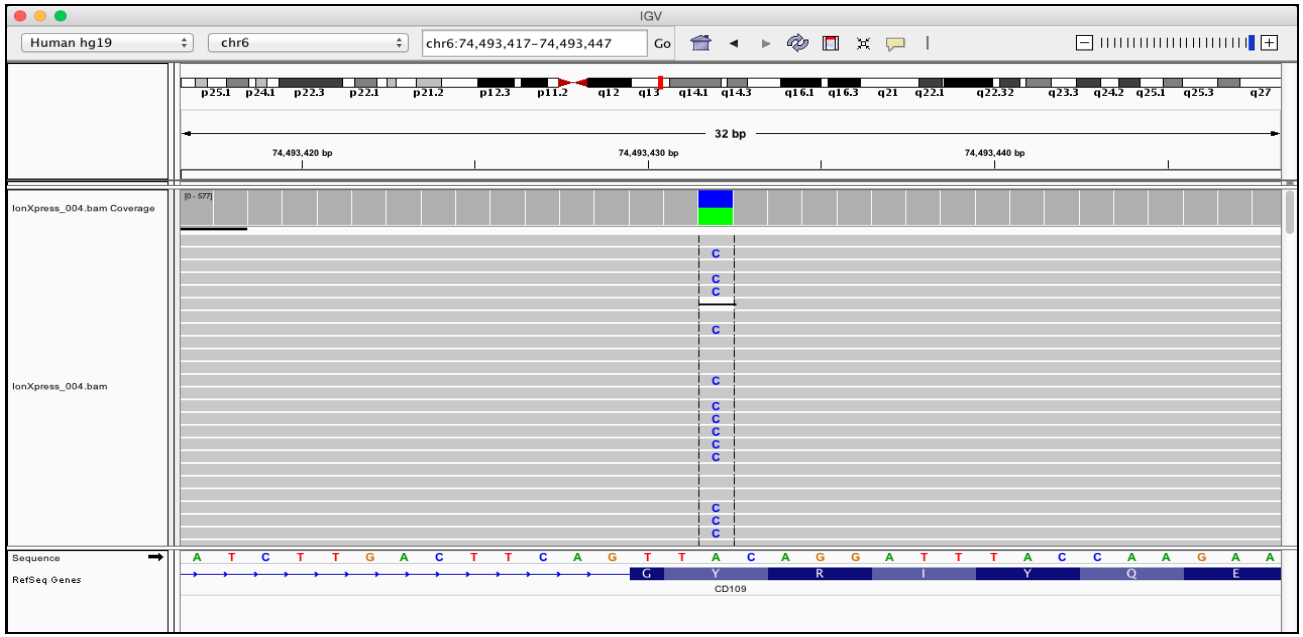


Figure 4.6 A heterozygous SNP for HPA-15a/15b genotype in the *CD109* gene.

Two colours (green for the reference nucleotide "A" and blue for the mutated nucleotide "C") are shown to indicate the heterozygous SNP. The percentage of this heterozygous SNP was 50% for the reference nucleotide and 50% for the other nucleotide. Further to that change in the nucleotide, the amino acid was substituted from tyrosine to serine at a position of 703. Output from IGV software demonstrates the visualisation of the sequencing data. Chromosomal locations are shown on the top and the reference gene in blue at the bottom.

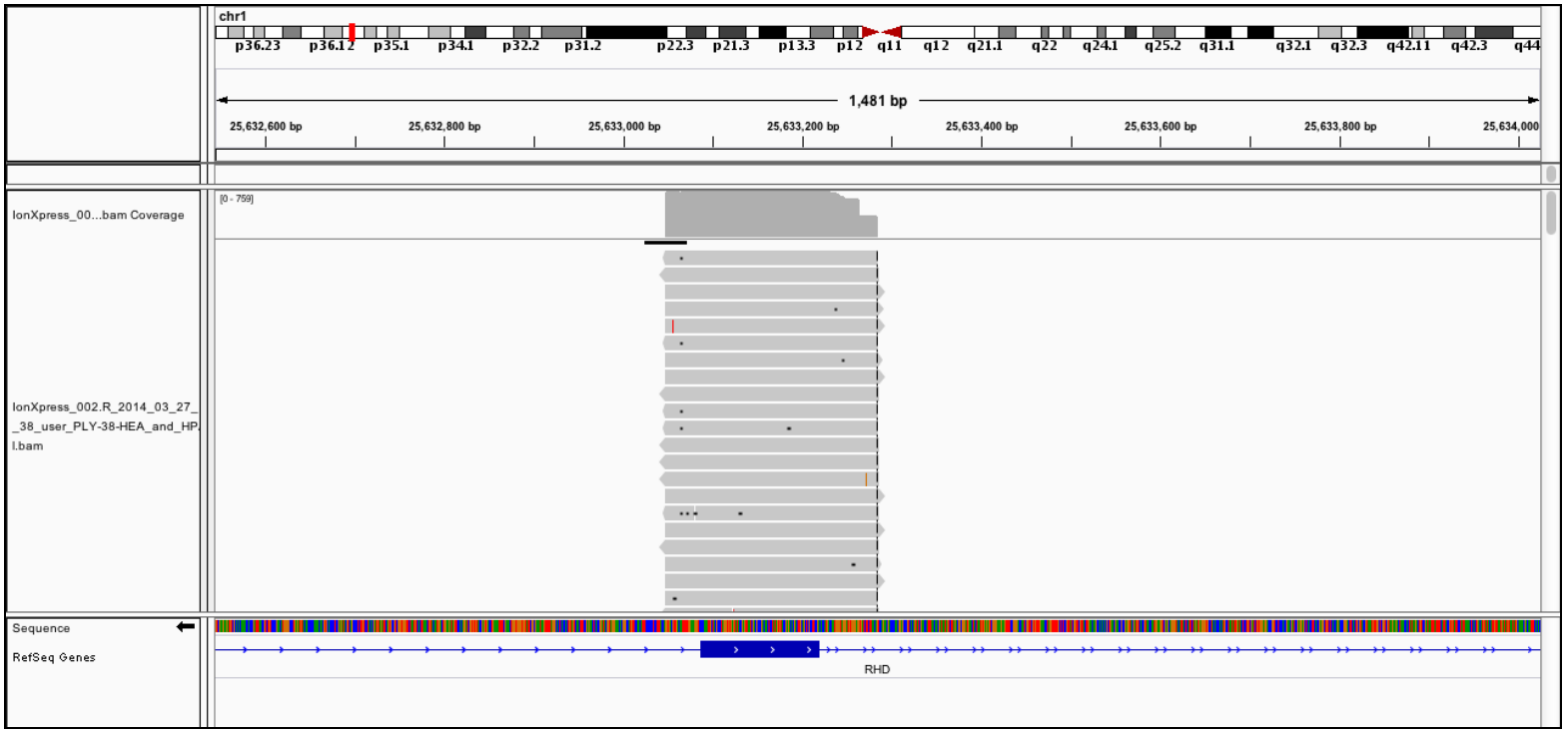


Figure 4.7 Good mapping quality for the sequencing reads

The screenshot shows a grey colour for the good quality mapping of the repeated sequencing reads. This screenshot demonstrates exon 7 of the *RHD* gene. Output from IGV software demonstrates the visualisation of the sequencing data. Chromosomal locations are shown on the top and the reference gene in blue at the bottom.

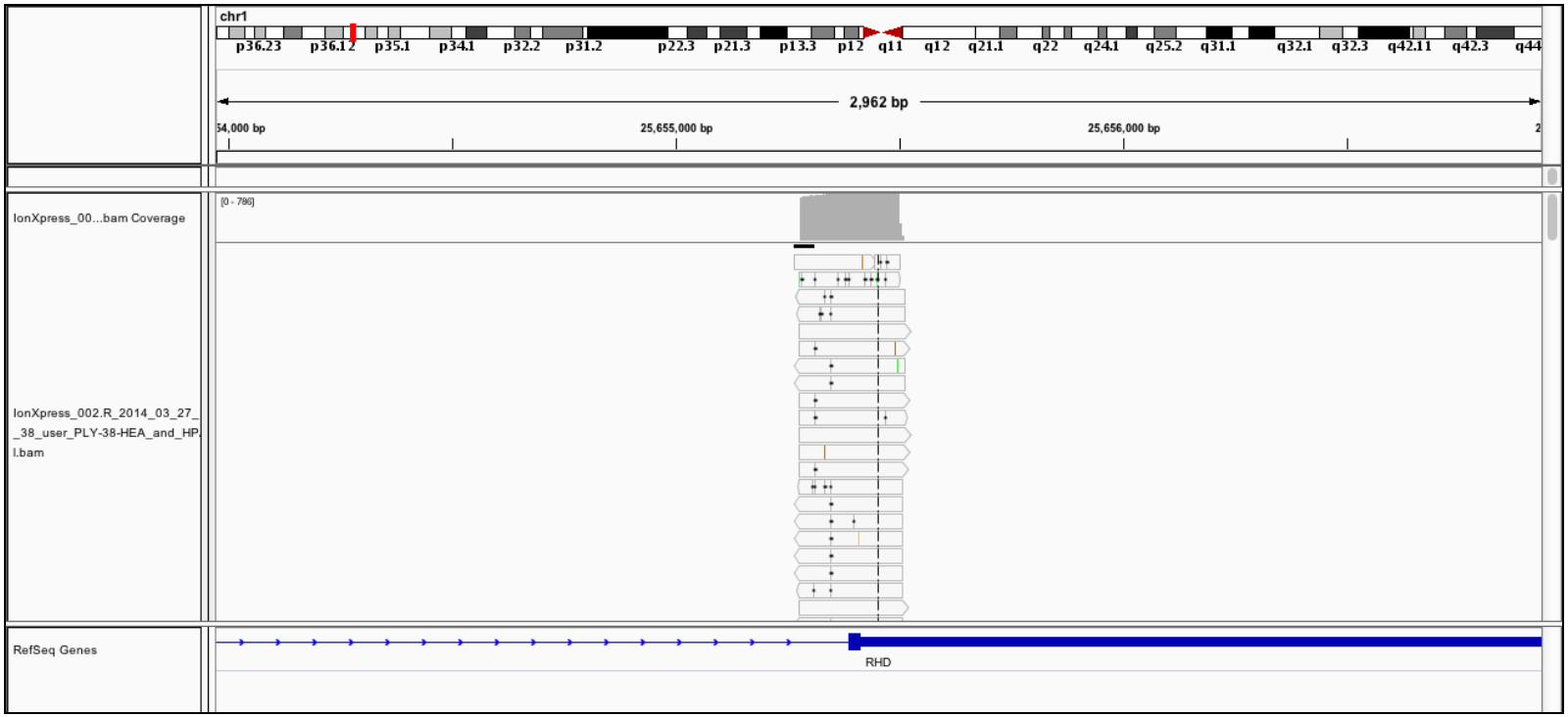


Figure 4.8 Poor mapping quality in exon 10 of the *RHD* gene.

The poor mapping reads can be seen in white colour. Output from IGV software demonstrates the visualisation of the sequencing data. Chromosomal locations are shown on the top and the reference gene in blue at the bottom.

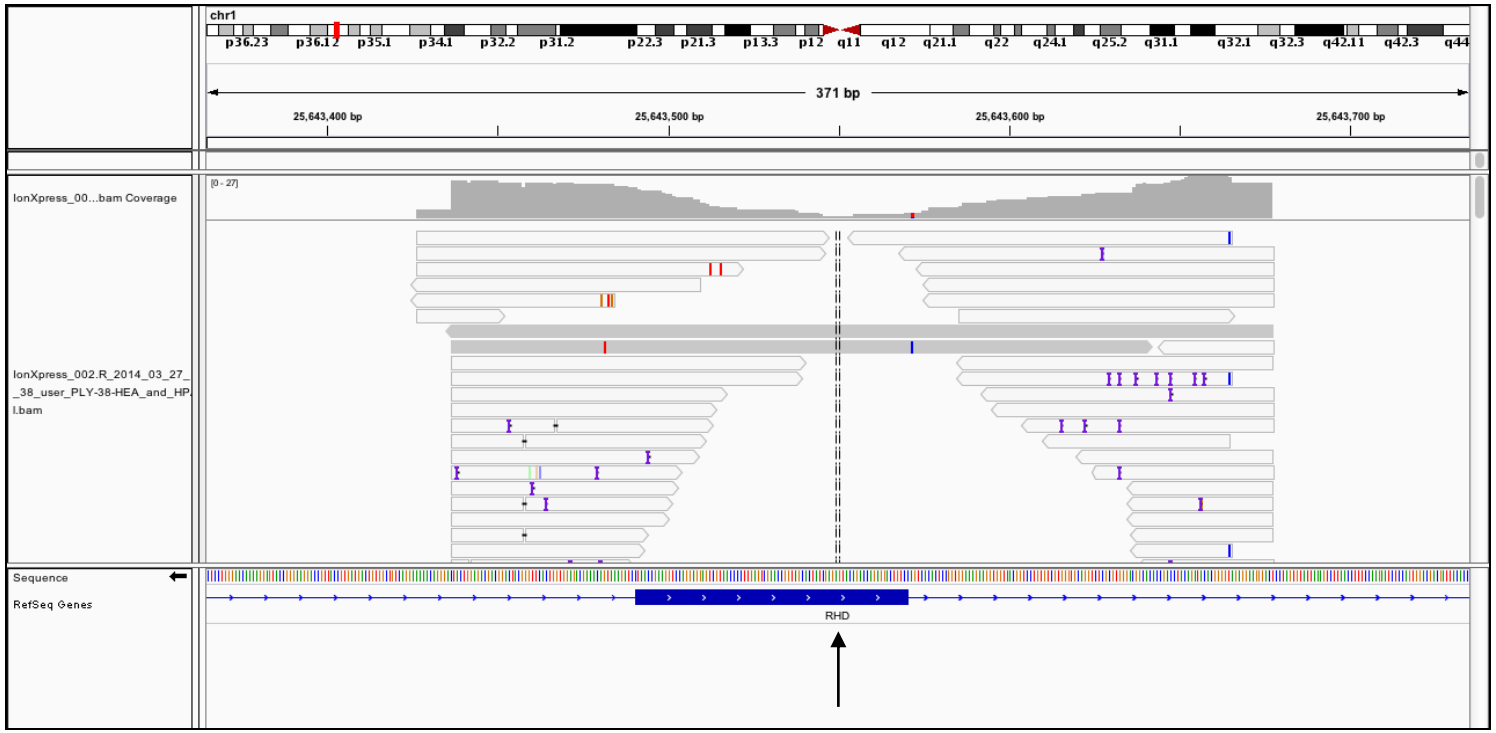


Figure 4.9 Poor mapping quality for the sequencing reads for exon 8 of the *RHD* gene.

The poor mapping reads can be seen in white colour. Furthermore, a low depth of coverage in exon 8 of the *RHD* gene in which there were less than 16 reads and only around two reads (in grey, indicated by a black arrow) at amino acid 379 which makes genotyping for the *RHD*DAU* allele impossible. Output from IGV software demonstrates the visualisation of the sequencing data. Chromosomal locations are shown on the top and the reference gene in blue at the bottom.

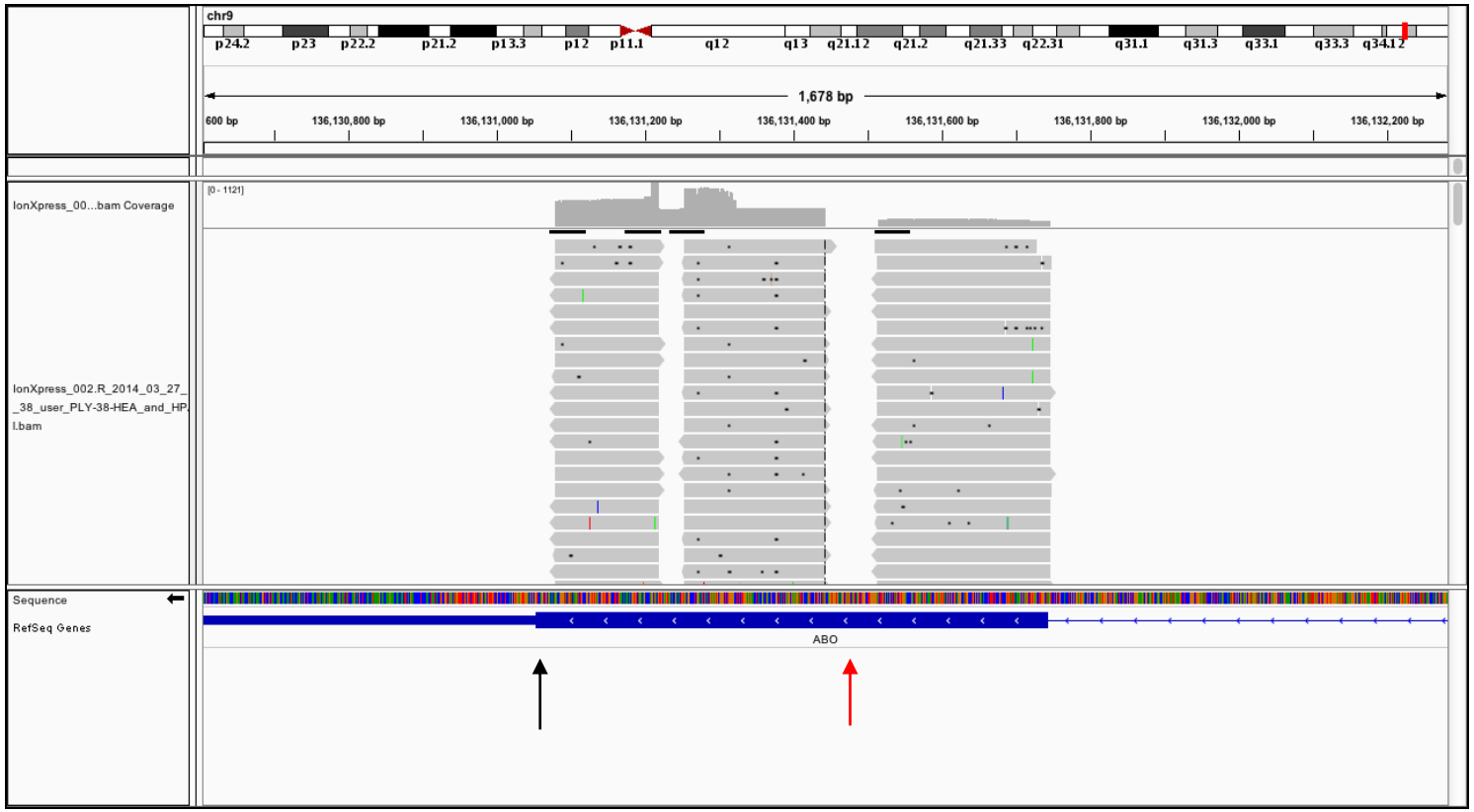


Figure 4.10 Missed regions within the exon 7 of the *ABO* gene.

These regions include amino acid positions (202-225) indicated by the red arrow and (347-354) indicated by the black arrow. Output from IGV software demonstrates the visualisation of the sequencing data. Chromosomal locations are shown on the top and the reference gene in blue at the bottom.

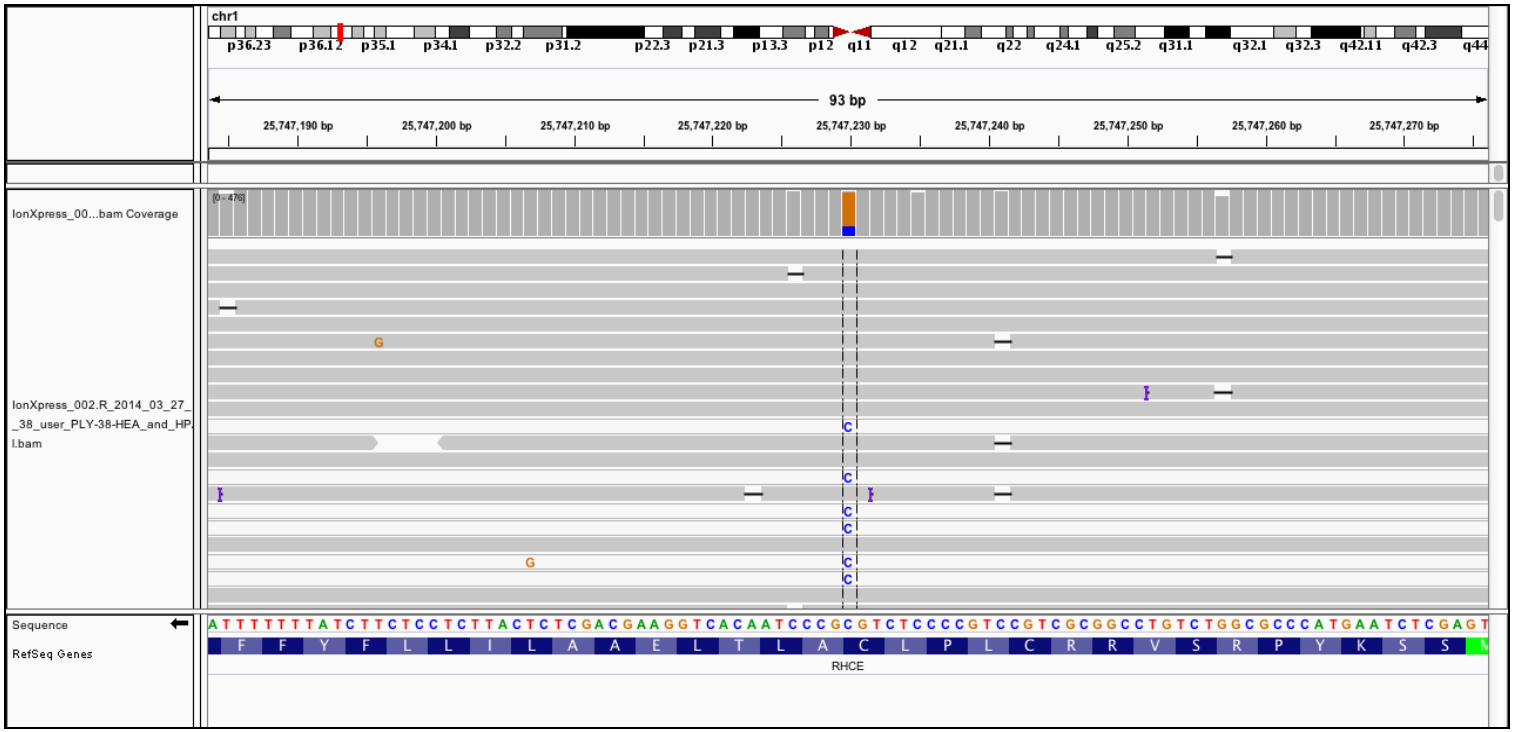


Figure 4.11 Allelic imbalance in exon 1 of the *RHCE* gene.

The reference nucleotide (C) represents 77% while the other nucleotide (G) represents 23%. Some reads with poor mapping quality can be shown in white colour. Output from IGV software demonstrates the visualisation of the sequencing data. Chromosomal locations are shown on the top and the reference gene in blue at the bottom.

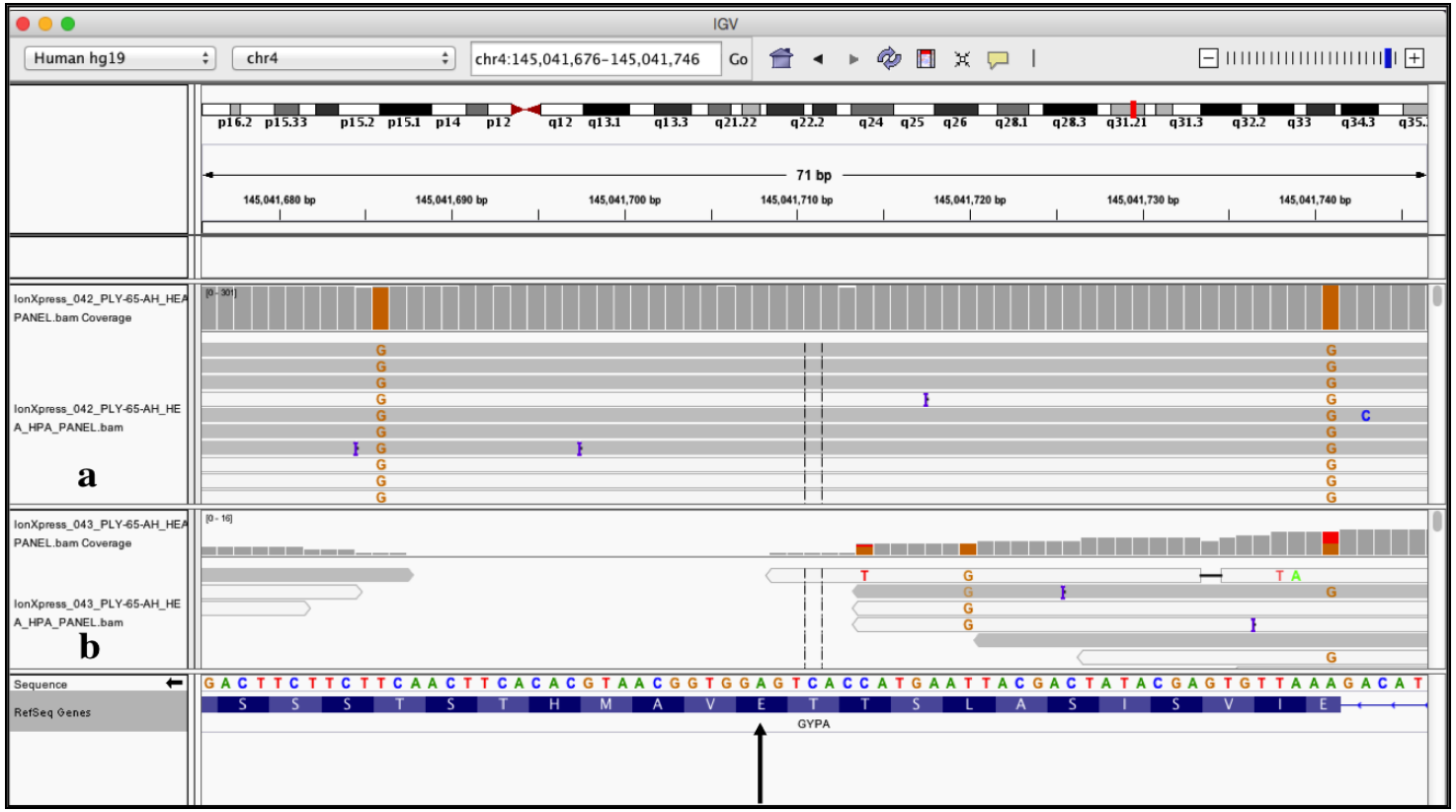


Figure 4.12 A comparison between two samples of the *GYPA* gene in the MNS blood group system.

There was no difference in the nucleotides between *GYPA*N* allele (a), which matches the reference gene, and the antithetical *GYPA*M* allele regarding the amino acid position 24 (b). The amino acid position 24 is indicated with a black arrow. Issues related to this sample include low coverage depth, low quality mapping and a missed region. This may be because non-specific primers were used for the amplification of both homologous genes, *GYPA* and *GYPB*. Output from IGV software demonstrates the visualisation of the sequencing data. Chromosomal locations are shown on the top and the reference gene in blue at the bottom.

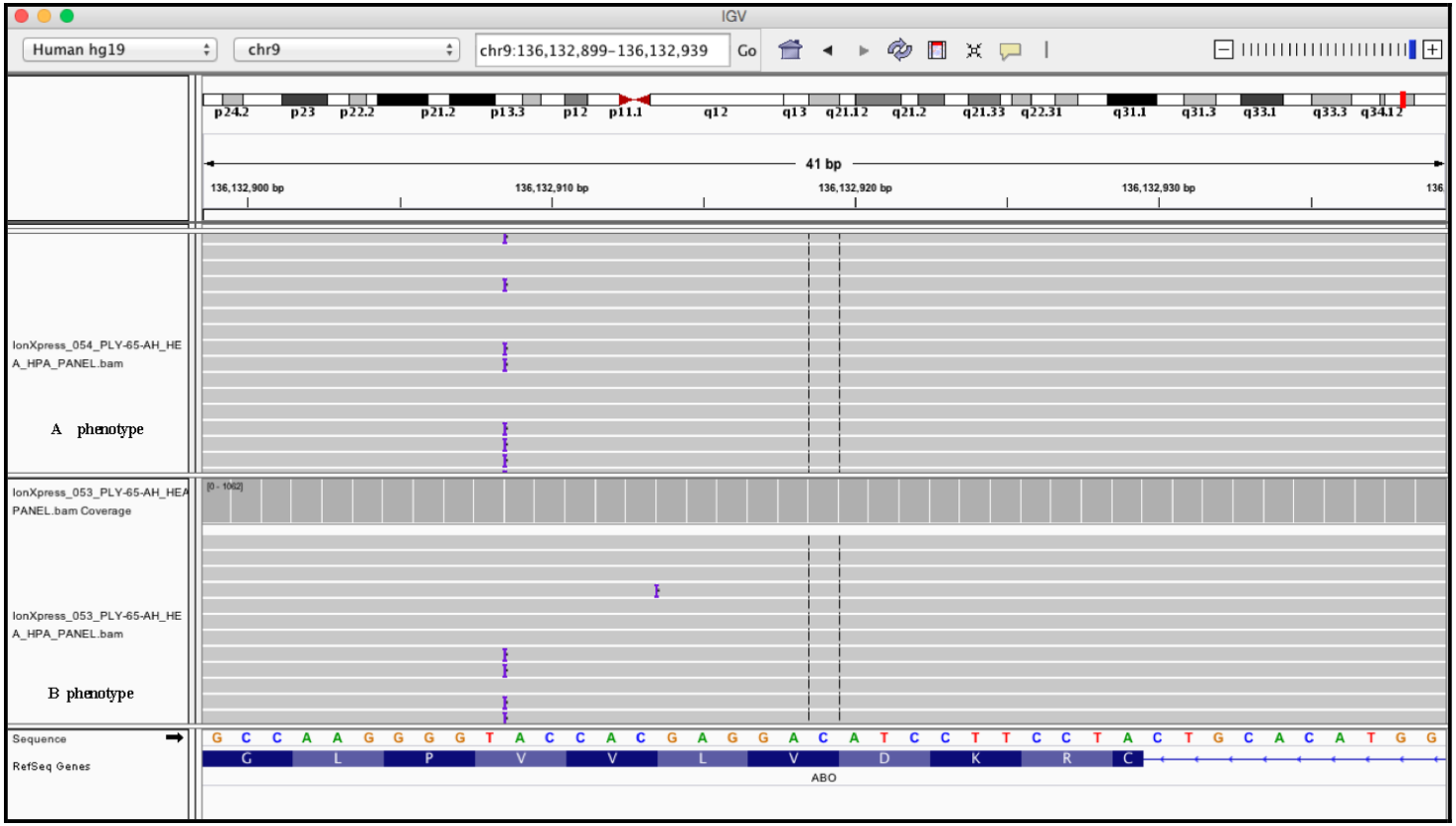


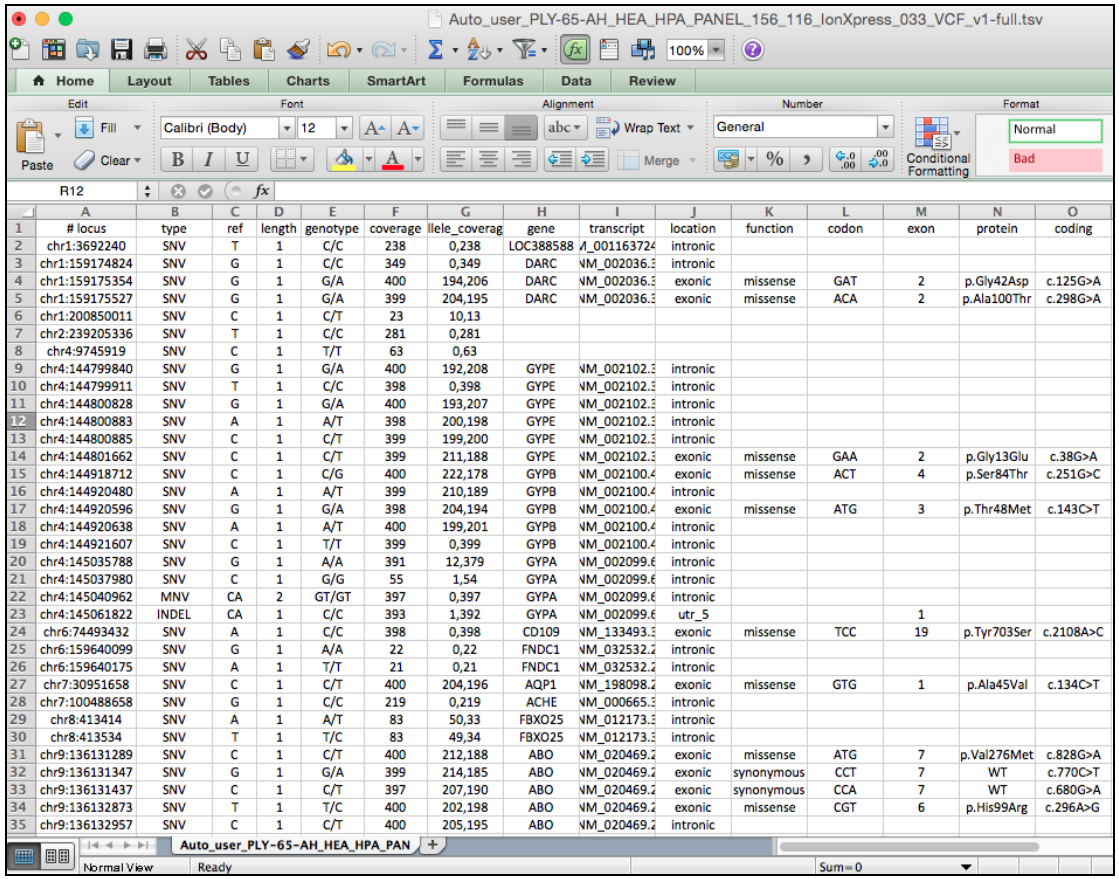
Figure 4.13 A nucleotide insertion with *ABO* gene in two samples with A and B phenotypes.

The purple sign indicates the insertion of a single nucleotide (G) at amino acid number 88. This is because the reference allele used is *O* allele. In this allele, a deletion of G nucleotide at position 261 causes a frameshift in the reading frame leading to 118 stop codon (88fs118stop). Output from IGV software demonstrates the visualisation of the sequencing data. Chromosomal locations are shown on the top and the reference gene in blue at the bottom.

4.3.4 Variant analysis using Ion Reporter™ software

In order to start the genotyping analysis, the VCF files were analysed using the Ion Reporter™ software. This software was used to annotate the sequencing data in order to genotype the blood group antigens and HPAs resulting from the HEA and HPA Panel. The data of the Ion Reporter™ can be exported in different formats including .pdf file format and in Excel sheets.

Figure 4.12 shows a screenshot of the report analysis for one sample of the HEA and HPA Panel that was exported as an Excel sheet. The Ion Reporter™ facilitated the analysis by giving the annotated information including chromosomal location, type of variant and whether it was a single nucleotide variant (SNV) or indels, gene, the reference nucleotide, the changing nucleotide, depth of coverage, the transcript used in the analysis based on the NCBI database, the location of the variant and whether it was (intronic, synonymous, exonic) SNPs, codon, exon number of that variant, amino acid substitution and position of the nucleotide change. The genotyping results are presented in the next section.



Auto_user_PLY-65-AH_HEA_HPA_PANEL_156_116_IonXpress_033_VCF_v1-full.tsv

# locus	type	ref	length	genotype	coverage	allele_coverag	gene	transcript	location	function	codon	exon	protein	coding
chr1:3692240	SNV	T	1	C/C	238	0,238	LOC388588	/_001163724	intronic					
chr1:159174824	SNV	G	1	C/C	349	0,349	DARC	NM_002036.3	intronic					
chr1:159175354	SNV	G	1	G/A	400	194,206	DARC	NM_002036.3	exonic	missense	GAT	2	p.Gly42Asp	c.125G>A
chr1:159175527	SNV	G	1	G/A	399	204,195	DARC	NM_002036.3	exonic	missense	ACA	2	p.Ala100Thr	c.298G>A
chr1:200850011	SNV	C	1	C/T	23	10,13								
chr2:239205336	SNV	T	1	C/C	281	0,281								
chr4:9745919	SNV	C	1	T/T	63	0,63								
chr4:144799840	SNV	G	1	G/A	400	192,208	GYPE	NM_002102.3	intronic					
chr4:144799911	SNV	T	1	C/C	398	0,398	GYPE	NM_002102.3	intronic					
chr4:144800828	SNV	G	1	G/A	400	193,207	GYPE	NM_002102.3	intronic					
chr4:144800883	SNV	A	1	A/T	398	200,198	GYPE	NM_002102.3	intronic					
chr4:144800885	SNV	C	1	C/T	399	199,200	GYPE	NM_002102.3	intronic					
chr4:144801662	SNV	C	1	C/T	399	211,188	GYPE	NM_002102.3	exonic	missense	GAA	2	p.Gly13Glu	c.38G>A
chr4:144918712	SNV	C	1	C/G	400	222,178	GYPB	NM_002100.4	exonic	missense	ACT	4	p.Ser84Thr	c.251G>C
chr4:144920480	SNV	A	1	A/T	399	210,189	GYPB	NM_002100.4	intronic					
chr4:144920596	SNV	G	1	G/A	398	204,194	GYPB	NM_002100.4	exonic	missense	ATG	3	p.Thr48Met	c.143C>T
chr4:144920638	SNV	A	1	A/T	400	199,201	GYPB	NM_002100.4	intronic					
chr4:144921607	SNV	C	1	T/T	399	0,399	GYPB	NM_002100.4	intronic					
chr4:145035788	SNV	G	1	A/A	391	12,379	GYPA	NM_002099.6	intronic					
chr4:145037980	SNV	C	1	G/G	55	1,54	GYPA	NM_002099.6	intronic					
chr4:145040962	MNV	CA	2	GT/GT	397	0,397	GYPA	NM_002099.6	intronic					
chr4:145061822	INDEL	CA	1	C/C	393	1,392	GYPA	NM_002099.6	utr_5			1		
chr6:74493432	SNV	A	1	C/C	398	0,398	CD109	NM_133493.3	exonic	missense	TCC	19	p.Tyr703Ser	c.2108A>C
chr6:159640099	SNV	G	1	A/A	22	0,22	FNDC1	NM_032532.2	intronic					
chr6:159640175	SNV	A	1	T/T	21	0,21	FNDC1	NM_032532.2	intronic					
chr7:30951658	SNV	C	1	C/T	400	204,196	AQP1	NM_198098.2	exonic	missense	GTG	1	p.Ala45Val	c.134C>T
chr7:100488658	SNV	G	1	C/C	219	0,219	ACHE	NM_000665.3	intronic					
chr8:413414	SNV	A	1	A/T	83	50,33	FBXO25	NM_012173.3	intronic					
chr8:413534	SNV	T	1	T/C	83	49,34	FBXO25	NM_012173.3	intronic					
chr9:136131289	SNV	C	1	C/T	400	212,188	ABO	NM_020469.2	exonic	missense	ATG	7	p.Val276Met	c.828G>A
chr9:136131347	SNV	G	1	G/A	399	214,185	ABO	NM_020469.2	exonic	synonymous	CCT	7	WT	c.770C>T
chr9:136131437	SNV	C	1	C/T	397	207,190	ABO	NM_020469.2	exonic	synonymous	CCA	7	WT	c.680G>A
chr9:136132873	SNV	T	1	T/C	400	202,198	ABO	NM_020469.2	exonic	missense	CGT	6	p.His99Arg	c.296A>G
chr9:136132957	SNV	C	1	C/T	400	205,195	ABO	NM_020469.2	intronic					

Figure 4.14 An Excel sheet for the genotyping analysis report using the Ion Reporter™ software.

Each variant was given by its locus at the chromosomal location, the type of variant including SNV or indels, the reference nucleotide (ref), the length of the variant, the genotype result, the depth of the coverage to that variant (coverage), allele coverage, gene, transcript which was based on the NCBI database, the location of the variant and whether it was intronic or within the exons, codon, exon number where the variant was located, protein changed or synonymous and the position of the coding nucleotide. This analysis represents a single sample, which was sequenced by the HEA and HPA Panel on the Ion PGM™.

4.3.5 Genotyping results of HEA and HPA Panel

Twenty-eight samples were examined for 11 blood group systems and 16 HPAs. The blood samples underwent serology testing at NHSBT Filton for the following blood groups ABO, Rh, MNS, P1, Lutheran, Kell, Lewis, Duffy, Kidd and Sickle cell status. Table 4.1 shows that the serological testing was performed for all 28 samples.

4.3.5.1 Genotyping of ABO Blood group system

Out of the 28 samples, 10 samples were serotyped as A phenotype, four samples as B phenotype and 14 samples as O phenotype. Table 4.2 lists the genotyping results for the ABO blood group system. Samples from all three phenotypes observed the following SNPs 106G>T (Val36Phe) in exon 3, 188G>A (Arg63His) in exon 4, 220C>T (Pro74Ser) in exon 5 and 261delG 88fs118stop in exon 6. These SNPs are related to O phenotype, which means the samples were *A/O*, *B/O* and *O/O* alleles.

Regarding the O phenotype, two homozygous samples and five heterozygous samples observed 829G>A (Val227Met) in exon 7, which is a SNP found to be associated with other SNPs in rare Japanese allele (Reid et al., 2012). In addition, a single sample observed a heterozygous SNP of 594C>T (Arg198Cys) in exon 7.

Regarding the A phenotype, in A_1 phenotype there was no association of any mutation in the exons, apart from a single sample which observed a heterozygous SNP 467C>T (Pro156Leu) and that SNP is found mainly in Asians (Reid et al., 2012).

Regarding the B phenotype, all four samples observed the four main SNPs that are associated with the B phenotype in exon 7. These included one homozygous and three heterozygous for 526C>G (Arg176Gly), one homozygous and three heterozygous for 703G>A (Gly235Ser), one homozygous and three heterozygous for 796C>A (Leu266Met) and one homozygous and three heterozygous for 803G>C (Gly268Ala). A SNP 296G>A (His99Arg) was found in B phenotype (two homozygous and two heterozygous) and in O phenotype (two homozygous and seven heterozygous).

4.3.5.2 Genotyping of RHD gene

Eighteen samples were RhD-negative samples, while 10 samples were serologically typed as weak D. Table 4.3 demonstrates the NGS genotyping for the weak D samples within the *RHD* gene including nucleotide changes and amino acid substitutions. Regarding the weak D samples, which were typed by serology, four samples were observed to be weak D type 1 (*RHD*01W.01)* 809T>G (Val270Gly), four samples of weak D type 2 *(RHD*01W.02)* 1154G>C in exon 9 (Gly385Ala) and one sample of DAR3.1 weak partial D 4.0 *(RHD*DAR3.01*). The latter had three nucleotides 602C>G (Thr201Arg) in exon 4, 667T>G (Phe223Val) in exon 5 and 819G>A (silent) in exon 6. Interestingly, one of the weak D samples, which was typed as weak D by conventional serology appeared to be an RhD-positive sample. In other words, no mutations were detected in any of the exons of the *RHD* gene.

4.3.5.3 Genotyping of other blood group systems

The antigens of the other blood group systems (RhCE, Kell, Kidd, Duffy, MNS, Diego, Dombrock, Colton. Yt and Vel) were genotyped and the predicted phenotypes are listed in Table 4.4. Regarding *RHCE* gene, four samples were genotyped as *RHCE*C* and were found to be heterozygous 48G>C (Trp16Cys) in exon 1. There was no difference observed between *RHCE*C* and *RHCE*c* alleles regarding exon 2. In other words, there were no SNPs related to the *RHCE*C* allele and it was only shown Pro103 in exon 2.

Five samples were genotyped as *RHCE*E*, in which they were heterozygous in exon 5 of the *RHCE* gene for 676C>G (Ala226Pro). The genotyping data by NGS for the *RHCE*C* and *RHCE*E* alleles confirmed the phenotypes provided by the serology. The following allele (*RHCE*ceVS.04* allele) was found in a single sample as 48G>C (Trp16Cys) in exon 1, 733C>G (Leu245Val) in exon 5 and 1025C>T (Thr342Ile) in

exon 7. This allele predicts partial e phenotype, V and VS antigens, which is found mainly in Africans (Reid et al., 2012).

Kell and Kidd blood groups

For the Kell blood group system, *KEL*02* (k), *KEL*02.04* (Kp^b) and *KEL*02.07* (Js^b) were found in all samples and confirmed the information obtained by serology. Regarding the Kidd blood group system, 10 samples were found to be heterozygous for *JK*A/B*, eight samples were homozygous for *JK*B* and five samples were homozygous for *JK*A*. Jk($a+^w$) were predicted in four samples and were found to be heterozygous with *JK*B*. The associated mutation with the *JK*01W.01* allele, encoding the Jk($a+^w$) phenotype, was 130G>A (Glu44Lys). The rest of the samples that performed by NGS were the same as the serology.

Duffy blood group

Ten samples were found to be homozygous for *FY*B* allele in which they observed 125G>A (Gly42Asp) in exon 2, while a single sample was homozygous for *FY*A* 125A>G (Asp42Gly). Seventeen samples were heterozygous for both *FY*A/B*. A SNP 298G>A (Ala100Thr) was found in seven samples which is associated with the *FY*02M.01* allele, in which six of them were in association with heterozygous *FY*A/B* and one with *FY*B* allele.

MNS blood group

Regarding the MNS blood group system, eight samples were found as *GYPA*N* allele by NGS which contradicted with the serology data, where they were typed as M antigen. Only four samples were genotyped with *GYPA*M* allele by NGS although they were observed in a low depth of coverage of five sequencing reads and some of them missed bp around amino acids (positions 23-31). Two non-valid samples were found in which no alleles could be identified. The rest of the samples matched the serology although most of them did not have the serological information. Twelve samples were

genotyped by NGS as *GYPB*s/s* with the SNP 143T>C (Met48Thr), two samples were found to be *GYPB*S/S* with the SNP 143C>T (Thr48Met) and 14 samples were heterozygous for both alleles *GYPB*S/s*. Out of the total of the 28 samples, two samples were typed by serology as S+ and they were revealed by NGS to be heterozygous *GYPB*S/s*. Interestingly, there was one sample found by NGS as *GYPB*S/S*, which confirmed the serotyping as S+.

Diego blood group

Regarding the Diego blood group system, 27 samples were homozygous for 2561T>C (Leu854Pro) in exon 19 of the *SLC4A1* gene which observes *DI*B* allele. On the other hand, only one sample was heterozygous *DI*A/B*.

Dombrock blood group

For the Dombrock blood group system, *DO*A* was found in a single sample as 793G>A (Asp265Asn) in exon 2 of the *ART4* gene. Eleven samples were homozygous for the *DO*B* allele observing 793A>G (Asn265Asp). One sample was heterozygous which had both alleles *DO*A/B*. All the samples were found the *DO*02.04* allele 323G<T (Val108Gly) in exon 2 and *DO*02.05* allele as 350T>C (Ile117Thr) in exon 2. These two alleles predict Hy and Jo^a antigens, respectively.

Colton blood group

For the Colton blood group system, 25 samples observed the *CO*A* allele and were found to be homozygous for the SNP 134T>C (Val45Ala). Only a single sample was found to be *CO*B*, which observed a homozygous SNP 134C>T (Ala45Val). In addition, two samples were heterozygous for both alleles *CO*A/B*.

Yt and Vel blood groups

Regarding the Yt blood group system, 26 samples were typed by NGS as *YT*A* which was observed as 1057A>C (Asn353His) in exon 2 of the *ACHE* gene. Furthermore, two

samples were heterozygous *YT*A/B*, observing the SNP 1057C>A, (His353Asn). For the Vel blood group system, all samples observed the Vel allele.

4.3.5.4 Genotyping of human platelet antigens

The HPAs were also tested using the HEA and HPA Panel. Table 4.5 demonstrates the phenotypes of HPAs 1 to 16 as well as the common form of the antigen (a) and the less common form of the platelet antigen (b). HPA-1 observed nine heterozygous for HPA-1a/1b and 19 homozygous for HPA-1a/1a. In HPA-2, five samples were heterozygous HPA-2a/2b and 23 were homozygous HPA-2a/2a. Regarding HPA-3, 14 individuals were homozygous for HPA-3a/3a, 10 homozygous for HPA-3b/3b and four samples were heterozygous HPA-3a/3b. In HPA-5, 25 samples were homozygous for HPA-5a/5a and three samples were heterozygous for HPA-5a/b. Regarding HPA-6, 22 samples were homozygous for HPA-6a/6a, one sample was homozygous for HPA-6b/6b and five samples were heterozygous for HPA-6a/6b. In HPA-15, six individuals were homozygous for HPA-15a/15a, 13 were homozygous for HPA-15b/15b and nine samples found to be HPA-15a/15b. All the samples were homozygous for HPA-4a/4a, HPA-7a/7a, HPA-8a/8a, HPA-9a/9a, HPA-10a/10a, HPA-11a/11a, HPA-12a/12a, HPA-13a/13a, HPA-14a/14a and HPA-16a/16a.

4.3.5.5 Novel alleles detected by the HEA and HPA Panel

Four novel alleles were identified by the HEA and HPA Panel. One was in the *RHCE* gene and three were in the *KEL* gene. Table 4.6 lists the four new SNPs investigated by the panel. The first SNP was detected in the *RHCE* in exon 2 heterozygous 208C>T (Arg70Trp). Initially, this SNP was assigned incorrectly to the *RHD* gene with an allelic imbalance ratio of (60:40) as demonstrated in Figure 4.15. This issue was resolved by amplification of both *RHD* and *RHCE* genes using LR-PCR approach following by sequencing on Ion PGM™ [see Chapter 6].

Regarding the *KEL* gene, the first SNP was found in sample 8 and was heterozygous 331G>A (Ala111Thr) in exon 4. The result of this SNP is shown in Figure 4.16. Sample 26 had another two novel SNPs. The first SNP was heterozygous 1907C>T (Ala636Val) in exon 17 as shown in Figure 4.17. The second SNP was in exon 19 and was found to be heterozygous 2165T>C (Leu722Pro) as demonstrated in Figure 4.18.

Table 4.1 The serological results provided by NHSBT Filton for the 28 samples.

These samples were tested by NGS using the HEA and HPA Panel. The samples indicated in blue were typed as weak D by serology, while the rest of the samples were RhD-negative.

	ABO	Rh	D	C	E	c	e	C^w	M	N	S	s	P1	Lu^a	Lu^b	K	k	Kp^a	Kp^b	Le^a	Le^b	Fy^a	Fy^b	Jk^a	Jk^b	A1
1	O+	R_0r	+	−	−	+	+	−	+	+	+	+	+	−	+	−		−	+		−	−	+	+	+	
2	A+	R_1r	+	+	−	+	+	−	+		+					−								+	+	
3	A+	R_2r	+	−	+	+	+	−	+															+	+	
4	O+	R_1r	+	+	−	+	+	−																		
5	O−	rr	−	−	−	+	+	−			+	+			+		+		+			+	+	+	−	
6	A−	rr	−	−	−	+	+	−	−	+	+		−		+	−	+			−	+			+	+	+
7	A−	rr	−	−	−	+	+	−	+		+	+				−						+	+	+	+	
8	A−	rr	−	−	−	+	+	−								−						+		+	+	
9	O−	rr	−	−	−	+	+	−	+		−	+	−		+	−			+	−		+		+	+	
10	A−	rr	−	−	−	+	+	−	+		−	+										+	+	+	+	
11	O−	rr	−	−	−	+	+		+						+	−			+					+	+	
12	O−	rr	−	−	−	+	+								+	−			+							
13	B−	rr	−	−	−	+	+	−	+		+					−								+	+	
14	O−	rr	−	−	−	+	+	−	−	+	−	+	+	−	+	−	+	−		−	+	−	+	+	−	
15	B−	rr	−	−	−	+	+	−	+		+					−								+	+	
16	O−	rr	−	−	−	+	+	−	+		+				+	−			+					+		
17	O−	rr	−	−	−	+	+									−										
18	O−	rr	−	−	−	+	+								+	−			+							
19	O−	rr	−	−	−	+	+								+	−			+							
20	B−	rr	−	−	−	+	+	−	+		−					−										
21	O−	rr	−	−	−	+	+	−	−		−	+		−	+	−		−	+	−		+	+	+	−	
22	O−	rr	−	−	−	+	+	−	+		+				+	−			+			+	+	+	+	
23	O+	R_2r	+	−	+	+	+	−	+		+	−	+			−						−	+	−	+	
24	A+	R_1r	+	+	−	+	+	−	−		−		+							−				−	+	
25	A+	R_1r	+	+	−	+	+									−								+	−	
26	B+	R_2r	+	−	+	+	+									−										
27	A+	R_2r''	+	−	+	+	+									−										
28	A+	R_2r	+	−	+	+	+	−	+		−					−								−	+	

Table 4.2 The genotyping results of the ABO blood group system obtained by the HEA and HPA Panel.

Phenotype	No. of samples	Exon 3	Exon 4	Exon 5	Exon 6	Exon 7	Possible Genotype
A	10	106G>T Val36Phe (9 hom[§])	188G>A Arg63His (10 hom)	220C>T Pro74Ser (10 hom)	261delG 88fs118stop (2 hom & 6 het[§])	467C>T Pro156Leu (1 het)	*A/O*
B	4	106G>T Val36Phe (2 hom & 2 het)	188G>A Arg63His (2 hom & 1 het)	220C>T Pro74Ser (2 hom & 2 het)	296G>A His99Arg (2 hom & 2 het) 261delG 88fs118stop (1 hom & 2 het)	829G>A Val227Met (2 het) 526C>G Arg176Gly (1 hom & 3 het) 703G>A Gly235Ser (1 hom & 3 het) 796C>A Leu266Met (1 hom & 3 het) 803G>C Gly268Ala (1 hom & 3 het)	3 samples of *B/O* 1 sample of *B/B*
O	14	106G>T Val36Phe (7 hom & 7 het)	188G>A Arg63His (8 hom & 6 het)	220C>T Pro74Ser (8 hom & 5 het)	296G>A His99Arg (2 hom & 7 het) 261delG 88fs118stop (1 het)	829G>A Val227Met (2 hom & 5 het) 594C>T Arg198Cys (1 het) 526C>G Arg176Gly (1 het)	*O/O*

[§]Hom and het refer to homozygous and heterozygous, respectively. fs= frameshift.

Table 4.3 Genotyping results using NGS of weak D samples obtained by the HEA and HPA Panel.

No. of Samples	Exon	Nucleotides	Amino acid	NGS Genotyping
1	4	602C>G	Thr201Arg	DAR3.1 weak partial D 4.0
	5	667T>G	Phe223Val	*(RHD*DAR3.01*)
	6	819G>A	Silent	
4	6	809T>G	Val270Gly	Weak D type 1 (*RHD*01W.01)*
4	9	1154G>C	Gly385Ala	Weak D type 2 *(RHD*01W.02)*

A single sample observed no related mutations within all the exons.

Table 4.4 The predicted phenotypes of NGS genotyping results for other blood groups by the HEA and HPA Panel.

The blood groups include RhCE, Kell, Kidd, Duffy, MNS, Diego, Dombrock, Colton, Yt and Vel blood group systems by the HEA and HPA Panel. § = low coverage depth, +w = weak form of the antigen, NV = non-valid. The blank fields are considered as negative results for clarity.

		RhCE							Kell						Kidd		Duffy			MNS					Diego		Dombrock				Colton		Yt		Vel
#	Serology	C	c	E	e	C^x	C^w	VS	K	k	Kp^a	Kp^b	Js^a	Js^b	Jk^a	Jk^b	Fy^a	Fy^b	Fy^x	M	N	S	s	U	Di^a	Di^b	Do^a	Do^b	Hy	Jo^a	Co^a	Co^b	Yt^a	Yt^b	Vel
1	R_0r		+		+			+		+		+		+	+	+		+			+	+	+	+			+	+	+	+	+			+		+
2	R_1r	NV	+		+					+		+		+	+w	+	+	+			+	+		+			+	+	+	+	+			+		+
3	R_2r		+	+	+					+		+		+	+	+	+	+		+		+	+	+			+	+	+	+	+			+		+
4	R_1r	NV	+		+					+		+		+	+	+	+	+			+		+	+			+		+	+	+			+		+
5	rr		+		+					+		+		+	+		+	+			+	+	+	+			+	+	+	+	+		+		+	+	+
6	rr		+		+					+		+		+	+	+	+	+			+	+	+	+			+	+	+	+	+			+		+	+	+
7	rr		+		+					+		+		+	+w	+	+	+			+	+	+	+			+		+	+	+		+		+		+
8	rr		+		+					+		+		+	+	+	+	+			+	+	+	+			+		+	+	+			+		+		+
9	rr		+		+					+		+		+	+	+		+			+		+	+			+		+	+	+			+		+		+
10	rr		+		+					+		+		+	+w	+	+	+			+		+	+			+	+		+	+			+		+		+
11	rr		+		+					+		+		+	+	+	+	+			+	+	+	+			+	+	+	+	+			+		+		+
12	rr		+		+					+		+		+	+w	+		+		NV	NV	+	+	+			+		+	+	+			+		+		+
13	rr		+		+					+		+		+	+	+		+		+§	+	+	+	+			+		+	+	+			+		+		+
14	rr		+		+					+		+		+		+	+	+			+		+	+	+	+	+	+	+	+	+			+		+		+
15	rr		+		+					+		+		+	+		+	+		+§		+	+	+			+	+	+	+	+		+		+		+
16	rr		+		+					+		+		+		+		+			+	+	+	+			+		+	+	+			+		+		+
17	rr		+		+					+		+		+	+			+			+		+	+			+	+	+	+	+			+		+		+
18	rr		+		+					+		+		+		+		+			+	+	+	+			+		+	+	+			+		+		+
19	rr		+		+					+		+		+		+	+	+			+	+	+	+			+	+	+	+	+			+		+		+
20	rr		+		+					+		+		+	+	+	+	+			+		+	+			+	+	+	+	+			+		+		+
21	rr		+		+					+		+		+	+			+			+		+	+			+		+	+	+			+		+		+
22	rr		+		+					+		+		+	+	+	+	+		NV	NV	+	+	+			+	+	+	+	+			+		+		+
23	R_2r		+	+	+					+		+		+		+		+			+	+		+			+	+	+	+	+			+		+		+
24	R_1r	NV	+		+					+		+		+		+	+	+			+		+	+			+		+	+	+			+		+		+
25	R_1r	NV	+		+					+		+		+	+		+	+		+§			+	+			+		+	+	+			+		+		+
26	R_2r		+	+	+					+		+		+	+	+	+	+		+§			+	+			+	+	+	+	+			+		+		+
27	R_2r''		+	+	+					+		+		+		+		+			+		+	+			+	+	+	+	+			+		+		+
28	R_2r		+	+	+					+		+		+		+	+				+		+	+			+	+	+	+	+			+		+		+

Table 4.5 The predicted phenotypes of the HPAs from the genotyping results using HEA and HPA Panel.

HPA																																
Sample	HPA-1		HPA-2		HPA-3		HPA-4		HPA-5		HPA-6		HPA-7		HPA-8		HPA-9		HPA-10		HPA-11		HPA-12		HPA-13		HPA-14		HPA-15		HPA-16	
	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b
1	+		+		+	+	+		+			+	+		+		+		+		+		+		+		+		+		+	
2	+	+	+			+	+		+			+	+		+		+		+		+		+		+		+		+	+	+	
3	+		+		+		+		+	+		+	+		+		+		+		+		+		+		+			+	+	
4	+		+		+	+	+		+			+	+		+		+		+		+		+		+		+		+	+	+	
5	+		+		+	+	+		+			+	+		+		+		+		+		+		+		+			+	+	
6	+		+			+	+		+			+	+		+		+		+		+		+		+		+		+		+	
7	+		+		+		+		+	+		+	+		+		+		+		+		+		+		+			+	+	
8	+		+		+		+		+			+	+		+		+		+		+		+		+		+		+	+	+	
9	+		+		+		+		+			+	+		+		+		+		+		+		+		+		+	+	+	
10	+	+	+		+	+	+		+	+		+	+		+		+		+		+		+		+		+		+	+	+	
11	+	+	+		+		+		+			+	+		+		+		+		+		+		+		+			+	+	
12	+		+	+	+		+		+			+	+		+		+		+		+		+		+		+		+	+	+	
13	+		+		+	+	+		+			+	+		+		+		+		+		+		+		+			+	+	
14	+		+		+		+		+			+	+		+		+		+		+		+		+		+		+		+	
15	+		+	+	+	+	+		+			+	+		+		+		+		+		+		+		+		+		+	
16	+	+	+		+	+	+		+			+	+		+		+		+		+		+		+		+		+	+	+	
17	+	+	+		+		+		+			+	+		+		+		+		+		+		+		+			+	+	
18	+		+	+	+		+		+			+	+		+		+		+		+		+		+		+			+	+	
19	+	+	+		+	+	+		+			+	+		+		+		+		+		+		+		+		+	+	+	
20	+		+	+	+		+		+			+	+		+		+		+		+		+		+		+			+	+	
21	+		+		+		+		+			+	+		+		+		+		+		+		+		+		+		+	
22	+		+		+		+		+			+	+		+		+		+		+		+		+		+			+	+	
23	+	+	+			+	+		+		+	+	+		+		+		+		+		+		+		+			+	+	
24	+	+	+		+		+		+		+	+	+		+		+		+		+		+		+		+			+	+	
25	+	+	+		+	+	+		+		+		+		+		+		+		+		+		+		+			+	+	
26	+		+			+	+		+		+	+	+		+		+		+		+		+		+		+		+	+	+	
27	+		+	+	+		+		+		+	+	+		+		+		+		+		+		+		+		+		+	
28	+		+		+	+	+		+		+	+	+		+		+		+		+		+		+		+			+	+	

The blank fields are considered as negative results for clarity.

Table 4.6 The novel alleles detected by the HEA and HPA Panel.

Sample number	Gene	Exon	Nucleotides	Zygosity	Amino acid
4	*RHCE*	2	208C>T	Heterozygous	Arg70Trp
8	*KEL*	4	331G>A	Heterozygous	Ala111Thr
26	*KEL*	17	1907C>T	Heterozygous	Ala636Val
26	*KEL*	19	2165T>C	Heterozygous	Leu722Pro

Two novel SNPs in the *KEL* gene were found in a single sample, which was sample 26.

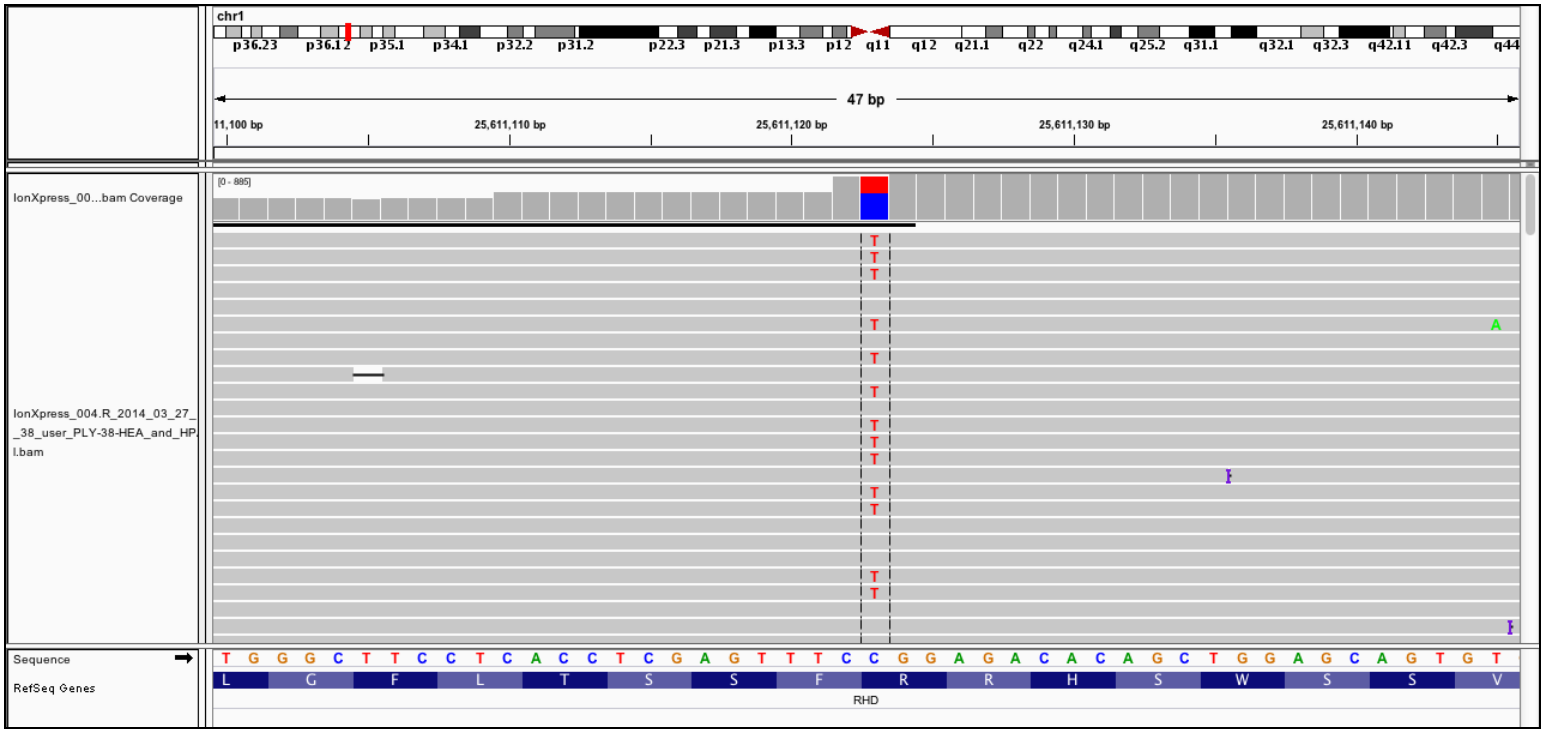


Figure 4.15 A novel SNP that was misaligned to the *RHD* gene instead of the *RHCE* gene with an allelic imbalance ratio of (60:40).

The SNP is 208C>T (Arg70Trp) in exon 2 of the *RHCE* gene [see Chapter 6]. Output from IGV software demonstrates the visualisation of the sequencing data. Chromosomal locations are shown on the top and the reference gene in blue at the bottom.

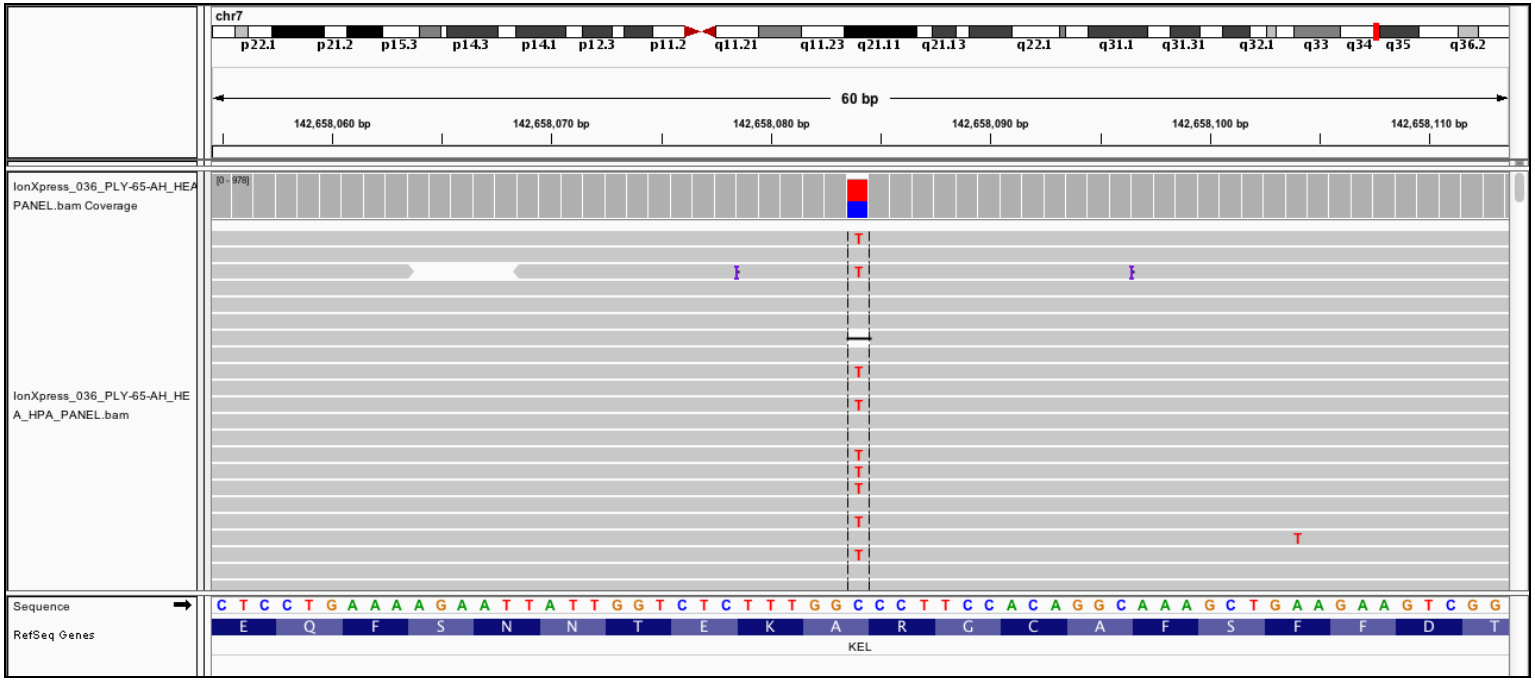


Figure 4.16 A novel SNP found in exon 4 of the *KEL* gene.

The SNP was heterozygous SNP 331G>A (Ala111Thr). The *KEL* gene is an anti-sense strand, therefore the SNP was shown as C>T on the reference gene instead of G>A. Output from IGV software demonstrates the visualisation of the sequencing data. Chromosomal locations are shown on the top and the reference gene in blue at the bottom.

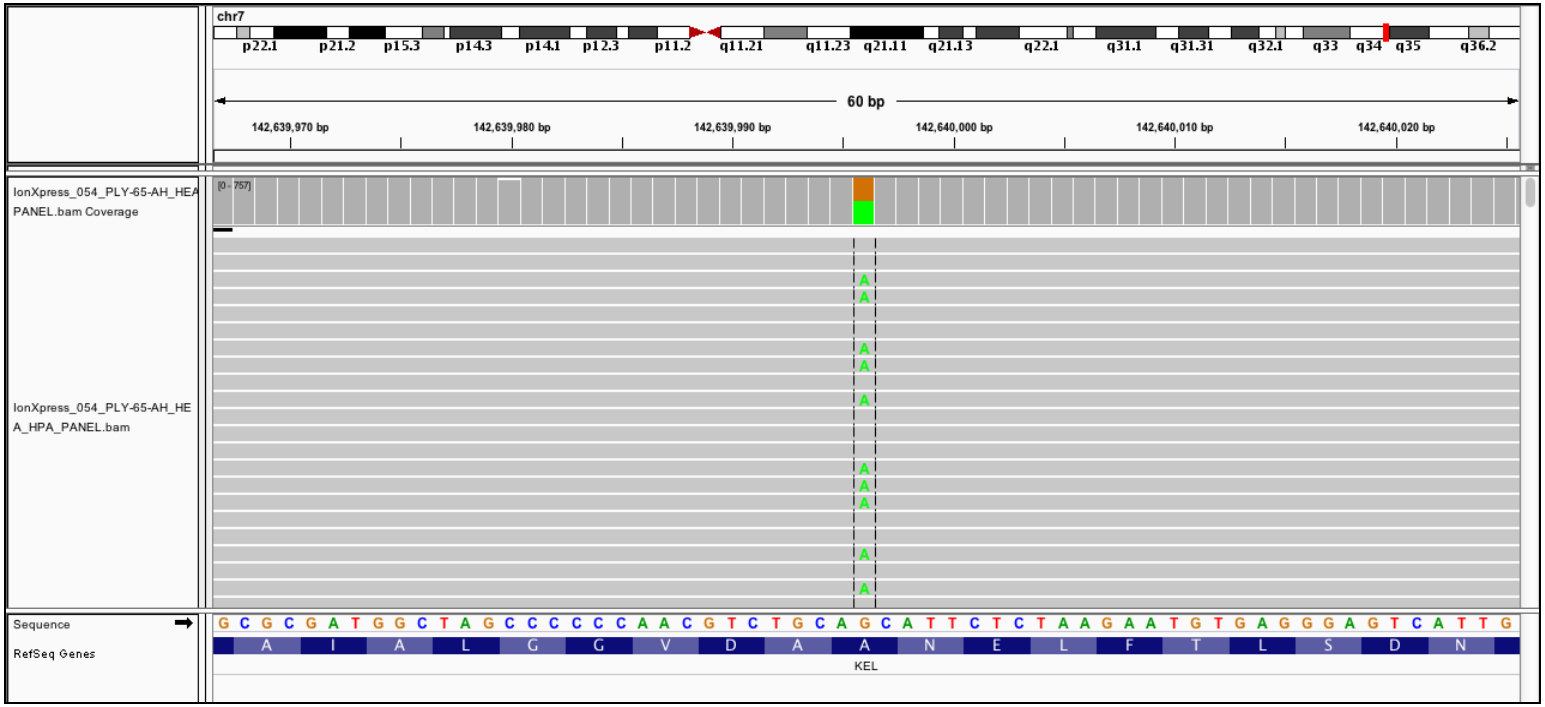


Figure 4.17 A novel SNP found in exon 17 of the *KEL* gene.

The SNP was heterozygous 1907C>T (Ala636Val). The *KEL* gene is an anti-sense strand, therefore the SNP was shown as G>A on the reference gene instead of C>T. Output from IGV software demonstrates the visualisation of the sequencing data. Chromosomal locations are shown on the top and the reference gene in blue at the bottom.

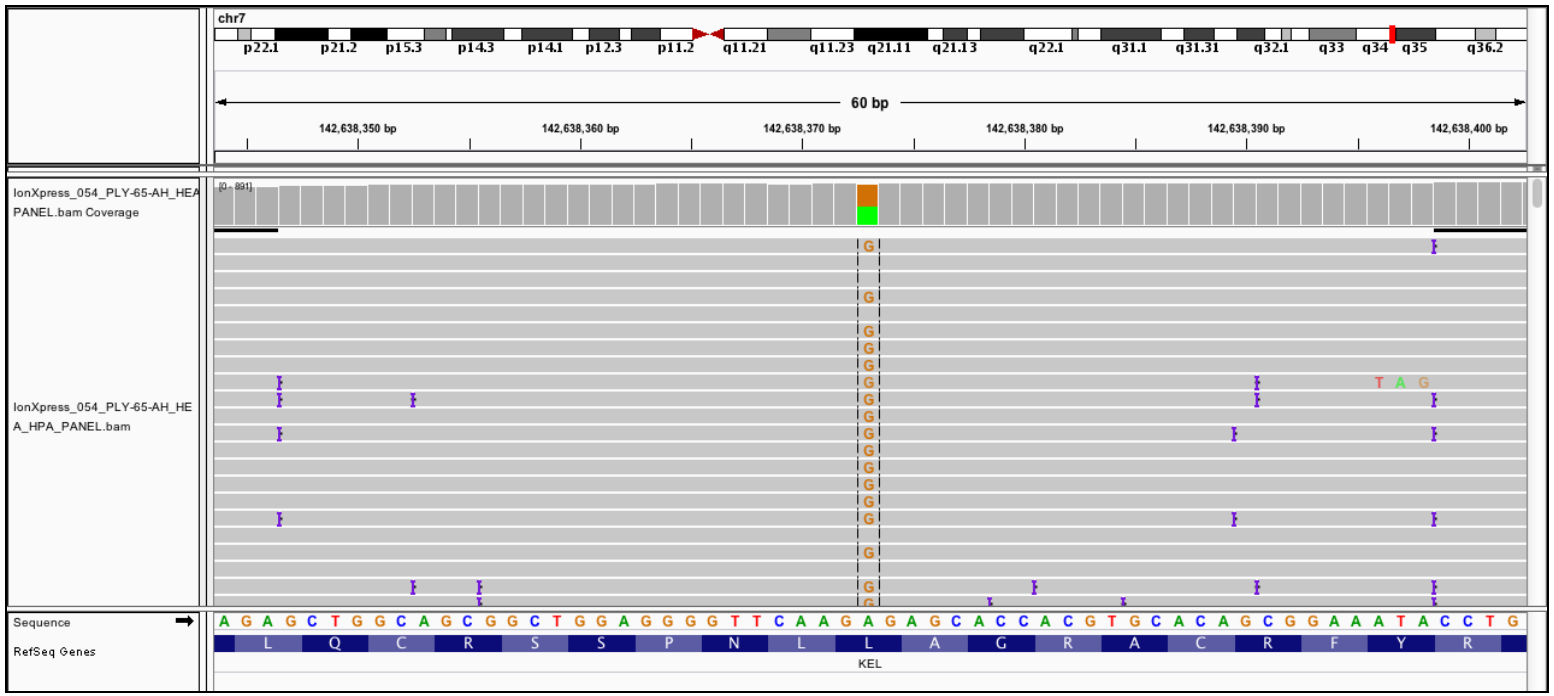


Figure 4.18 A novel SNP found in exon 19 of the *KEL* gene.

The SNP was heterozygous SNP 2165T>C (Leu722Pro). The *KEL* gene is an anti-sense strand, therefore the SNP was shown as A>G on the reference gene instead of T>C. Output from IGV software demonstrates the visualisation of the sequencing data. Chromosomal locations are shown on the top and the reference gene in blue at the bottom.

4.4 Discussion

The HEA and HPA Panel is a customised panel that was based on the Ion Ampliseq™ Custom Panel. The panel can be used to test 11 blood group systems (ABO, Rh, Kell, Kidd, Duffy, MNS, Diego, Dombrock, Colton, Yt, Vel) and HPAs 1-16 in a single assay (Halawani et al., 2014). In this study, 167 amplicons were amplified using two primer pools with ultra-high multiplex amplification. Primer digestion then took place and was followed by ligation of the barcoded adaptors. Unlike the other protocol of NGS, the advantages of this panel include the non-requirement for fragmentation step as well as the multiple purification steps. Therefore, this will make the panel the method of choice in the future for typing the blood group and HPAs.

Moreover, this panel was based on NGS and it is a sequence-based typing which is able to detect the nucleotide sequence, in particular for the novel SNPs. Four novel SNPs were identified, one in the *RHCE* gene 208C>T (Arg70Trp) and three in the *KEL* gene. The first SNP was 331G>A (Ala111Thr) in exon 4. The second SNP was 1907C>T (Ala636Val) in exon 17 and the third SNP was 2165T>C (Leu722Pro) in exon 19. Unfortunately, due to poor electropherogram results the data of Sanger sequencing were non-conclusive.

4.4.1 Genotyping by the HEA and HPA Panel

The sequencing data obtained by the HEA and HPA Panel led to successful genotyping and the phenotypes were extrapolated easily. The Ion Reporter™ software facilitated the annotation of the sequencing data and accordingly the genotyping data were produced by the software. The prediction of the phenotypes was performed according to the Blood Group Antigen Factsbook (Reid et al., 2012). In this study, the *ABO* genotyping data successfully confirmed the serology information provided [Table 4.2] in

comparison to previous work by Fichou group where it was not possible to genotype for the ABO blood group system (Fichou et al., 2014).

10 weak D samples were assessed by the HEA and HPA Panel. Nine samples were successfully genotyped [Table 4.3]. These were four samples of weak D Type 1 (*RHD*01W.01*), four samples of weak D Type 2 (*RHD*01W02*) and one sample of DAR3.1 weak partial D 4.0 (*RHD*DAR3.01*). Interestingly, in the last sample no mutations were observed within any of the 10 exons of the *RHD* gene. This could be mistyped by serology as a weak D, while it is indeed an RhD-positive sample. On the other hand, it could be mistyped by NGS and due to the issue of unspecific primers no mutations were detected. However, the 10 exons need to be sequenced using Sanger sequencing to identify whether the sample is weak D or RhD-positive. Because of the time factor this sequencing was not performed.

The alleles of the *RHCE* gene were successfully genotyped, especially the weak form that could not be identified by serology. The predicted antigens included partial e, V and VS antigens. The *RHCE*C* and *RHCE*c* alleles could not be identified correctly by the HEA and HPA Panel, apart from the SNP of 48C>G (Cys16Trp) in exon 1. The same outcome was obtained by Fichou and co-workers, in which they could not distinguish between the *RHCE*C* and *RHCE*c* alleles (Fichou et al., 2014).

However, in our data a novel heterozygous SNP 208C>T (Arg70Trp) was detected in the *RHCE* gene. However, this SNP was misaligned to the *RHD* gene in an allelic imbalance ratio of 60:40 [Figure 4.13]. This issue was revealed using specific primers along with the LR-PCR approach followed by sequencing on the Ion PGM™ [see Chapter 6]. It was very obvious that the primer design for the HEA and HPA Panel did not take the gene homology into consideration. The same issue was found by Fichou et al. (2014) in which the primer design did not predict the typing of C, c, M and N antigens due to non-specific primers of their design.

The genotyping of the other blood group systems was properly achieved by the panel. The *JK*01W.01* allele encoding the weak form (Jka+^{w}) was easily identified in four samples. The low prevalence form of the alleles can possibly be detected by the panel, such as the *YT*B* and *CO*B* alleles. However, some of the attempts to investigate the *GYPA*M* allele failed and this may be due to the non-specific primers for the homologous genes of the MNS blood group system [Table 4.4].

The HPAs were also included in the panel and they were assessed easily. All the genotyping results of the 16 HPAs are demonstrated in [Table 4.5]. Along the high clinically significant platelet alleles, HPA-1a was found in nine heterozygous and 19 homozygous samples. The HPA-5b was detected in three samples in a heterozygous form.

4.4.2 Missed regions in the designed panel

One of the disadvantages of the HEA and HPA Panel was the issue of missed regions in the initial design by the Ampliseq™ Designer of non-covered targets. The missed areas in the initial design include 30 bp in the *ABO* gene, 45 bp in exon 18 in *SLC4A1* of the Diego blood group system and 63 bp in exon 7 of the *KEL* gene. Fortunately, the missed regions of the Kell and Diego blood group systems did not belong to any predefined antigens, although it might be risky for any novel alleles that could be discovered in the future.

Among the missed regions within the ABO blood group system, two missed regions were found in exon 7, which shows a high level of polymorphism. Such an issue was an obstacle to genotype the variants in the ABO blood group system, although the genotyping data was obtained. Fichou et al. (2014) designed a customised panel using the Ion Ampliseq™ Custom Panel and found that the missing regions were not amplified in the *ABO* gene although the initial design had full coverage of those areas. The authors stated that the consecutive stretches of homopolymers within numerous exons

and flanking introns were the reason for the complexity involved in amplifying the exons of the *ABO* gene (Yamamoto et al., 1990). However, in this study *ABO* genotyping was successfully performed, as listed in Table 4.2.

4.4.3 Low depth of coverage

One of the disadvantages of this study was that the depth of coverage in exon 8 of the *RHD* was very low which was around 16 repeated reads or fewer. In addition, the area of exon 8 of the *RHD* gene was poorly mapped [Figure 4.9]. This may affect the genotyping of *RHD*DAU* allele and make it unfeasible. The reason for this problem might be because of the unspecific primers and the mapping of the homologous genes *RHD* and *RHCE*.

Regarding the MNS blood group system, some samples that observed the *GYPA*M* allele had a lower depth of coverage of around five sequencing reads. Furthermore, some samples observed missed regions between amino acid positions (23-31) [Figure 4.10]. The only clear explanation for such an issue is the non-specificity of the primer design regarding homologous genes *GYPA* and *GYPB.*

4.4.4 The issue of unspecific primers

The missed regions and the poor mapping quality in the *GYPA* gene as well as the mistyping of the novel allele on the *RHD* gene instead of the *RHCE* gene are due to the unspecific primers provided by the Ion Ampliseq™ Designer. Fichou et al. (2014) designed a panel for blood group antigens based on the Ion Ampliseq™ Custom Panel to detect 15 blood group systems. The panel was for ABO, MNS, Rh, Lutheran, Kell, Duffy, Kidd, Diego, Yt, Scianna, Dombrock, Colton, Landsteiner-Wiener, Cromer and Knops. A similar issue with non-specific primers was found regarding homologous genes *RHD, RHCE, GYPA* and *GYPB* by Fichou and co-workers as an allelic imbalanced ratio of (80:20) instead of (50:50) for the heterozygous SNP.

To overcome this issue, Fichou et al. designed a set of specific primer pairs per sample for both exon 1 and exon 2 for *RHD, RHCE, GYPA* and *GYPB.* Stabentheiner et al. (2011) designed specific primers for the *RHD* gene and the amplification and ultimately the sequencing data were successfully obtained without any possible contamination to amplify the *RHCE* gene. By using our LR-PCR approach to amplify both *RHD* and *RHCE* genes, the sequencing data were aligned successfully for the gene of interest in consideration of masking the undesired gene [see Chapter 6]. The novel allele that was discovered in the *RHCE* gene by the HEA and HPA Panel was aligned incorrectly to the *RHD* gene. However, in consideration of using the specific primer of both genes in LR-PCR approach, *RHD* and *RHCE*, the SNP was aligned properly to the *RHCE* gene.

Weinzek-Lischka and co-workers designed a panel to genotype four HPAs and 10 targets from the blood groups. The panel comprised HPA-1, HPA-3, HPA-5 and HPA-15 alleles. The panel was mainly designed in order to avoid the risk of FNAIT in babies to HPA-1b/b mothers who developed anti-HPA-1a antibodies. Regarding the blood group antigens, the targets included exon 4 and 7 of the *RHD* gene and the SNP of the following antigens *RHCE*C/c, RHCE*E/e, GYPA*M/N, GYPB*S/s, FY*A/FY*B, KEL*01/02, JK*A/B, FY*01N.01*. To test the presence of foetal DNA, an internal control was used, including the *SRY* gene and eight anonymous SNPs from the Ion AmpliSeq™ Sample ID-Panel kit to prevent any false negative results. The authors stated that this method does not require paternal zygosity testing and provides non-invasive prenatal testing for the HPAs and the blood groups (Wienzek-Lischka et al., 2015). The advantage of the Ion Ampliseq™ Custom Panel is that it uses short amplicons ranging from 100 to 250 bp. However, the designed panel has limited numbers of SNPs which needs to be expanded to obtain more genotyping analysis, as the primer pool of Ampliseq only requires 10 ng of genomic DNA. Moreover, by using

a small number of SNPs the price of the NGS will be high and the benefit of the NGS will not be invested.

4.4.5 Population screened

Such a panel can be used as required by a laboratory and this gives the flexibility to determine the blood group antigens and the HPAs of interest. The choosing of this panel was similar to the Bloodgen project, which is more suitable for the European population (Avent, 2009). However, the Vel blood group system was added because of its clinical significance and the fact that it is encoded by *SMIM1* gene which is 4187 bp. Furthermore, the custom panel can be adjusted to screen certain populations that need specific genotyping such as the Asian population. McBean et al. (2014) stated that different populations' need for specific genotyping could vary because the ethnicity and the microarray platforms are not the right choice for the Australian population. The author stated that BeadChip™ was based in the United States and depends on African-American ethnicity, while BLOODchip® was designed for the European population (McBean et al., 2014).

4.4.6 Scalability of the number of samples

The HEA and HPA Panel were based on NGS and can be used for sequencing on Ion PGM™ or Ion Proton™. Therefore, scalability can be managed according to the number of screened samples in consideration of using the barcoded adaptors. The size of the used panel in this project was 31.18 Kb. The number of samples can be scaled depending on the type of platform used and the type of chip used.

Table 4.7 demonstrates the number of samples that can be analysed with the different chips with a depth of coverage of 100× on the Ion PGM™ platform, as well as the cost of the sequencing run. The size of the panel can be scaled depending on the sample number that needs to be screened. In other words, for a small number of donors, the Ion 314™ chip Version 2 can be used to analyse 16 samples. For a large cohort, 160 samples

and 320 samples can be tested with the Ion 316™ Chip Version 2 and Ion 318™ Chip Version 2, respectively. The throughput can be increased to 3,207 samples if the Ion Proton™ platform using Ion PI™ Chip Version 2. However, there remains an issue of having a sufficient number of barcoded adaptors because there are still only 96 adaptors commercially available.

This number of samples can be achieved with a 100× depth of coverage. This means the same target was sequenced at least a 100 times. There is an argument regarding which depth of coverage needs to be used. Bentley et al. (2008) published the first paper on sequencing the human genome using short read sequences with Illumina technology (Bentley et al., 2008). He stated that a coverage depth of 15× and 33× is suitable regarding accuracy, for homozygous and heterozygous SNPs, respectively. Consequently, these numbers of depth of coverage were applied on both HEA and HPA Panel and LR-PCR approach.

WGS was performed for an Asian individual and obtained high accuracy sequencing reads with an average 36× depth of coverage (Wang et al., 2008). Ajay et al. (2011) concluded that using a depth of coverage of 50× sequencing reads is required for higher accuracy for WGS. Then, the author found that the same level of accuracy can be achieved by using a 35× depth of coverage when new software and new improved sequencing chemistry were used in order to reduce GC bias (Ajay et al., 2011). Thermo Fisher Scientific has released a new sequencing polymerase called Ion Hi-Q™ chemistry, which can be used for both Ion PGM™ and Ion Proton™. The improvements in this new sequencing chemistry include a development in variant detection and a reduction of false positives in indels for microbial *de novo* assembly (Thermo Fisher Scientific, 2015b).

Table 4.7 Number of samples can be performed on the HEA and HPA Panel using different chips with a depth of coverage of 100X on the Ion PGM.

Chip type	Capacity	Number of samples	Cost per run
Ion 314™	30-50 Mb	16	£271.55
Ion 316™	300-500 Mb	160	£444.80
Ion 318™	600 Mb – 1Gb	320	£619.67

4.4.7 The cost of HEA and HPA

The price per sample of the HEA and HPA Panel is around £118 for the sample preparation and the sequencing on the Ion PGM™ costs £3.22 per sample using the Ion 318™ Chip. The price per SNP might be difficult to calculate. This is due to the sequencing which captures many targets within the gene and it is indeed a sequence-based typing and requires an amplification process prior to the sequencing procedure. Therefore, many targets can be taken into consideration, including entire genes rather than a single SNP. In comparison, the BLOODchip® service mainly tests specific targets, mainly SNPs, for 10 blood group systems including ABO, Rh, Kell, Kidd, Duffy, MNS, Diego, Dombrock, Colton and Lutheran as well as the HPA1-11 and HPA-15. The cost is £150 per sample (Redfern, L., Personal Communication, 17th July 2015). The price per sample for the HEA and HPA Panel is lower and the information received from the sequencing more comprehensive as it includes 11 blood group systems and the HPAs 1-16. Interestingly, the panel assay captures the entire exons of the blood group genes and targets the SNPs associated with HPAs.

Overall, the HEA and HPA Panel provided a rapid protocol for genotyping of the blood group antigens and HPAs. The approach was based on deep sequencing using semiconductor technology. One of the advantages was that the scalability of the chip would provide high-throughput screening of many samples as long as the barcoded adaptors are used. Furthermore, this method of genotyping was able to identify novel alleles easily, as four alleles were identified in this study. Finally, the price per sample is extremely affordable in contrast to the microarray technology. Although some issues were faced in relation to the specificity of the primer sequences, the future of this assay will supplant the microarray technique and the conventional serology. This will assist to reduce the risk of alloimmunisation by screening many antigens, especially low prevalence ones. Therefore, this will help multiply transfused patients such as SCD

individuals. Moreover, using a combination of this assay to include HPA, the risk of MPR, PTP and FNAIT will be diminished. Interestingly, BGG using NGS technology will pave the way to genotype both the blood group antigens and the HPAs.

Chapter 5 : Genotyping the Kell Blood Group by Next-generation Sequencing

5.1 Introduction

The Kell blood group system (ISBT 006) is a highly polymorphic blood group system [see section 1.6]. Most of the Kell blood group antigens have been associated with single nucleotide changes that give rise to amino acid substitutions, apart from KEL20, which require the presence of the Xk protein (Lee et al., 1997; Daniels, 2013a). Numerous genetic mechanisms may lead to the lack of all Kell antigens, which is designated as K_{null} (K_0). These mechanisms include amino acid substitutions, premature stop codons, and a 5' splicing-site mutation in exons and introns of the *KEL* gene (Lee et al., 2001). Different missense mutations result in the weak expression of Kell antigens, which is known as K_{mod} (Lee et al., 2003b).

Anti-K is the most clinically significant antibody, which can cause HTR and HDFN. BGG paves the way to genotype the blood group antigens and to overcome conventional typing by serology (Avent et al., 2007). NGS is an emerging technique that enables to capture many targets in the genome and to sequence them repeatedly (Mardis, 2008).

5.2 Aims

The aim of this project was to investigate whether NGS is a suitable platform for genotyping the Kell blood group system. This design worked as a model for the Rh blood group genes to sequence the entire gene and to genotype its alleles using NGS in order to investigate the mutations within the gene.

The entire *KEL* gene of the Kell blood group system was amplified by designing two primer pairs and was then sequenced by NGS using Ion PGM™. Using the LR-PCR approach assisted to obtain extensive genotyping of all the alleles of the Kell blood

group system, in particular the high prevalence ones. Intronic SNPs can be investigated, especially for those areas that might be accompanied by the alleles responsible for K_{null} and K_{mod} phenotypes due to giving rise to several molecular mechanisms including intronic polymorphisms. This approach was used as a model to assess whether such a procedure can be applicable to the Rh blood group system, which possesses two highly homologous genes.

5.3 Results

5.3.1 Long range PCR for *KEL* gene

Twenty samples with known serology were chosen for this experiment. Table 5.1 shows the serology information for the Kell blood group system. Two primer pairs, about 12 Kb of each amplicon, were designed per sample to amplify the whole *KEL* gene [Figure 5.1]. Figure 5.2 shows the PCR results of the *KEL* gene on 0.7% agarose gel, which runs on gel electrophoresis at 80V for 1 hour and 40 minutes.

The PCR mastermix used was LongAmp Hot Start Taq 2X Master Mix (New England Biolabs, United Kingdom). This mastermix possesses a fidelity of two times in comparison to the standard Taq polymerase according to the manufacturer's instructions. This polymerase has the capability to amplify up to 20 Kb from human and 40 Kb from bacterial DNA.

Table 5.1 Serology information for the 20 samples regarding the Kell phenotypes.

No.	Sample ID	K	k	Kp^a	Kp^b
1	5686	−			
2	7268	−			
3	469X	−			
4	3857	−			
5	9139	−			
6	736G	−			+
7	616Z	−			
8	8387	−			+
9	566L	−			
10	799W	−			
11	431S	−			
12	004B	−		−	+
13	227P	+			
14	905N	+	+		
15	1516	−			
16	330C	−			
17	619U	−			
18	486Y	−			
19	7496	−			
20	733M	−			

These samples were used in the amplification of the entire *KEL* gene and that had been sequenced by NGS.

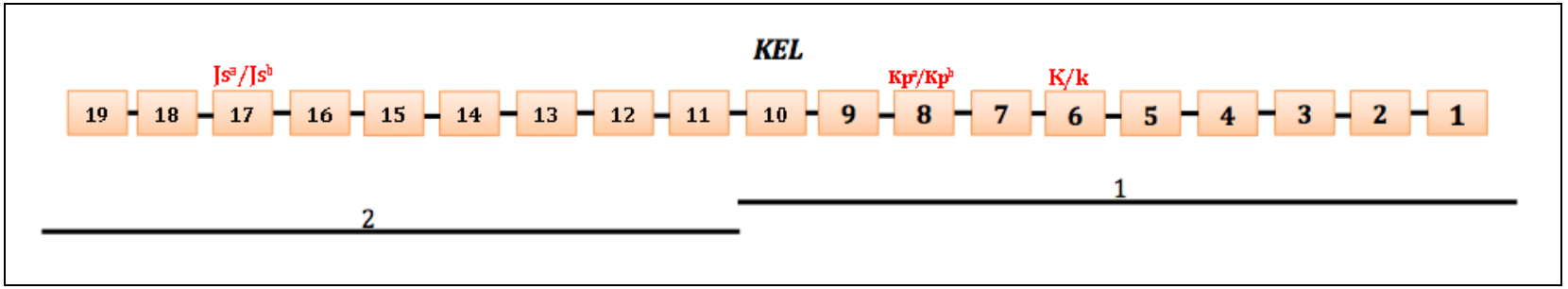


Figure 5.1 Two LR-PCR products covered the whole *KEL* gene.

The *KEL* gene is approximately 21.3 Kb possessing 19 exons. Each PCR product is around 12 Kb with an overlap of 652 bp. The boxes show the exons (coding area), while the lines are the introns (non-coding area). The numbering of the exons is in reverse direction because the *KEL* gene is an anti-sense gene. The major antigens (K/k, Kp^a/Kp^b and Js^a/Js^b) are indicated in red.

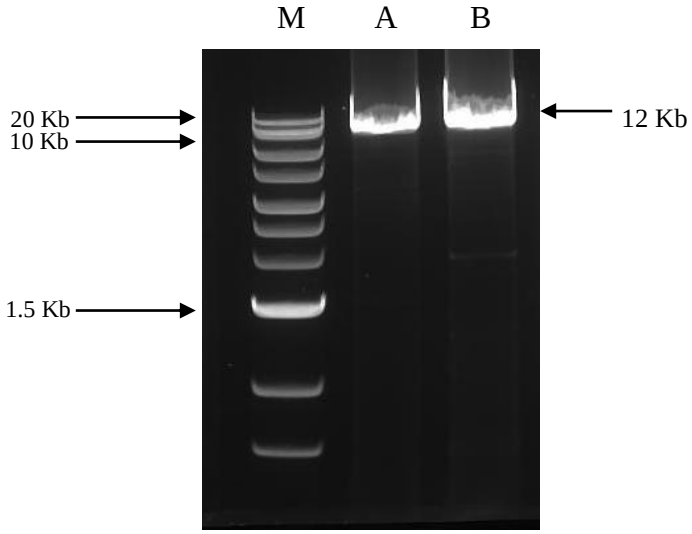


Figure 5.2 Amplification of the entire *KEL* gene by two products of LR-PCR.

Each sample had two products (A and B), about 12 Kb each. The PCR products were purified using 0.7% agarose which runs on gel electrophoresis at 80V for 1 hour and 40 minutes. The sample used in this example was identified by serology as kk antigen. The two samples of the Kk antigen had identical results.

5.3.2 Sequencing Libraries

5.3.2.1 Fragmentation

The sequencing libraries of the NGS were constructed following the purification of the amplicons. Then, the fragmentation was performed using the Ion Xpress™ Plus Fragment Library Kit. The shearing enzyme was utilised for 15 minutes. After the fragmentation had taken place, the purification was carried out using Agencourt® AMPure® XP reagent. The results were assessed using the high sensitivity DNA Kit which was run on Agilent® 2100 Bioanalyzer. A peak of fragmented DNA was achieved by the shearing enzyme and was around 50 to 500 bp. Figure 5.3 shows the fragmentation result for one of the *KEL* sequencing libraries.

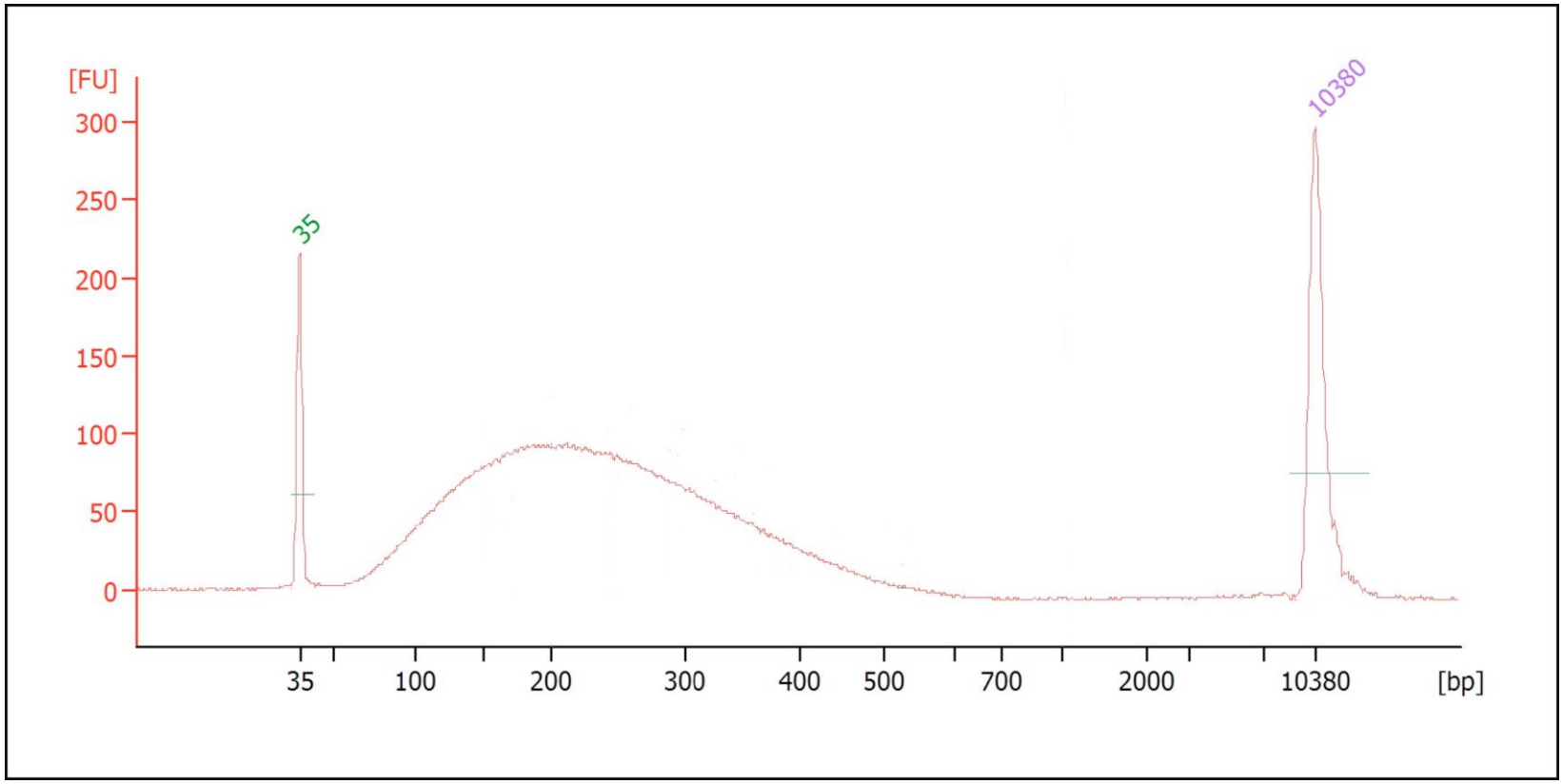


Figure 5.3 An electropherogram of the fragmented *KEL* sequencing library (a pool of two *KEL* amplicons).

The method was carried out using the Ion Xpress™ Plus Fragment Library Kit for 15 minutes. The assay was assessed using Agilent® 2100 Bioanalyzer instrument. The green and purple numbers show the lower marker and upper marker in bp, respectively.

5.3.2.2 Ligation

Figure 5.4 demonstrates the ligation result of one of the *KEL* sequencing libraries. The process of ligation was carried out using the Ion Xpress™ Plus Fragment Library Kit. Both adaptors, P1 and the barcode adaptor, were ligated using the ligase enzyme with the purified fragmented DNA.

5.2.2.3 Size selection using SPRIselect magnetic beads

Following the purified ligated products, the sequencing libraries underwent size selection. This procedure was performed using SPRIselect® magnetic beads. A peak of around 200 bp was achieved for the size selected sequencing libraries. This was in order to run it on a 200 bp reading length. Figure 5.5 demonstrates the size selection of one sample of the *KEL* sequencing libraries.

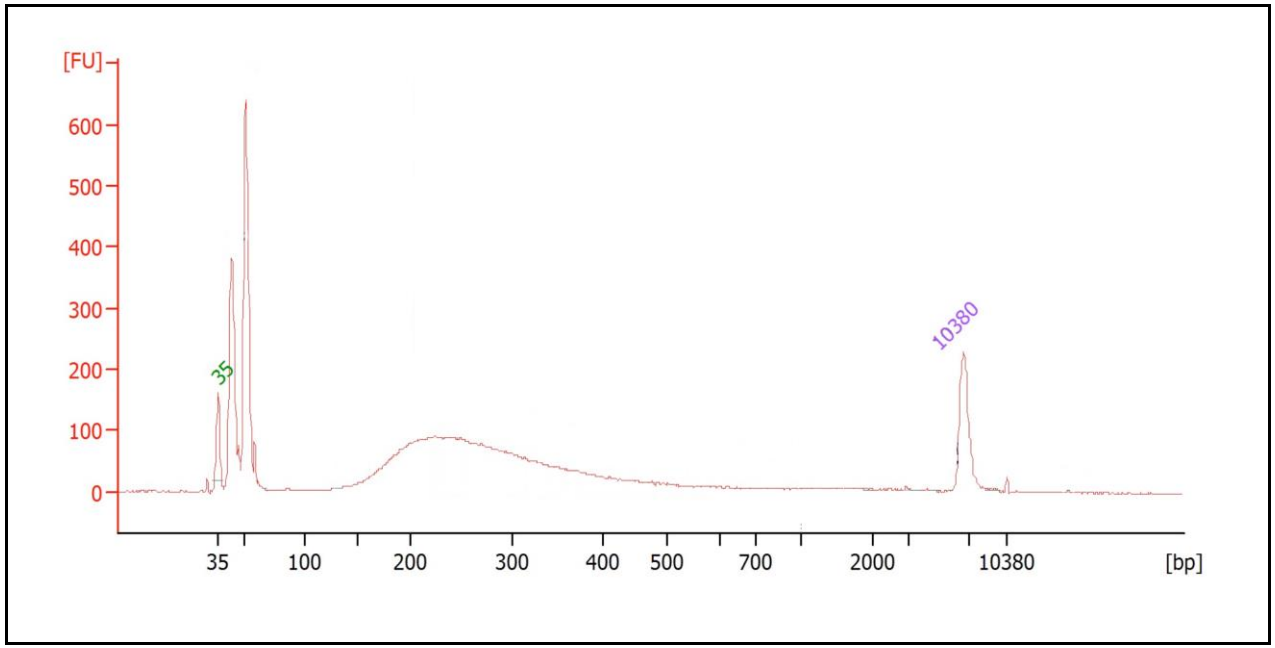


Figure 5.4 An electropherogram of the ligated products of the *KEL* sequencing library.

The ligation was performed using the Ion Xpress™ Plus Fragment Library Kit. The assay was assessed by the Agilent® 2100 Bioanalyzer instrument. The green and purple numbers show the lower marker and upper marker in bp, respectively.

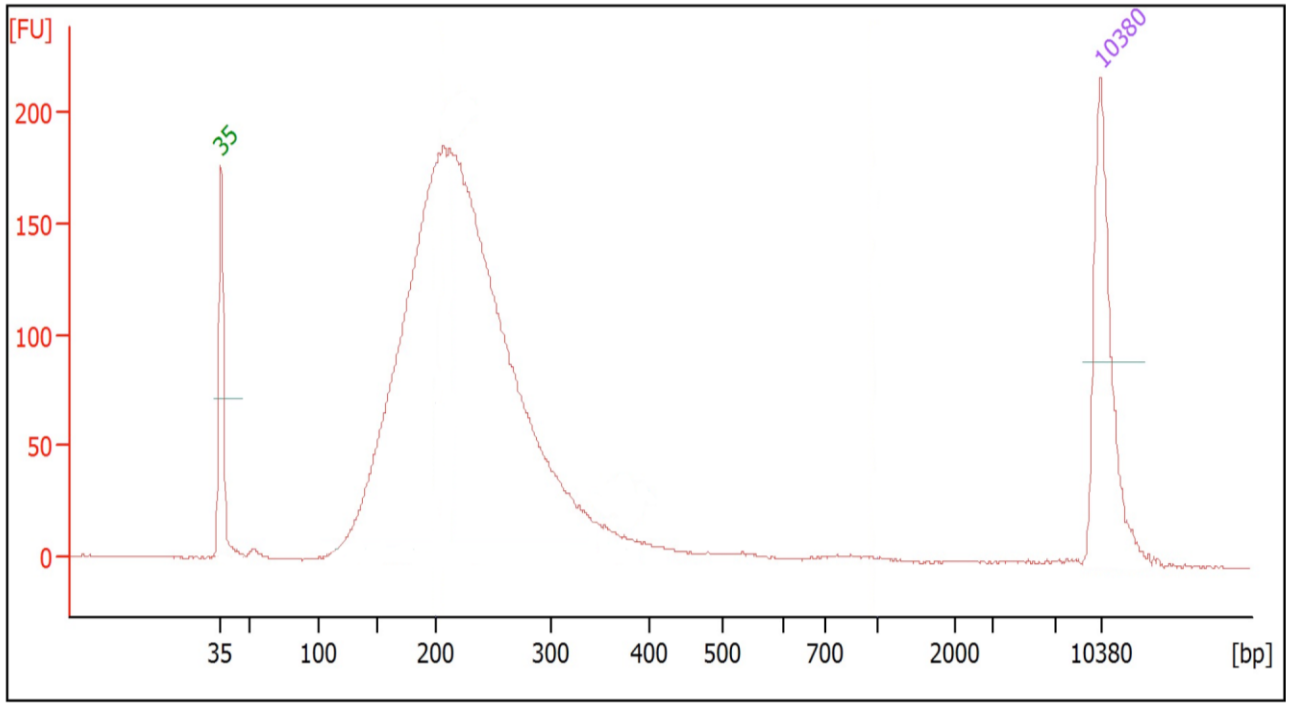


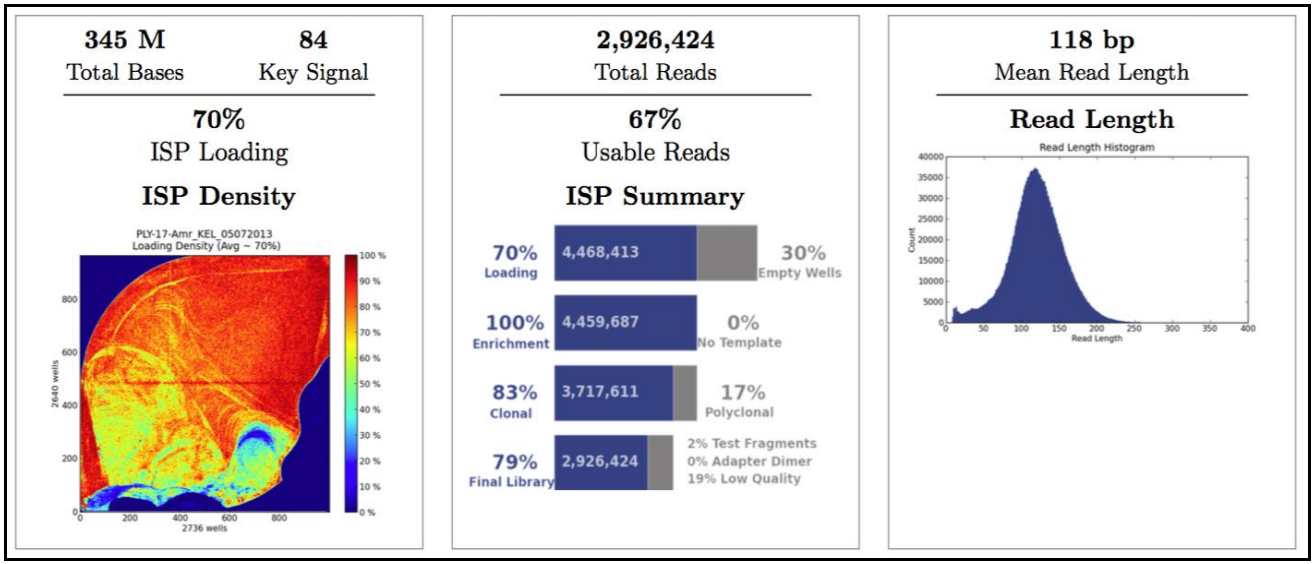
Figure 5.5 An electrophergram of the size selected *KEL* sequencing library.

A peak of 200 bp was achieved for a 200 bp reading length. This procedure of size selection was performed using SPRIselect® magnetic beads. The assay was assessed by the Agilent® 2100 Bioanalyzer instrument. The green and purple numbers show the lower marker and upper marker in bp, respectively.

5.3.3 Analysis of NGS sequencing data

Following the library construction for the *KEL* gene and sequencing them on the Ion PGM™, the sequencing data were generated in the first instance by Torrent Suite Version 4.4. Adaptors were trimmed by default by the Torrent Server software. The total reads for the 20 samples of the *KEL* gene were about 2.92 million reads with a mean coverage depth of around 840 sequencing reads. Figure 5.6 shows the sequencing report for the sequencing run on the Ion PGM™. The loading on Ion 316™ chip was 70% in which the wells of the chip were addressed by ISPs. This was in comparison to a percentage of 30% of the empty wells. The percentage of usable reads was 67% to be used for further downstream analysis. Two sets of software were used; the first was well classification and the other was responsible for filtering and trimming. The well classification considered the addressable wells in contrast to the empty wells and was reported as bead loading density. Regarding the filtering and trimming software, the filter removed the polyclonal reads with more than one template that attached to the beads, reads with low quality or less than 4 bp and a primer dimer of less than 8 bp.

The percentage of the templates enriched with sequencing libraries was 100%. The clonal reads comprised 83%, while the percentage of polyclonal reads was 17%. The percentage of clonal templates demonstrates that the template had a single sequencing library attached to the ISP prior to starting the emulsion PCR. This is in comparison with the polyclonal templates, which had more than one sequencing library attached to the ISP. The final library was 79% and the mean and median of the read length were both 118 bp.



(a) (b) (c)

Figure 5.6 An overview of a report on the sequencing run of the *KEL* gene.

(a) Loading density of ISP that was addressed by the chip wells, which was 70%.
(b) Total of 2,926,424 reads after filtering and trimming.
(c) A histogram of the read length of the sequencing libraries with a mean of 118 bp, on which the y-axis demonstrates the read count, while the x-axis shows the read length in the bp.

5.3.4 Quality Control

5.3.4.1 Per base sequence quality

A plugin on the Torrent Suite called FastQC was run to assess the quality of the generated sequencing according to the Phred score. Figure 5.7 demonstrates the Phred score on the y-axis across the position in reads per base pair. In this study, the sequencing quality obtained a Phred score of around 29-30. Accordingly, this means accuracy of about 99.9% and the probability of the base being called incorrect was 1 in 1000. This outcome enabled the user to carry out further analysis including the genotyping.

The background of the graph splits the y-axis into three different regions with different colours; green for the best quality reads, orange for the reasonable reads and red for the poor quality reads. Box and Whisker plots were drawn on every position per base pair. The inter-quartile range (25-75%) was represented by yellow boxes. The upper and lower whiskers demonstrate 10% and 90% points.

The blue line shows the mean Phred score across the reads in bp, while the central red line demonstrates the median. The mean for the sequencing results is around 29 on the Phred score, which gives accuracy more than 99%. The yellow boxes stopped after reaching the maximum range of the sequencing read length.

5.3.4.2 Per Sequence Quality Scores

Figure 5.8 demonstrates average quality using Phred score per reads. The FastQC tool was utilised to assess if a subset of the sequences has poor quality reads. The results here demonstrate that the quality of most of the reads was 30 according to the Phred score, which indicates high quality reads with an accuracy of 99.9%. This means the probability of the base called to be incorrect was 1 in 1000.

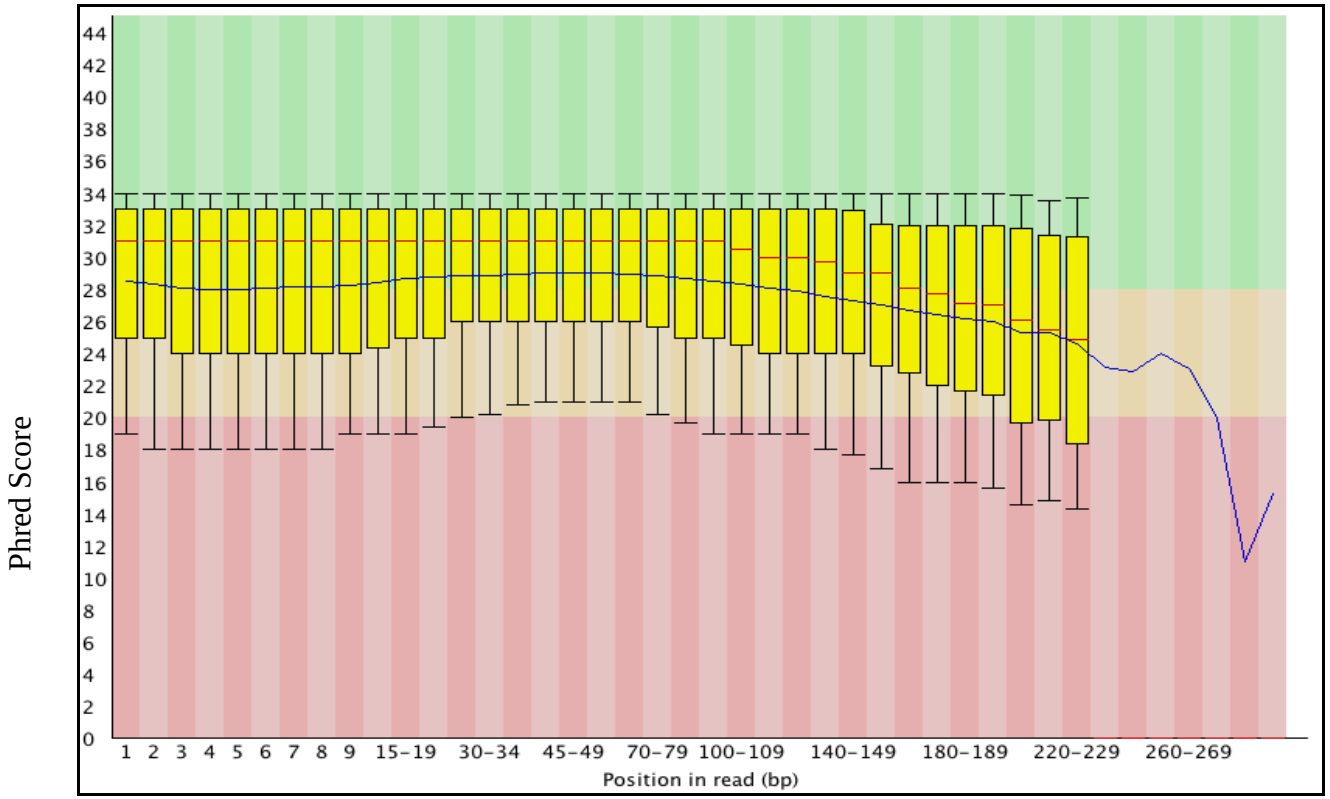


Figure 5.7 Phred quality scores across all the bases for a single sample of the *KEL* gene.

The background of the y-axis demonstrates the Phred score, while the x-axis shows the position in reads per bp. The background of the graph splits the y-axis into three different regions with different colours; green for the best quality reads, orange for the reasonable reads and red for the poor quality reads. Box and Whisker plots were drawn on every position per bp. The inter-quartile range (25-75%) was represented by yellow boxes. The upper and lower whiskers demonstrate 10% and 90% points. The blue line represents the mean of the quality according to the Phred score across the reads in position. All the analysed sample has the same Phred scores.

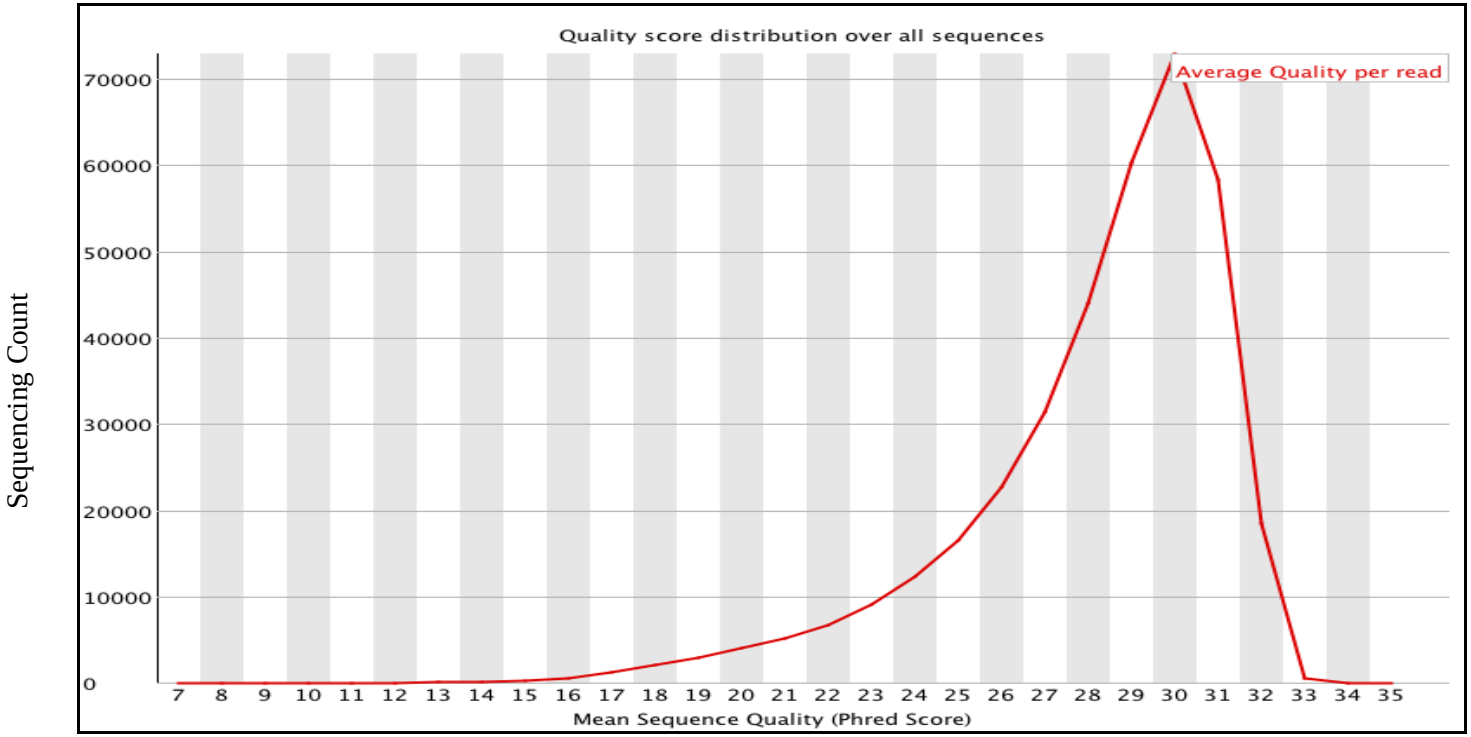


Figure 5.8 Quality scores per sequencing count for a single sample of the *KEL* gene.

The mean sequence quality according to the Phred score is demonstrated across the read counts. This tool assesses if a subset of the sequencing reads is of poor quality. The results here demonstrate that the quality of most of the reads was 30 according to the Phred score, which indicates high quality reads with an accuracy of 99.9%. This means the probability of the base called to be incorrect was 1 in 1000. All the analysed sample has the same Phred scores.

5.3.5 Sequencing visualisation

The generated sequencing data were visualised using IGV Version 2.3.46. The IGV software aligns the sequencing data against a reference gene showing the chromosomal locations of nucleotides and assesses the genomic variations such as SNP changes including zygosity determination, insertion or deletions and the coverage depth of the sequencing. BAM files were utilised in order to visualise the sequencing data. **Figure 5.9** shows the heterozygous SNP for *KEL*01.01/02* allele at a chromosomal position of chr7:142,655,008 with a coverage depth of 817. This confirmed the phenotyping by serology in which the sample types for both K and k phenotypes. The *KEL* gene is an anti-sense strand, therefore the SNP was shown as G>A instead of C>T at nucleotide position 578. Consequently, this missense mutation gives an amino acid substitution of Thr193Met.

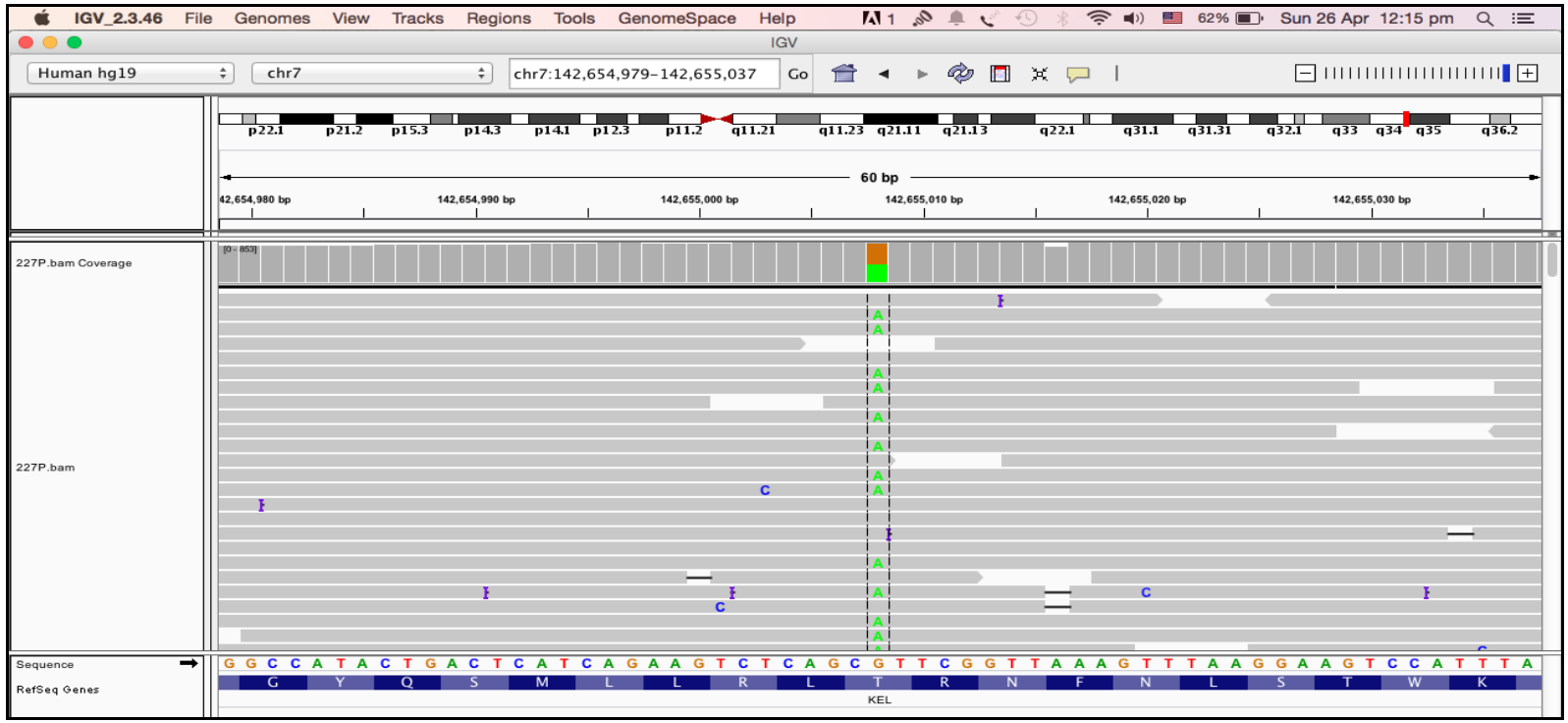


Figure 5.9 A heterozygous SNP in the *KEL* gene for the *KEL*01.01/02* genotype.

This SNP was in chromosomal position chr7:142,655,008 with a coverage depth of 817. The *KEL* is an anti-sense strand, therefore the SNP was shown as G>A on the reference gene instead of 578C>T, which encoded an amino acid substitution to Thr193Met. Output from IGV software demonstrates the visualisation of the sequencing data. Chromosomal locations are shown on the top and the reference gene in blue at the bottom.

5.3.6 Variant analysis

The VariantCaller plugin in the Torrent Suite Software was used to annotate the SNPs against the reference gene and to produce the VCF files. Then, those files were uploaded for annotation using the SeattleSeq Annotation 141 website to annotate the SNPs (SeattleSeq Annotation Tool 141, 2014). **Figure 5.10** illustrates the table that resulted from the website for the annotated SNP.

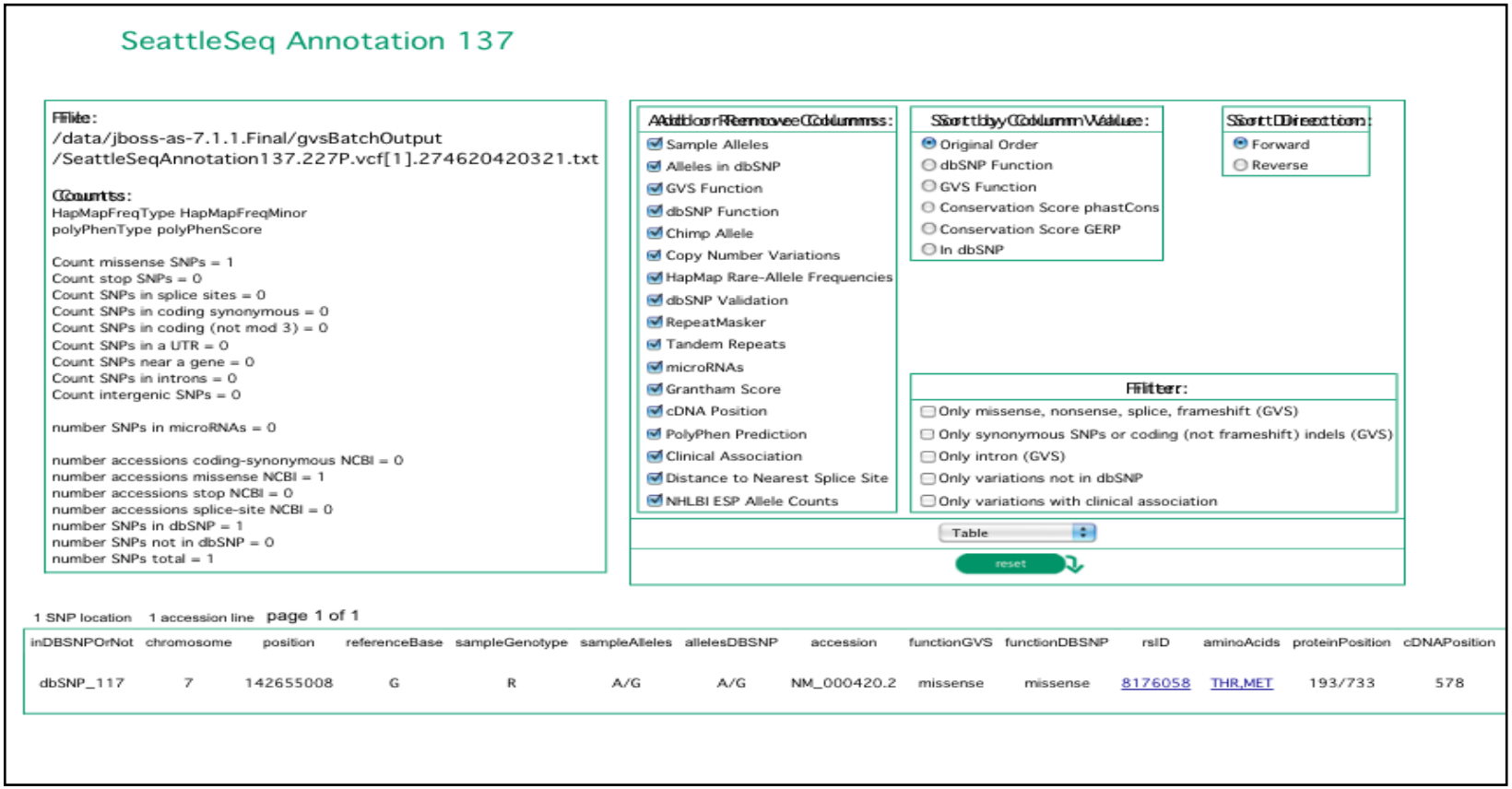


Figure 5.10 The genotyping results of the annotated SNP for the *KEL*01.01/02* allele using SeattleSeq Annotation 141 website.

The results show the chromosomal location of the SNP, the reference base, the sample alleles, the cDNA position, the type of mutation and the amino acid change. The reference nucleotide shown in this example was (G) and it was substituted by (A). This change in deed was C>T rather than G>A. This was because the *KEL* gene is an anti-sense strand, therefore the analysis was carried out on the sense strand by default.

5.3.7 Genotyping of the *KEL*

Twenty samples with known serology were sequenced by the Ion PGM™ platform. Two of these samples were phenotyped by serology as K. The NGS data confirmed the serology for these two samples. A heterozygous SNP was detected at 578C>T in exon 6 which gives rise to the amino acid substitution Thr193Met. An interesting finding was that the *KEL*02.03* allele encoding the Kp^a antigen was genotyped in a sample as 841C>T in exon 8 (Arg281Trp) which was not identified by serology. Table 5.2 demonstrates the main alleles of the Kell blood group system that were genotyped by NGS. One silent mutation was found in two samples A>C at 1680 in exon 15 with no change to the amino acid Pro560. The following high prevalence antigens k, Kp^b, Js^b, KEL11, KEL12, KEL13, KEL14, KEL18, KEL19, KEL22, KEL26, KEL27, KEL29, KEL30, KEL32, KEL33, KEL34, KEL35, KEL36, KEL37 and KEL38 were predicted in all samples. Table 5.3 lists 57 Intronic SNPs and six new intronic SNPs were found in a number of samples.

Table 5.2 Genotyping of the main alleles of the Kell blood group system by NGS.

Allele	Antigen	Necletide	Exon	Amino acid	Zygosity	Allelic frequency (Caucasians)	No. of Samples	Serological matching
KEL*01.01	K	578C>T	6	Thr193Met	Heterozygous 54.3% Heterozygous 48%	9.02%	2	Yes
KEL*02.03	Kp^a	841C>T	8	Arg281Trp	Heterozygous 50%	2.28%	1	Not identified
KEL*02.04	Kp^b	C at 841 G at 842	8	Arg281	Homozygous	100%	19	Yes

Table 5.3 A list of intronic SNPs with the *KEL* gene.

These SNPs were found in a number of samples. The rsID number is the number of the identified SNP in the NCBI database. The non-applicable (N/A) were not yet included in the NCBI database.

#	rsID	Zygosity	Base change	Intron	position	Distance to Splice	Samples found
1	N/A	Heterozygous	G>A	4	142657245	768	1
2	N/A	Heterozygous	G>T	4	142656011	494	1
3	8175963	Heterozygous	G>A	4	142657775	238	2
4	8175962	Heterozygous	A>G	4	142657915	98	2
5	8175972	Heterozygous	A>G	4	142655717	200	1
6	8175978	Heterozygous	C>T	5	142655078	16	1
7	8175977	Heterozygous	G>A	5	142655097	35	1
8	8175976	Heterozygous	G>C	5	142655113	51	1
9	N/A	Heterozygous	T>A	6	142652941	1325	1
10	71545388	Heterozygous	C>T	6	142654335	577	4
11	7799481	Heterozygous	C>T	6	142651868	252	3
12	6954192	Heterozygous	T>G	6	142652378	762	1
13	6958998	Heterozygous	A>G	6	142652850	1234	1
14	8175984	Heterozygous	G>A	6	142653478	1434	1
15	8175985	Heterozygous	T>C	6	142653424	1488	2
16	N/A	Heterozygous	G>C	6	142653070	1454	1
17	N/A	Heterozygous	G>A	6	142653514	1398	1
18	3757853	6 Heterozygous & 6 Homozygous	G>C	9	142650475	418	12
19	8176003	Heterozygous	G>A	10	142647564	2030	2
20	145755643	Heterozygous	A>G	10	142646888	2706	1
21	8176063	Heterozygous	G>A	10	142643870	464	5
22	150938699	Heterozygous	G>A	10	142649218	376	1
23	8176061	Heterozygous	T>C	10	142646386	2980	2
24	79243129	Heterozygous	C>T	10	142644474	1068	1
25	11980340	Heterozygous	G>A	10	142644521	1115	1
26	8176008	Heterozygous	A>G	10	142645944	2538	1
27	73464414	Heterozygous	A>G	10	142646081	2675	1
28	8176006	Heterozygous	T>C	10	142646995	2599	1
29	8176005	Heterozygous	G>A	10	142647428	2166	1
30	7806742	Heterozygous	G>A	10	142647720	1874	3
31	6974425	Heterozygous	A>G	10	142648273	1321	1
32	8176001	Heterozygous	T>C	10	142649292	302	1
33	8175999	Heterozygous	T>C	10	142649564	30	1
34	10261078	Heterozygous	G>A	10	142643676	270	2
35	8176018	Heterozygous	C>T	10	142643774	368	2
36	8176012	Heterozygous	C>A	10	142644900	1494	2
37	8176010	Heterozygous	G>A	10	142645614	2208	2
38	N/A	Heterozygous	G>A	10	142649435	159	1
39	7781817	Heterozygous	A>G	11	142642940	352	1
40	7780729	Heterozygous	C>A	11	142642985	307	1
41	7807823	Heterozygous	A>G	11	142643099	193	3
42	8176027	Heterozygous	C>G	11	142641917	87	2
43	8176026	Heterozygous	T>C	11	142642098	268	2
44	8176023	Heterozygous	T>C	11	142642596	696	3
45	7784417	Heterozygous	C>T	11	142643092	200	2
46	187074728	Heterozygous	T>A	13	142641174	202	1
47	8176031	Heterozygous	T>C	13	142641382	26	3
48	7810229	Heterozygous	C>T	13	142641087	115	1
49	7797896	Heterozygous	A>G	13	142641189	217	2
50	7797908	Heterozygous	A>G	13	142641207	201	2
51	187074728	Heterozygous	C>T	13	142641174	202	1
52	8176037	Heterozygous	A>G	16	142640277	84	1
53	8176044	12 Heterozygous & 3 Homozygous	A>G	18	142639321	198	15
54	N/A	Heterozygous	G>A	18	142639472	47	1
55	143804816	Heterozygous	G>C	18	142639446	73	1
56	6949788	Heterozygous	A>G	18	142638697	195	1
57	7787760	Heterozygous	A>G	18	142639155	364	3

5.3.8 Validation of the Kell Antigens

Following the analysis of the NGS data, Sanger sequencing was carried out to validate the SNP of the *KEL*01.01* allele encoding the K antigen. The primers used for this validation are shown in [section 2.5.6]. Figure 5.11 displays the validated SNP for the *KEL*01.01* allele. The sample was heterozygous for that SNP (578 C>T) and considered to be a missense mutation, which gives an amino acid substitution, Thr193Met.

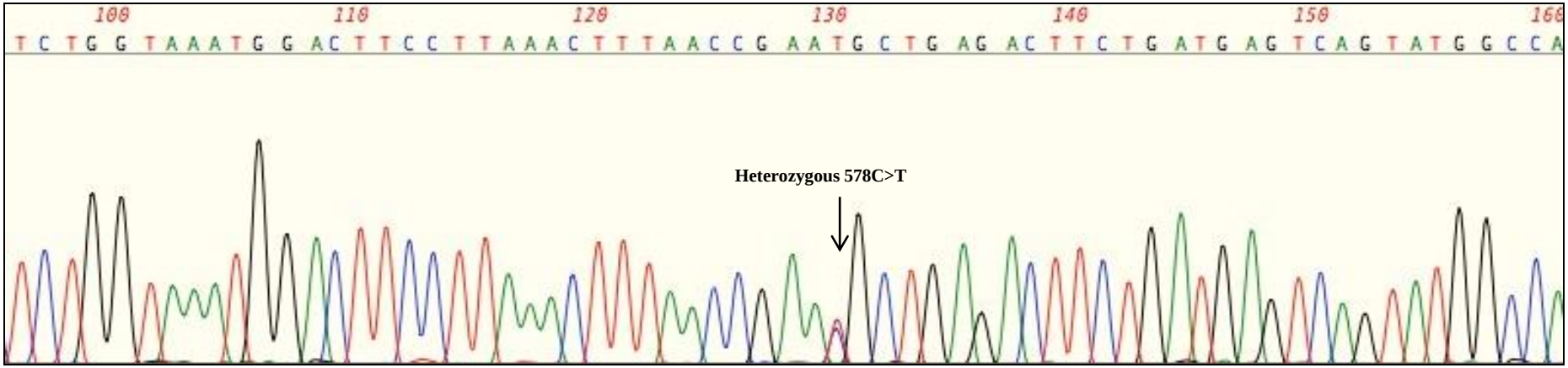


Figure 5.11 Validation of the SNP for the *KEL*01.01* allele encoding the K antigen.

A heterozygous SNP (578 C>T) was confirmed using Sanger sequencing in order to validate the NGS data [see Figure 5.9]. The serology provided with this sample was typed for K and k antigens. The electropherogram was generated using MacVector Version 12.7.

5.4 Discussion

5.4.1 Quality control of the NGS

Quality control is the first step which should be taken to analyse any sequencing run from the high-throughput platforms. FastQC is a tool used to perform the test to evaluate if the sequencing possesses good or poor reads. It is available as standalone free software or as a plug-in with Torrent Suite software (Andrews, 2010).

5.4.2 NGS data visualisation

IGV software was used to visualise the sequencing reads against the reference gene of interest in order to facilitate the analysis. The software is free and easy to use and does not require any previous knowledge. It can be used for assessing the data to demonstrate the mutation with distinctive colours, to determine the zygosity as well as find out the count of the depth of coverage of the sequencing reads (Robinson et al., 2011). The challenging issue in this project was that the *KEL* gene is an anti-sense strand in comparison to the sequencing data and reference gene which are all assigned in the sense direction. This issue needs to be noticed, otherwise a false identification of new alleles will be possible.

5.4.3 Genotyping of the Kell antigens

Out of 20 samples, two samples were predefined by serology as K phenotype. The NGS results confirmed the serological data and genotyped both samples to have a heterozygous SNP C>T578 (Thr193Met) in exon 6. This antigen was reported to be involved in causing of HTR and HDFN by both haemolysis and suppression of erythropoiesis (Vaughan et al., 1998). The first sample of K phenotype did not observe any of the intronic SNPs, while the second one had three heterozygous SNPs. These

three intronic SNPs were in intron 9 G>C (chr7: 142650475), in intron 10 G>A (chr7: 142647564) and in intron 18 A>G (chr7: 142639321).

*KEL*02.03* allele, encoding the Kp^a antigen, was genotyped in a single sample as heterozygous SNP 841C>T (Arg281Trp) in exon 8. This antigen was not detected by standard serology. Although the Kp^a antigen very rarely causes HDFN, it has the possibility to cause a severe type of HDFN that requires a blood transfusion (Costamagna et al., 1997). A case of hydrops fetalis was reported by Smoleniec et al. (1994) that was caused by anti-Kp^a. The mother was typed by serology as RhD-positive and Kp(a−b+), while the father was typed as RhD-positive Kp(a+b+) which gave a 50% chance of the baby being heterozygous Kp^a positive. The authors stressed that the low frequency antigens need to be serologically screened in a reference laboratory prior to a diagnosis of non-immune hydrops (Smoleniec et al., 1994). This shows the requirement for NGS platforms to involve as many antigens as possible. For example, the HEA and HPA Panel [see Chapter 4], which include 11 blood group systems and 16 HPAs.

Furthermore, another case has been reported by Koshy et al. (2009) of a Caucasian woman suffering from delayed haemolytic transfusion reactions attributed to anti-Kp^a after receiving multiple transfusion units. The DAT and the antibody screening were negative. One of the units was incompatible with antihuman globulin for Kp^a phenotype. The authors suggested to include the crossmatch testing for all the transfusion units with consideration of which antigen should be included in the antibody screening cells (Koshy et al., 2009). It should be noted that the implication of BGG will get rid of such an issue caused by the standard serology.

Another case of severe HDFN was reported by Rossi *et al.* (2013) which was caused by anti-Kp^a. Rossi and co-workers recommended that extended phenotyping is required for low prevalence antigens in particular when a DAT is positive for the foetal or neonatal cells. They indicated that even if there is other etiology causing the foetal anaemia, a

DAT is recommended before intrauterine transfusion (Rossi et al., 2013). Furthermore, an interesting case was reported by Tuson et al. (2011) of a mother pregnant with twins and her serum contained anti-Kp^a. The first child was anemic whose phenotyped (Kp^a+) and the other twin (Kp^a-) was not (Tuson et al., 2011). The constant decrease in bilirubin levels was coincident with the falling of the hemoglobin level in addition to a decrease in absolute reticulocyte count. These laboratory outcomes and clinical course of the affected twin were consistent with the suppression of erythropoiesis as well as immune RBCs destruction. Although anti-K is the antigen that is associated with erythropoiesis suppression, this case may confirm that the other antigens of the Kell blood group system can be involved in addition to K antigen (Tuson et al., 2011). The results may approve the interpretation of the Kell antigens that suppress the erythropoiesis due to the early representation of its protein in the red blood cells (Southcott et al., 1999; Daniels and Green, 2000). Consequently, the absence of hyperbilirubinemia in the patient stressed that the erythroid progenitor cells do not contain haemoglobin in the early phase of erythropoiesis (Tuson et al., 2011).

5.4.4 Genotyping of the *KEL*

The issue of having a test to detect the low prevalence antigens can be overcome by high-throughput sequencing. To our knowledge this is the first work that has utilised LR-PCR to amplify the entire *KEL* gene. The method helps to examine the *KEL* gene in high resolution by deep sequencing. This technique provides comprehensive genotyping for all the SNPs which can be found within the gene, and in particular the high prevalence antigens. In addition, all intronic SNPs can be examined, which will aid the detection of all the polymorphic mechanisms in rare cases that are associated with K_{mod} and K_{null}. The benefits of the intronic SNPs also include identification the area should be avoided when designing the primer and preventing any allelic dropout. The advantages of this approach can be invested in typing the low prevalence and rare Kell

antigens, which can be found in different ethnic backgrounds and used by different laboratories. Overall, it will facilitate the discovery of new alleles of the Kell system. This is due to its principle of genotyping by sequencing in comparison with the other limited molecular techniques such as microarray platforms, which requires previous knowledge of the alleles (Hurd and Nelson, 2009).

Rieneck and co-workers (2013) used NGS to genotyping the *KEL*01.01*/02 encoding K/k phenotypes to predict the foetal phenotype. The method is based on the primer fusion to the barcoded adaptors to distinguish between the different samples and to amplify the PCR products. The products were run on agarose gel electrophoresis to assess the required size and a further restriction enzyme experiment was implemented that used *Bsm*I to digest the K product (Rieneck et al., 2013). After that, the samples were sequenced on Illumina GAIIx which is no longer available. The drawback of this procedure was in observation of unexpected base calls, less than 3000 unpredicted base calls along the primer sequences and approximately 20000 in a number of positions. This outcome can lead to obtaining false-positive or false-negative results for the genotyping although the genotyping for *KEL*01.01* was precise. In the LR-PCR study, no such issue was reported.

The advantage of using NGS to screen many individuals in a single assay is due to the high capacity of the chip [see Table 5.4]. Furthermore, the sequencing chip can be scaled depending on the number of individuals required. On the other hand, the big target in our case to amplify the entire gene will reduce the number of individuals on the sequencing run. Such an issue can be resolved by the scalability of the chip in case the sequencing is run on Ion Torrent™ platforms.

The run cost for the prenatal analysis of the K/k experiment by Rieneck and co-workers (2013) was around $1500 (£971). On the other hand, the cost of the sequencing run on the Ion PGM™ on the medium size chip, Ion 316™ Chip, costs around £444.80. This

enormous difference in price allows the Ion PGM™ platform to do more sequencing runs because it is more affordable in comparison to the Illumina GAIIx.

The cost per sample preparation and library construction to test the entire *KEL* is approximately £48. This cost involved the two reactions of PCR per sample, the library construction using the Ion Xpress™ Fragment Kit including the barcoded adaptor and two runs of the assessment of the constructed library on the Bioanalyzer using the high sensitivity DNA kit.

The capacity of the Ion chips varies according to the scalability. Table 5.4 demonstrates the three Ion chips, the number of samples that can be obtained by a chip as well as the sequencing cost of each chip. By using a depth of coverage of 100×, 20 samples of sequencing for the entire *KEL* gene can be sequenced on Ion 314™ Chip. In order to increase the number of samples, 204 and 408 samples can be scaled and sequenced using Ion 316™ Chip and Ion 318™ Chip, respectively.

The cost of the run ranges from around £271.55 to £619.67 depending on the chip used. The sequencing cost is around £2.12 per sample. NGS using the Ion Torrent PGM™ offers scalability to examine as many individuals as is necessary. The cost of the sequencing run included the template preparation using the Ion OneTouch™ 2 system as well as the sequencing cost of the Ion PGM™ platform.

Table 5.4 Number of NGS samples can be achieved for *KEL* gene using different chips.

Chip type	Capacity	Number of samples	Cost per run
Ion 314™	30-50 Mb	20	£271.55
Ion 316™	300 Mb-500 Mb	204	£444.80
Ion 318™	600 Mb – 1Gb	408	£619.67

A list of the three types of Ion chip, their capacities and the number of samples can be obtained with a depth of coverage of 100× in addition

In summary, NGS using the approach of LR-PCR was capable of genotyping all the Kell blood group alleles among all the exons of the *KEL* gene, especially the high prevalence ones. Therefore, any low prevalence alleles of the Kell blood group system can be detected with this technique. Moreover, the intronic polymorphisms are easy to investigate and any possible mutations that may be associated with K_{null} or K_{mod} alleles can be detected. The finding of the heterozygous SNP for the *KEL*02.03* allele encoding Kp^a phenotype showed the power of this approach in overcoming the standard phenotyping by serology. Proof of principle of the LR-PCR approach for Kell blood group genotyping makes the application of this technique more feasible in complicated blood groups such as the Rh blood group system, which will be discussed in the next chapter.

Chapter 6 : Genotyping the Rh Blood Group by Next-generation Sequencing

6.1 Introduction

The Rh blood group system is the most complex blood group due to the high number of polymorphisms. These polymorphisms are in fact because this system is encoded by two highly homologous genes, *RHD* and *RHCE,* which share of identity 92% (Cartron, 1994). There are different mechanisms involved in the expression of the Rh antigens. For example, the presence of the entire *RHD* gene gives rise to the RhD-positive phenotype. On the other hand, in Caucasians, the deletion of the whole *RHD* gene leads to the RhD-negative phenotype (Avent et al., 2006). Genetic mechanisms such as point mutation, gene conversion and gene rearrangement leads to weak D and partial D, hybrid genes and *RHCE* variants including C^X and C^W antigens (Avent et al., 2006). Such variants lead to alloimmunisation in patients with SCD [see section 1.5 for further detail].

BGG has aided in recent years to provide more compatible blood units in comparison to the conventional serology. There are numerous techniques have been developed following the cloning of most the blood group genes. SSP-PCR is a simple technique and is robust, but in some ambiguous cases it is not sufficient to type the variants and requires sequencing (Prager, 2007). Therefore, the sequencing-based typing may be the optimal solution to genotype the Rh blood group system. Array-based methods have been developed to facilitate the genotyping work for both donors and patients (Hashmi et al., 2005; Avent et al., 2007). The obstacles of these array-based platforms are that they cannot identify new alleles and they need to be updated for the newly discovered alleles (Avent et al., 2015).

6.2 Aims

Here, the Rh blood group genes were entirely amplified and sequenced in high resolution on the Ion PGM™ platform. In the first instance, the *KEL* gene worked as a model and was amplified by two products and was fully sequenced on Ion PGM™ [see Chapter 5]. Such an approach gave a high resolution to genotype all the Kell antigens in addition to the high prevalence ones [see Chapter 5]. The same approach was applied to the *RHD* and *RHCE* genes in order to resolve the complicated variants of the Rh blood group system. Furthermore, NGS with LR-PCR is able to identify new alleles and works in 'a discovery mode' in contrast to array-based technology (Avent et al., 2015).

6.3 Results

6.3.1 Long range PCR for Rh genes

10 samples were chosen for this experiment according to serology typing, five RhD-positive samples and five weak D samples. Table 6.1 shows the Rh status of these samples by conventional serology. A total of seven primer pairs were used to amplify the entire genes of the Rh blood group system which ranged from 11,215 bp to 24,387 bp. Three primer pairs were used for the *RHD* gene and four for the *RHCE* gene. The primer pairs are listed in Table 2.3 and Table 2.4. Figure 6.1 illustrates a schematic diagram for the primer pairs used to amplify both genes. LR-PCR was performed using PrimeSTAR GXL Polymerase (Takara, Japan) as explained in [section 2.5.2.3]. Figure 6.2 shows the results of the PCR products for the Rh blood group system run on 0.5% agarose gel electrophoresis. The figure demonstrates the negative control for the *RHD* gene, which was RhD-negative sample and shows no bands [Figure 6.2 a]. The three bands to amplify the entire of the *RHD* gene are shown in [Figure 6.2 b]. The products of the *RHCE* gene are displayed in [Figure 6.2 c].

Table 6.1 Serology information provided by NHSBT Filton for the Rh status.

#	D status	Rh	D	C	E	c	e	C^w	Defined zygosity[§]
1	D+	R_0r	+	−	−	+	+	−	Hemizygous
2	D+	R_1R_1	+	+	−	−	+	−	Homozygous
3	D+	R_1R_1	+	+	−	−	+	−	Homozygous
4	D+	R_2R_2	+	−	+	+	−	−	Homozygous
5	D+	R_2R_2	+	−	+	+	−	−	Homozygous
6	Weak D	R_0r	+	−	−	+	+	−	Hemizygous
7	Weak D	$R_1R_2^*$	+	+	+	+	+	−	Hemizygous
8	Weak D	R_2r	+	−	+	+	+	−	Hemizygous
9	Weak D	$R_1R_2^*$	+	+	−	+	+	−	Hemizygous
10	Weak D	R_1r	+	+	−	+	+	−	Hemizygous

§ = The defined zygosity for these samples were performed by Ms. Kelly Sillence. Samples 7 and 9 were typed as weak D by serology and it is known that weak D samples have only one copy of the *RHD* gene. The zygosity testing confirmed that samples 7 and 9 have a single copy of the *RHD* gene. The phenotype of sample 7 and 9 may be R_2r', R_1r'', R_zr or R_0r_y.

* Weak D samples should never be described as *RHD* homozygous as they are by definition hemizygous, unless in very rare instances two variants *RHD* alleles are co-inherited.

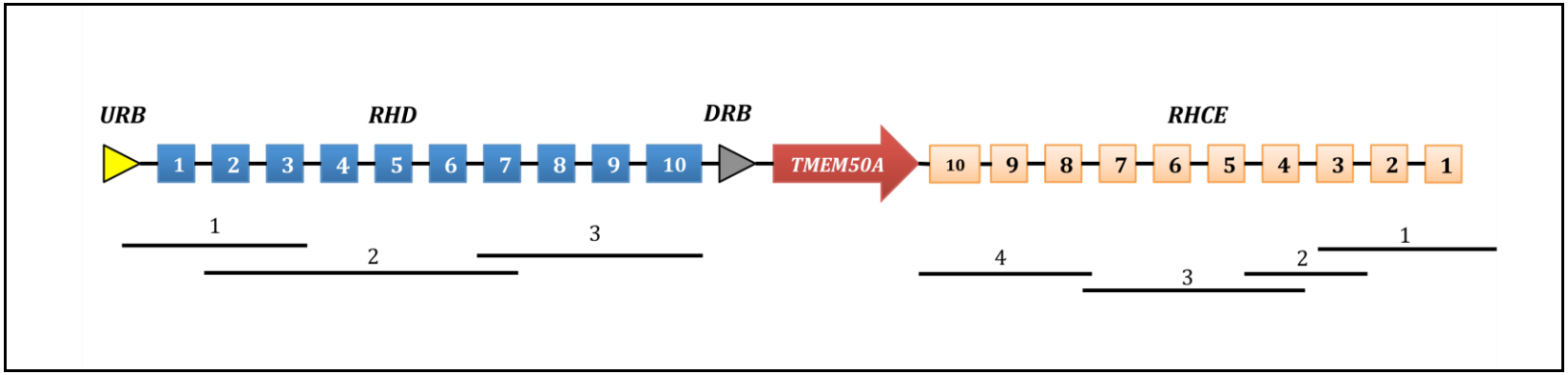


Figure 6.1 Seven LR-PCR products covered the entire of the *RHD* and *RHCE* genes.

The boxes show the exons (coding areas) of the genes, in which each gene has 10 exons, while the lines attached to the boxes are the introns (non-coding areas). *TMEM50A* is a gene flanked by the *RHD* and *RHCE* genes. It was previously called *SMP1*.
URB= upstream rhesus box.
DRB= downstream rhesus box.

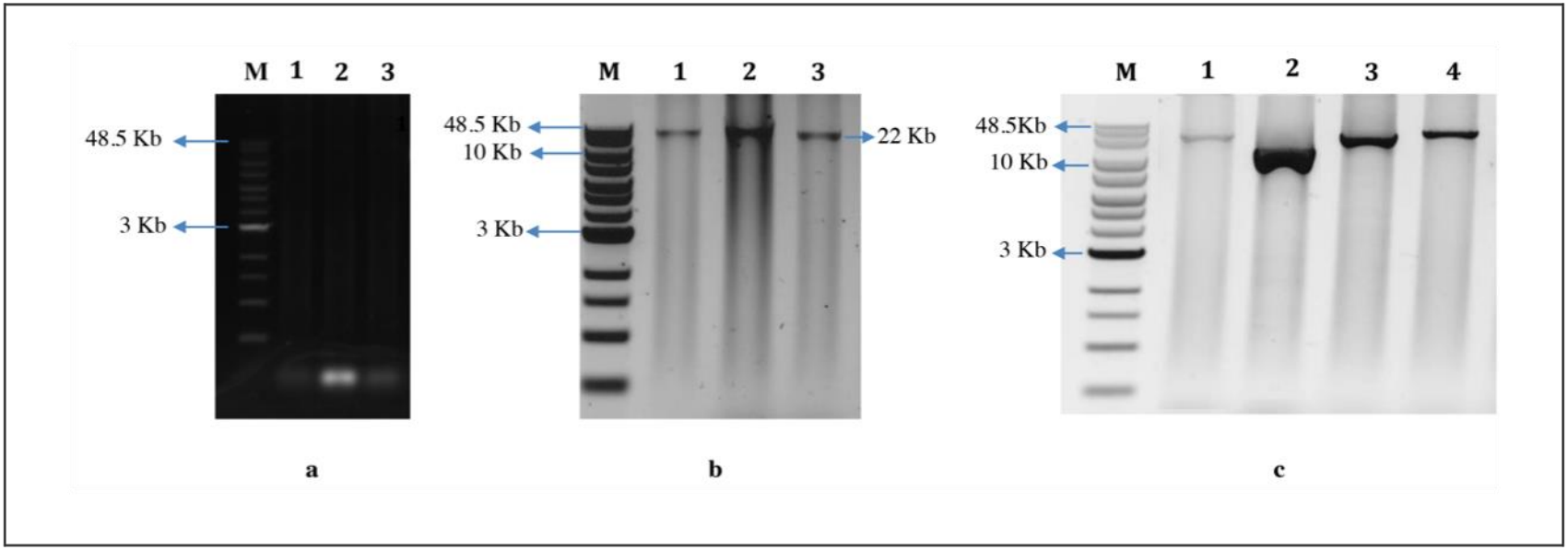


Figure 6.2 LR-PCR products for the Rh blood group system.

(a) The negative control for the *RHD* gene, which is the RhD-negative sample shows no bands in the three lanes 1, 2 and 3.
(b) Three long-range amplicons (1, 2, and 3) were designed to amplify the entire of the *RHD* gene. The figure shows three products were amplified from the RhD-positive samples in which each product was about 22 Kb.
(c) Four long-range amplicons (1, 2, 3, and 4) were amplified the entire *RHCE* gene from a single sample. The products range as following; 1= 19 Kb, 2= 11.2 Kb, 3= 17.8 Kb and 4= 24.3 Kb.

6.3.2 Sequencing Libraries

6.3.2.1 Fragmentation

Following the purification of the seven LR-PCR products of the Rh blood group system [section 2.5.2.4], the products of each gene were pooled separately from the other gene. In other words, the three products of the *RHD* gene were pooled in a single tube and the four products of the *RHCE* gene were pooled in another one [section 2.5.2.5]. Then, the sequencing libraries were constructed using the Ion Xpress™ Plus Fragment Library Kit. The shearing enzyme was used for 15 minutes regarding the 200 bp sequencing reading length on the Ion PGM™ [section 2.5.2.6].

Following the fragmentation procedure, the purification was performed utilising the Agencourt® AMPure® XP reagent. The high sensitivity DNA Kit was used in order to evaluate the fragmented DNA which was run on the Agilent® 2100 Bioanalyzer [section 2.5.2.7]. Figure 6.3 and Figure 6.4 demonstrate the results of the fragmented DNA for the *RHD* and *RHCE* genes, respectively.

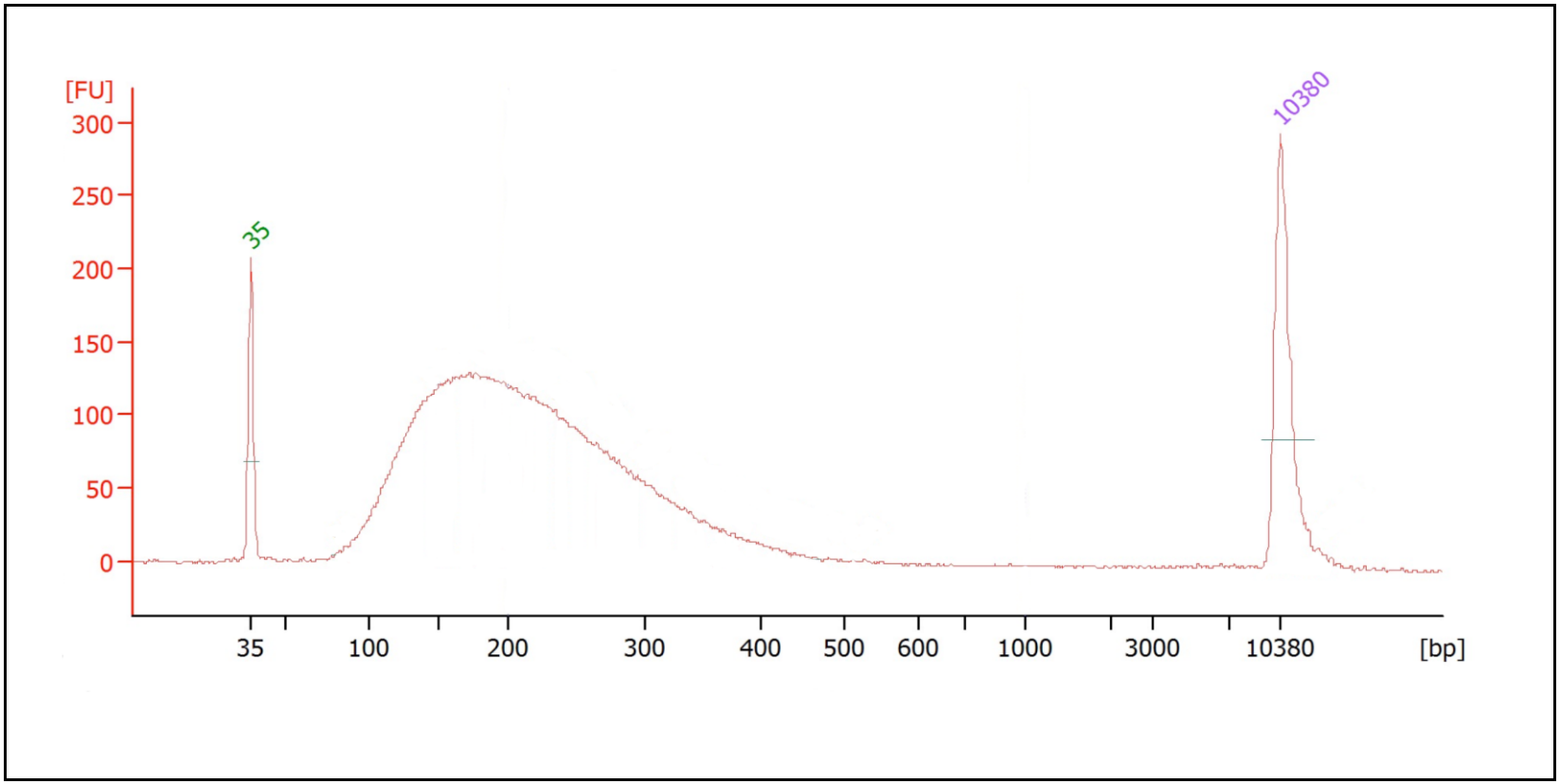


Figure 6.3 An electropherogram of the fragmented *RHD* LR-PCR products.

These fragments range (82-511 bp) using the Ion Xpress™ Plus Fragment Library Kit. The shearing enzyme of the kit was used for 15 minutes. The assay was assessed using the high sensitivity DNA Kit which was run on the Agilent® 2100 Bioanalyzer instrument. The green and purple numbers show the lower and upper markers in bp, respectively.

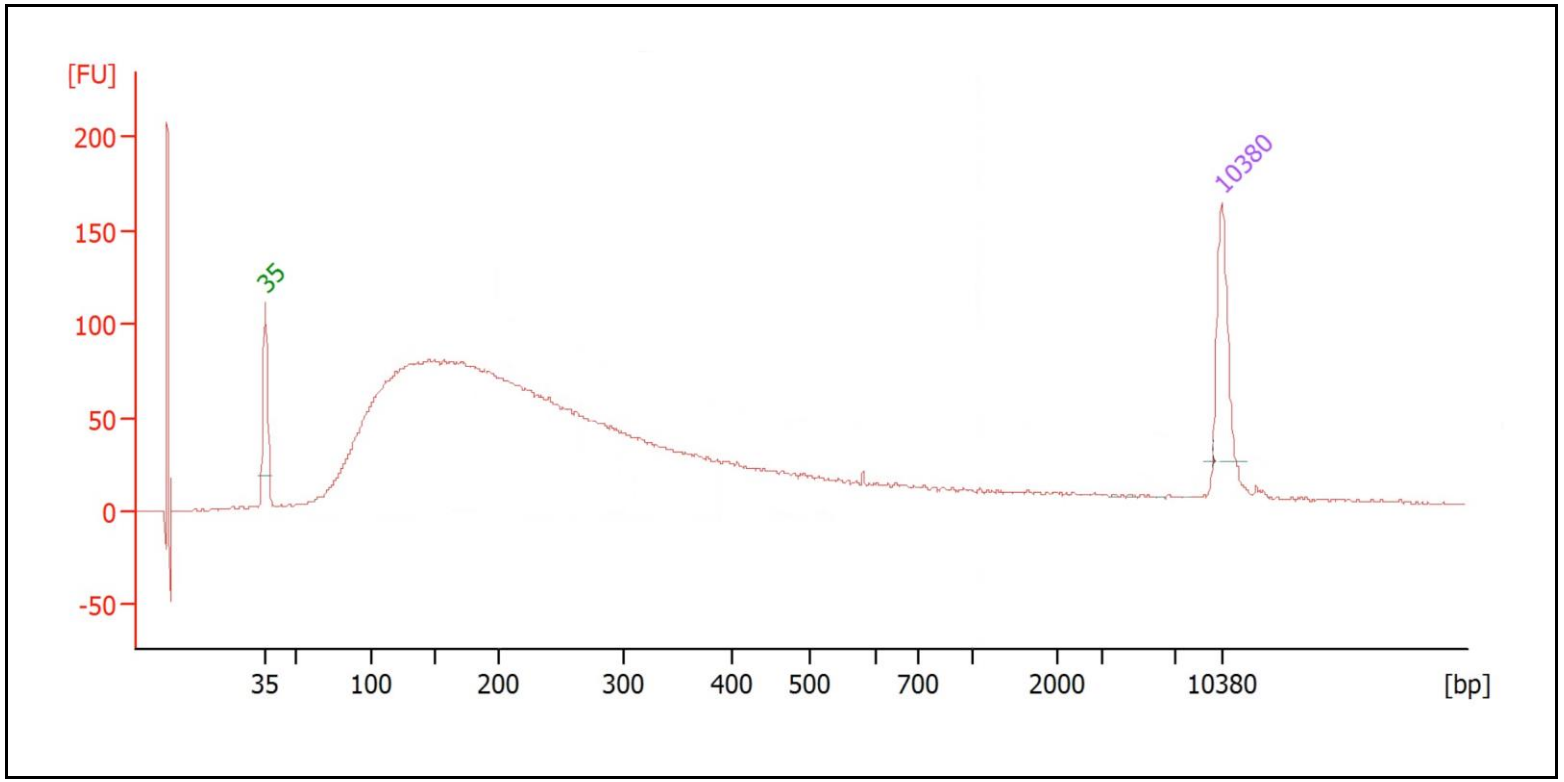


Figure 6.4 An electropherogram of the fragmented *RHCE* LR-PCR products.

The fragments range (50-522 bp) using the Ion Xpress™ Plus Fragment Library Kit. The shearing enzyme of the kit was used for 15 minutes. The assay was assessed using the high sensitivity DNA Kit which was run on the Agilent® 2100 Bioanalyzer instrument. The green and purple numbers show the lower and upper markers in bp, respectively.

6.3.2.2 Ligation

Following the purification of the fragmented DNA of both genes, the products were ligated to P1 and barcoded adaptors as discussed in [section 2.5.2.8]. After that, the ligated products were purified using the Agencourt® AMPure® XP reagent. Figure 6.5 and Figure 6.6 demonstrate the assessment runs on the Agilent® 2100 Bioanalyzer instrument of the purified ligated products for both the *RHD* and *RHCE* genes using the high sensitivity DNA Kit.

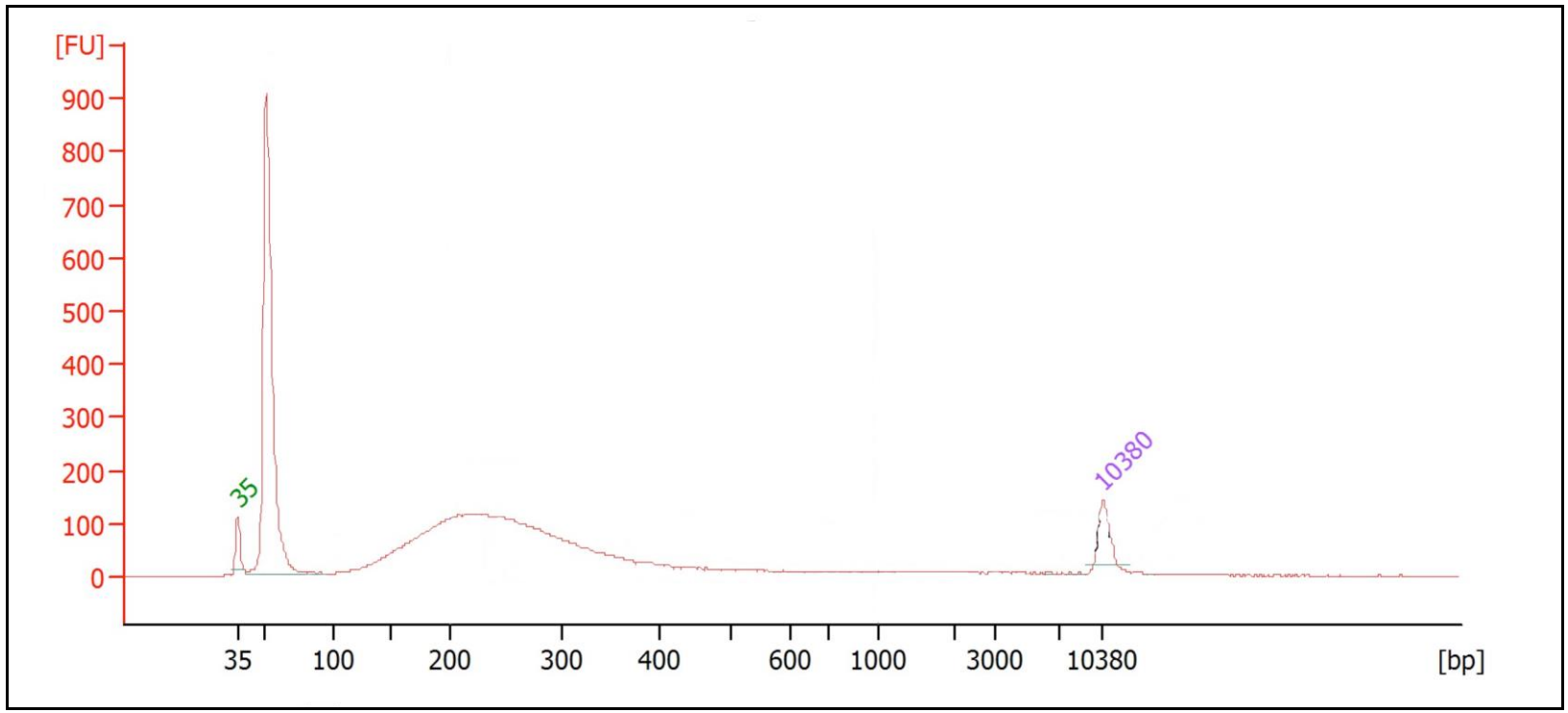


Figure 6.5 An electropherogram of the ligation of the *RHD* sequencing library

A peak of approximately 208 bp can be seen in this figure. The ligation was performed using the Ion Xpress™ Plus Fragment Library Kit. The assay was assessed using the high sensitivity DNA Kit that runs on the Agilent® 2100 Bioanalyzer instrument. The green and purple numbers show the lower marker and upper marker in bp, respectively.

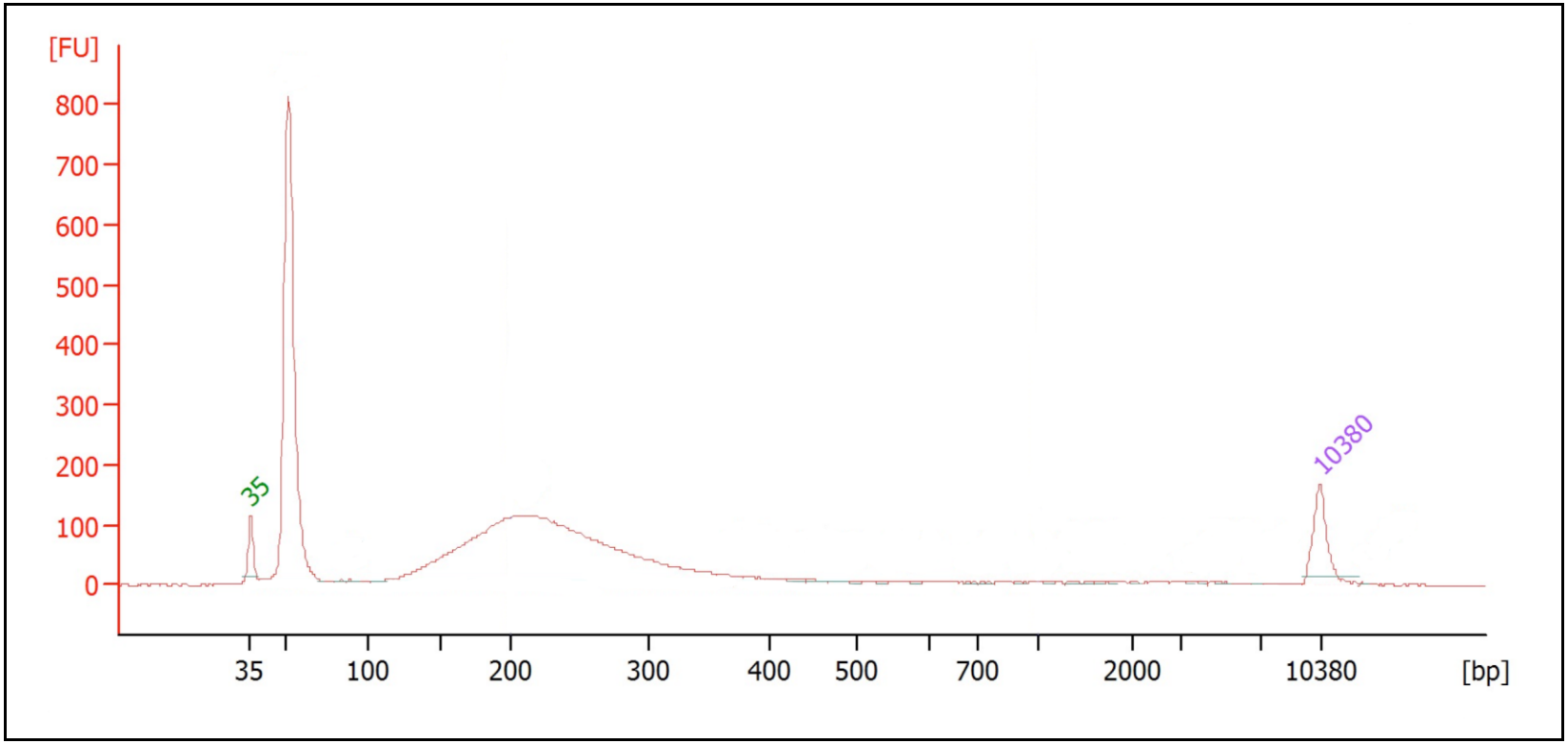


Figure 6.6 An electropherogram of the ligation of the *RHCE* sequencing library.

A peak of approximately 207 bp can be seen in this figure. The ligation was performed using the Ion Xpress™ Plus Fragment Library Kit. The assay was assessed using the high sensitivity DNA Kit that runs on the Agilent® 2100 Bioanalyzer instrument. The green and purple numbers show the lower marker and upper marker in bp, respectively.

6.3.2.3 Size selection using SPRIselect® magnetic beads

The size selection of the sequencing libraries was carried out for both the *RHD* and *RHCE* genes using SPRIselect® beads [section 2.5.2.8]. Figure 6.7 and Figure 6.8 show the results of the size selection for *RHD* and *RHCE* samples, respectively. The targets of the size selection ranged from 100 to 300 bp regarding 200 bp sequencing read lengths.

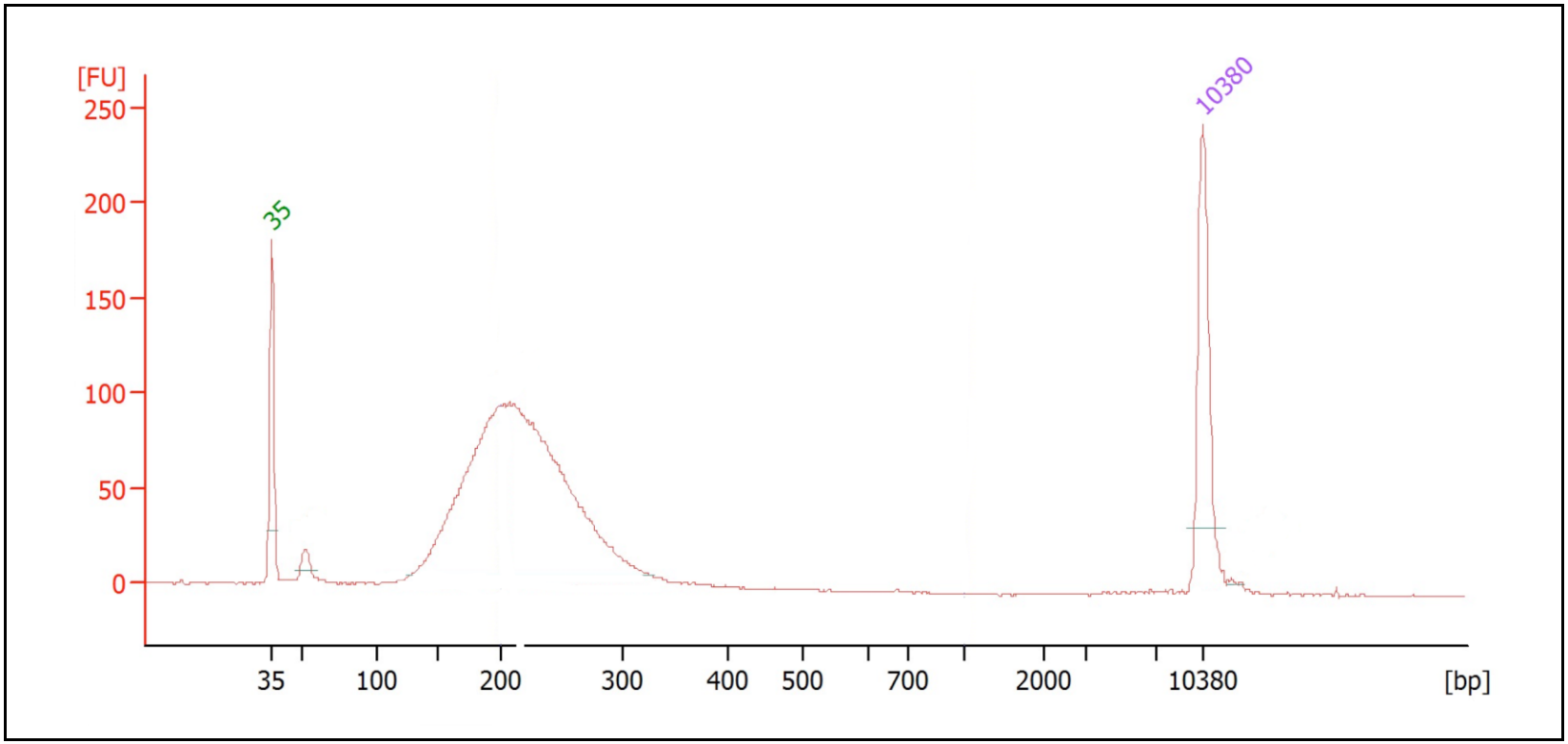


Figure 6.7 An electropherogram of the size-selected *RHD* sequencing library.

A peak of 207 bp was achieved for a 200 bp reading length. This procedure of size selection was carried out using SPRIselect® magnetic beads. The assay was assessed using the high sensitivity DNA Kit that runs on the Agilent® 2100 Bioanalyzer instrument. The green and purple numbers show the lower marker and upper marker in bp, respectively.

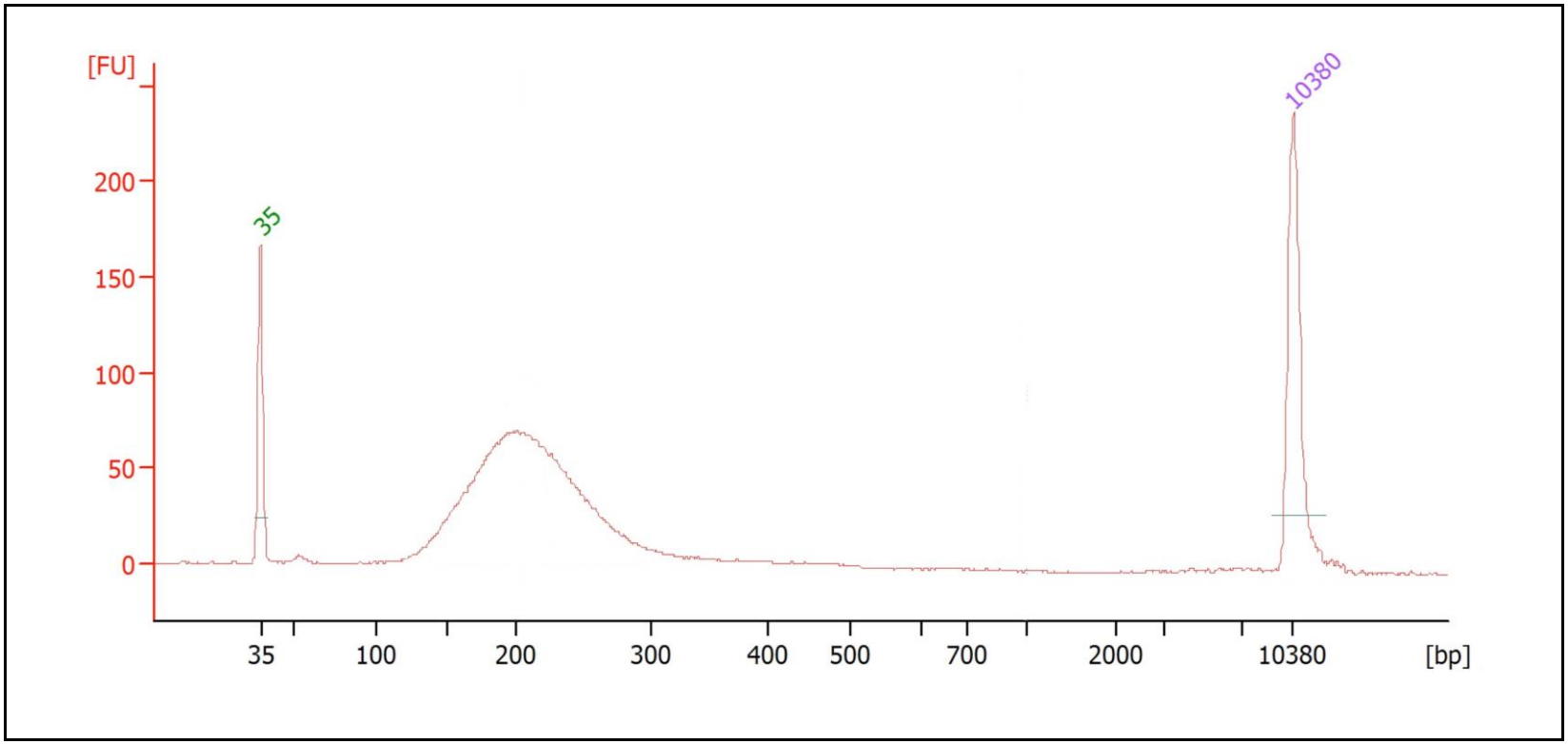
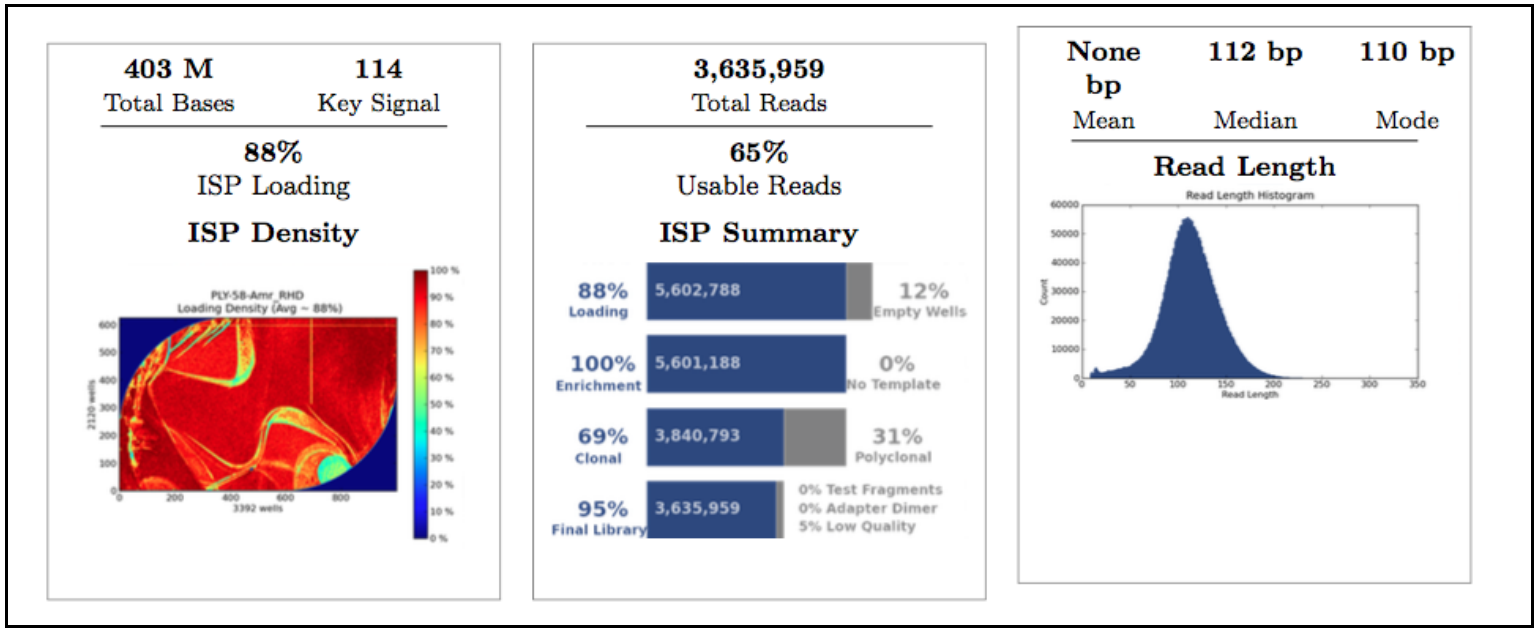


Figure 6.8 An electropherogram of the size-selected *RHCE* sequencing library.

A peak of 201 bp was achieved for a 200 bp reading length. This procedure of size selection was performed using SPRIselect® magnetic beads. The assay was assessed using the high sensitivity DNA Kit that runs on the Agilent® 2100 Bioanalyzer instrument. The green and purple numbers show the lower marker and upper marker in bp, respectively.

6.3.3 Analysis of NGS sequencing data

Ten samples in total were sequenced for both the *RHD* and *RHCE* genes in a single sequencing run on the Ion PGM™. Figure 6.9 demonstrates a summary of the sequencing report. The loading density was 88% in which the ISPs was addressed by the chip wells in comparison to 12% of the empty wells. The total reads were 3,635,959, of which the usable reads were 65%. The enrichment of the clonal templates was successfully achieved a high percentage of 100%. The clonal amplification achieved 69% in comparison to polyclonal amplification, which was 31%. The percentage of the final library was 95% of which 5% was a low quality library. Finally, the median of the read length was 112 bp [Figure 6.9].



(a) (b) (c)

Figure 6.9 An overview of a report on the sequencing run of the genes of the Rh blood group system.

(a) The loading density of the ISP that was addressed by the chip wells was 88%.
(b) Total of 3,635,959 reads after filtering and trimming.
(c) A histogram of the read length of the sequencing libraries with a median of 112 bp, in which the y-axis demonstrates the read count, while the x-axis shows the read length in the base pair.

6.3.4 Quality Control

6.3.4.1 Per base sequence quality

The quality control of the generated sequences of the *RHD* and *RHCE* genes was assessed using FastQC software. Figure 6.10 shows the Phred score across the *RHD* sequences position in bp. The quality of *RHD* sequencing started at a Phred score of 32 and decreased gradually until it reached 24 at the end of the sequencing.

Regarding the *RHCE* gene, the Phred score of the quality started at a score of 32 and then decreased gradually to reach a score of 22 by the end of the sequencing. Figure 6.11 demonstrates the Phred score across the sequences of *RHCE* in bp.

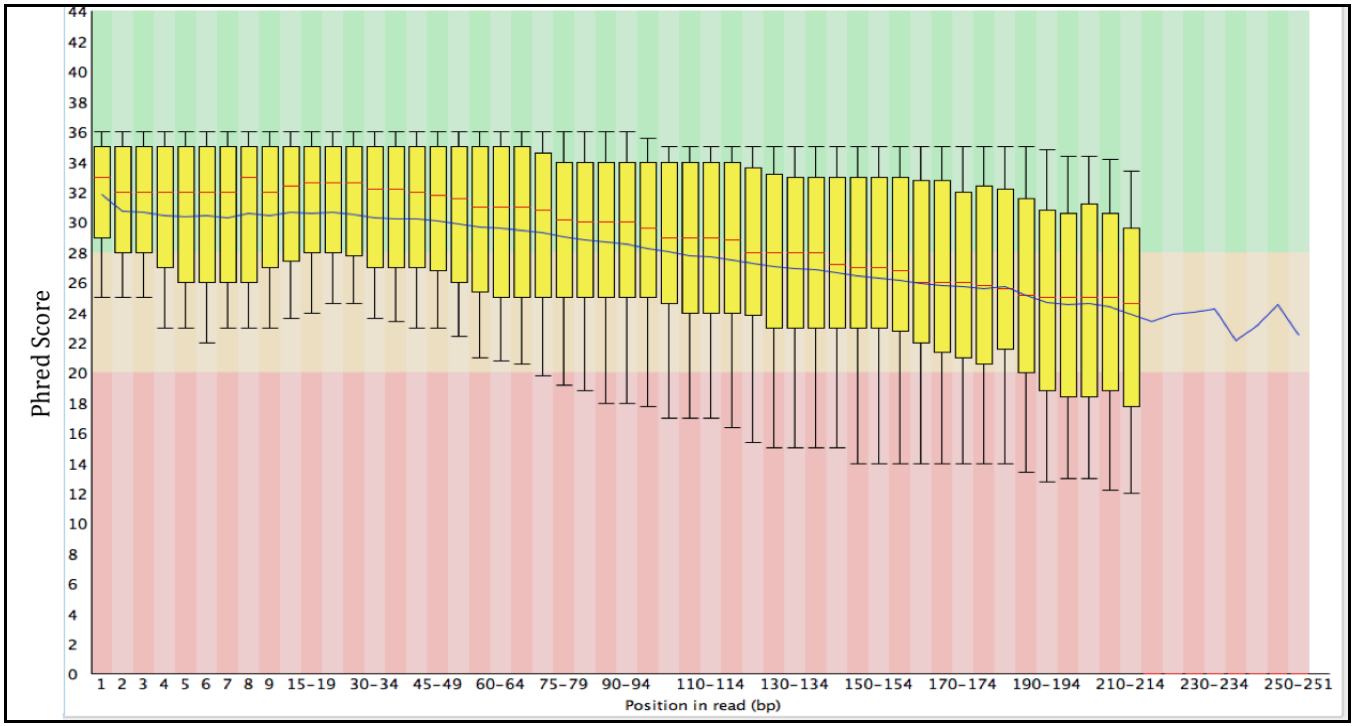


Figure 6.10 Phred quality scores across all the bases for a single sample of the *RHD* gene.

The x-axis represents the position of the reads in the bp, while the y-axis represents the Phred score of quality. The background of the y-axis is divided into three colour areas: green for the best reads, orange for the reasonable reads and red for the low quality reads. Box and Whisker plots were drawn on every position per bp. The inter-quartile range (25-75%) was represented by yellow boxes. The upper and lower whiskers demonstrate 10% and 90% points. The blue line shows the mean of the Phred quality score across the sequencing reads in the bp starting at a high quality of 32 and it was decreased gradually until it reached a score of around 24. The central red line demonstrates the median of the quality score. All the analysed samples had the same Phred score.

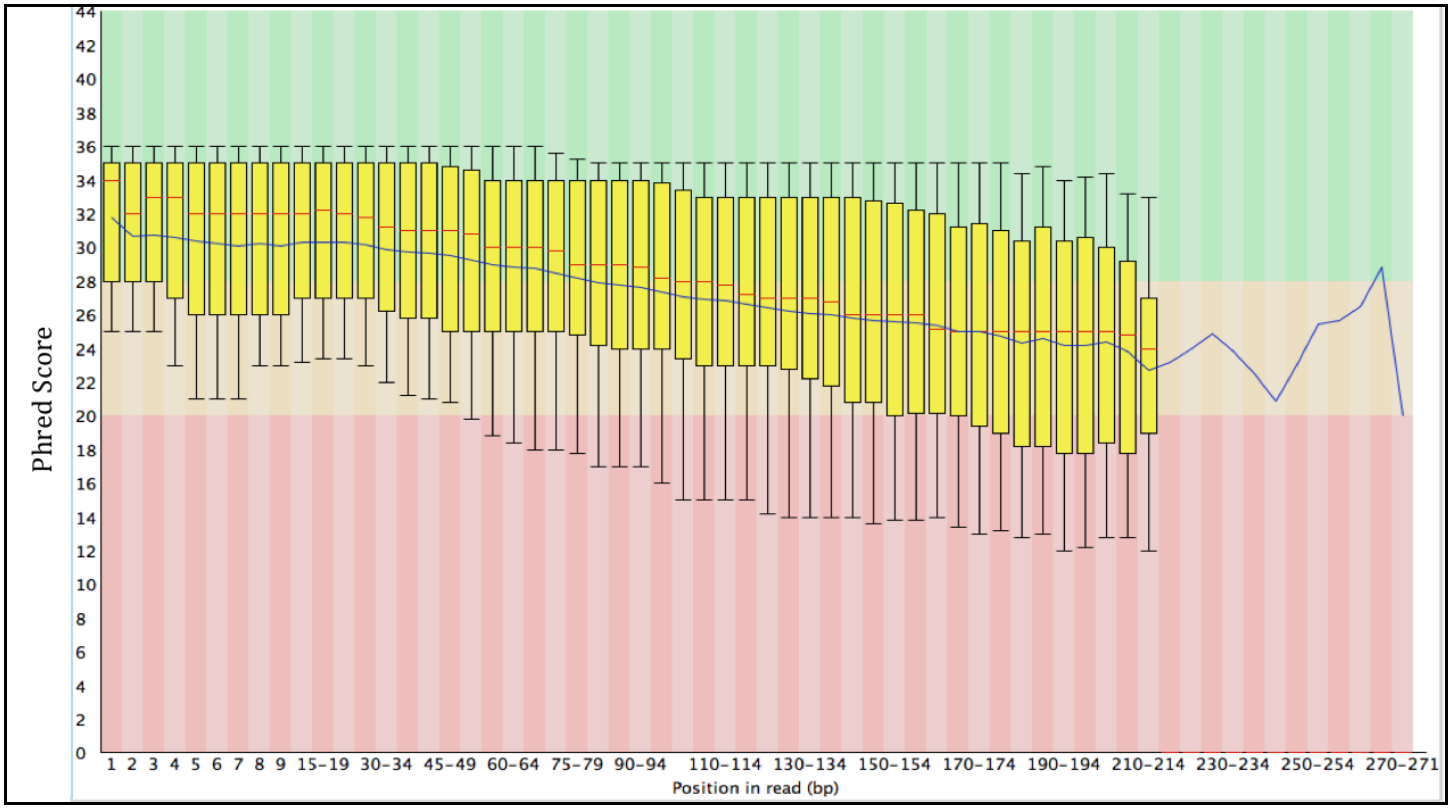


Figure 6.11 Phred quality scores across all the bases for a single sample of the *RHCE* gene.

The x-axis represents the position of the reads in the bp, while the y-axis represents the Phred score of quality. The background of the y-axis is divided into three colour areas: green for the best reads, orange for the reasonable reads and red for the low quality reads. Box and Whisker plots were drawn on every position per bp. The inter-quartile range (25-75%) was represented by yellow boxes. The upper and lower whiskers demonstrate 10% and 90% points. The blue line shows the mean of the Phred quality score across the sequencing reads in the bp starting at a high quality of 32 and it was decreased gradually until it reached a score of around 22. The central red line demonstrates the median of the quality score. All the analysed samples had the same Phred score.

6.3.4.2 Per Sequence Quality Scores

Another tool within the FastQC software called per sequence quality scores was used to analyse whether a subset of one of the sequencing reads represents poor quality reads. Figure 6.12 and Figure 6.13 illustrate the mean of the Phred quality sequence per reads across the count of the reads for the *RHD* and *RHCE* genes, respectively. The results of both the *RHD* and *RHCE* genes show that the quality of most reads was 30 according to the Phred score. Therefore, this indicates high quality reads were achieved with an accuracy of 99.9%.

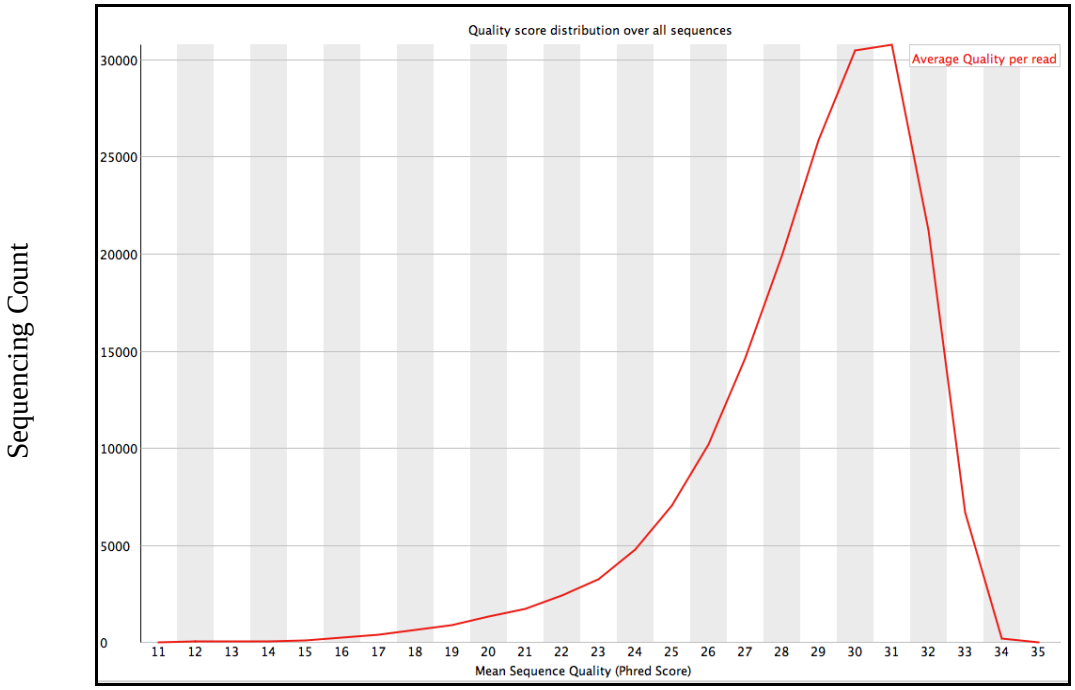


Figure 6.12 Quality scores per sequencing count for a single sample of the *RHD* gene.

The figure illustrates the mean sequence quality according to the Phred score across the read counts. It was used to assess if a subset of the sequencing reads is of poor quality. The results show that the quality of most reads was 30 according to the Phred score, which indicates high quality reads with an accuracy of 99.9%. All the analysed samples had the same quality scores.

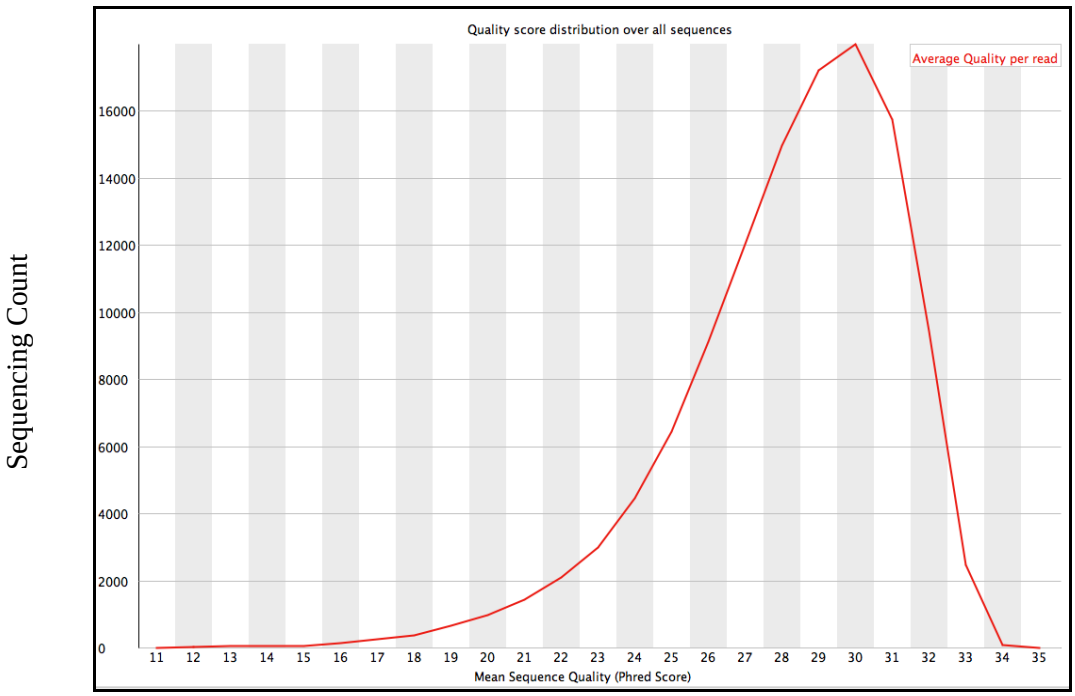


Figure 6.13 Quality scores per sequencing count for a single sample of the *RHCE* gene.

The figure shows the mean sequence quality according to the Phred score across the read counts. It was used to assess if a subset of the sequencing reads is of poor quality. The results show the quality of most reads was 30 according to the Phred score, which indicates high quality reads with an accuracy of 99.9%. All the analysed samples had the same quality scores.

6.3.5 Sequencing visualisation

Due to the extremely high homology between the *RHD* and *RHCE* genes, when the NGS data were visualised and analysed for one gene, the other needed to be masked. This procedure was kindly performed by Dr. Xinzhong Li at Plymouth University using maskfasta utility from the bedtools website as explained in [section 2.5.5.1] (Bedtools, 2015).

IGV software was utilised to visualise the sequencing data of the *RHD* and *RHCE* genes. The reference of the human genome (hg19) was used to analyse the sequencing data. The software was used to assess the whole coverage of both genes, missed regions if found, visualisation of the variants, the depth of coverage of the sequencing data and the zygosity of the SNP.

Figure 6.14 demonstrates an overview image for the entire *RHD* gene. An extremely low depth of coverage in exon 1 and intron 1 can be seen as well as some missed regions. The depth of coverage for those regions ranged between one to seven sequencing reads. In the remaining areas of the gene the depth of coverage varied from around 16 to 35 sequencing reads.

Regarding the *RHCE* gene, the area between exon 1 and exon 3 expressed a very low depth of coverage of around 8-10 sequencing reads as well as some missed regions in the areas. Figure 6.15 shows a screenshot of the IGV software for sequencing reads for the regions that covered exon 1, intron 1, exon 2 and intron 2 of the *RHCE* gene.

The regions that covered exon 3 and exon 4 show full coverage with no missed regions and high depth of coverage. The sequencing reads for those areas ranged from around 130 to 191. Figure 6.16 demonstrates a screenshot of the IGV software for the regions of exon 3 and exon 4 of the *RHCE* gene.

The depth of coverage became low again after exon 4 of the *RHCE* gene. Around 4-21 sequencing reads were obtained in the regions that covered exon 5 to exon 7. The depth

of coverage was decreased again to around four to eight sequencing reads in exons 8-10. Figure 6.17 illustrates the low depth of coverage from the regions that covered from exon 5 to exon 10 in the *RHCE* gene.

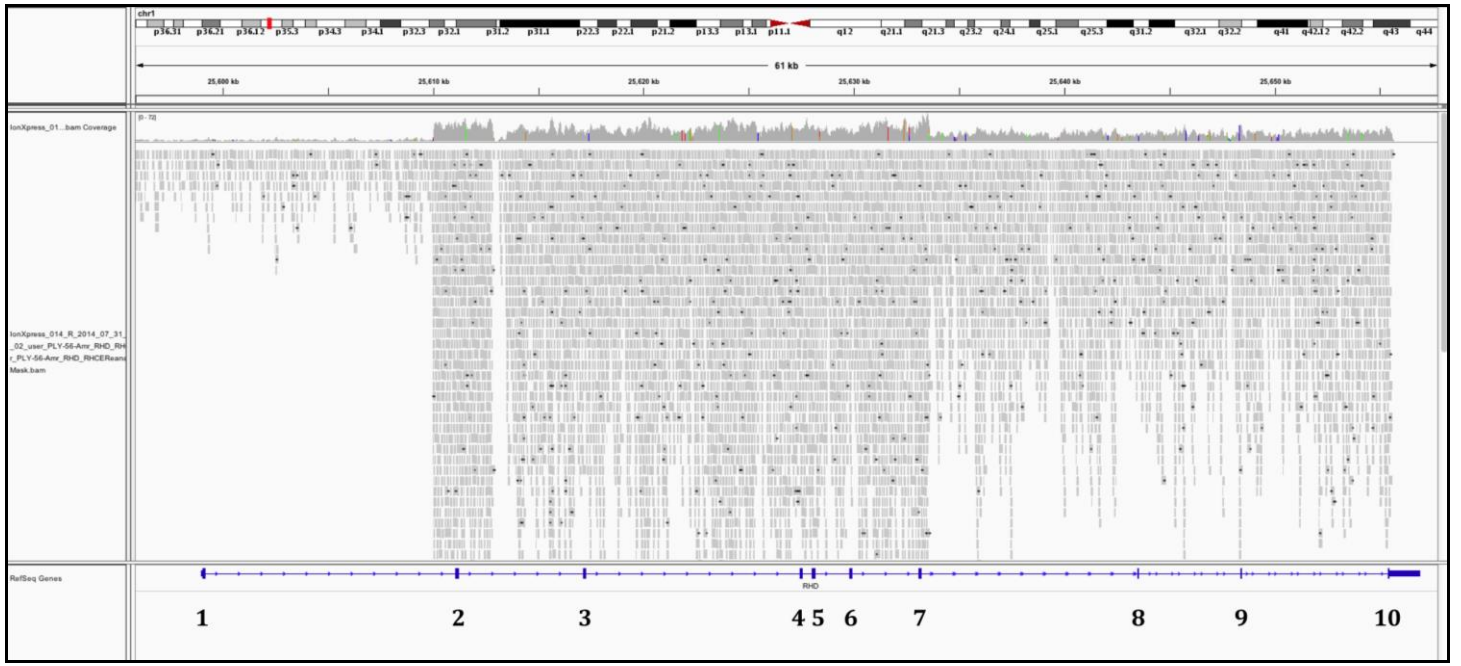


Figure 6.14 An overview image of the coverage by sequencing for the entire *RHD* gene.

The ten exons of the *RHD* gene are shown in number. There is a very low depth of coverage in exon 1 and intron 1, which was amplified by the first amplicon. This sample is the RhD-positive (R_2R_2) phenotype. Schematic diagram for the amplicons that covered both *RHD* and *RHCE* genes is demonstrated in Figure 6.1. Output from IGV software demonstrates the visualisation of the sequencing data. Chromosomal locations are shown on the top and the reference gene in blue at the bottom.

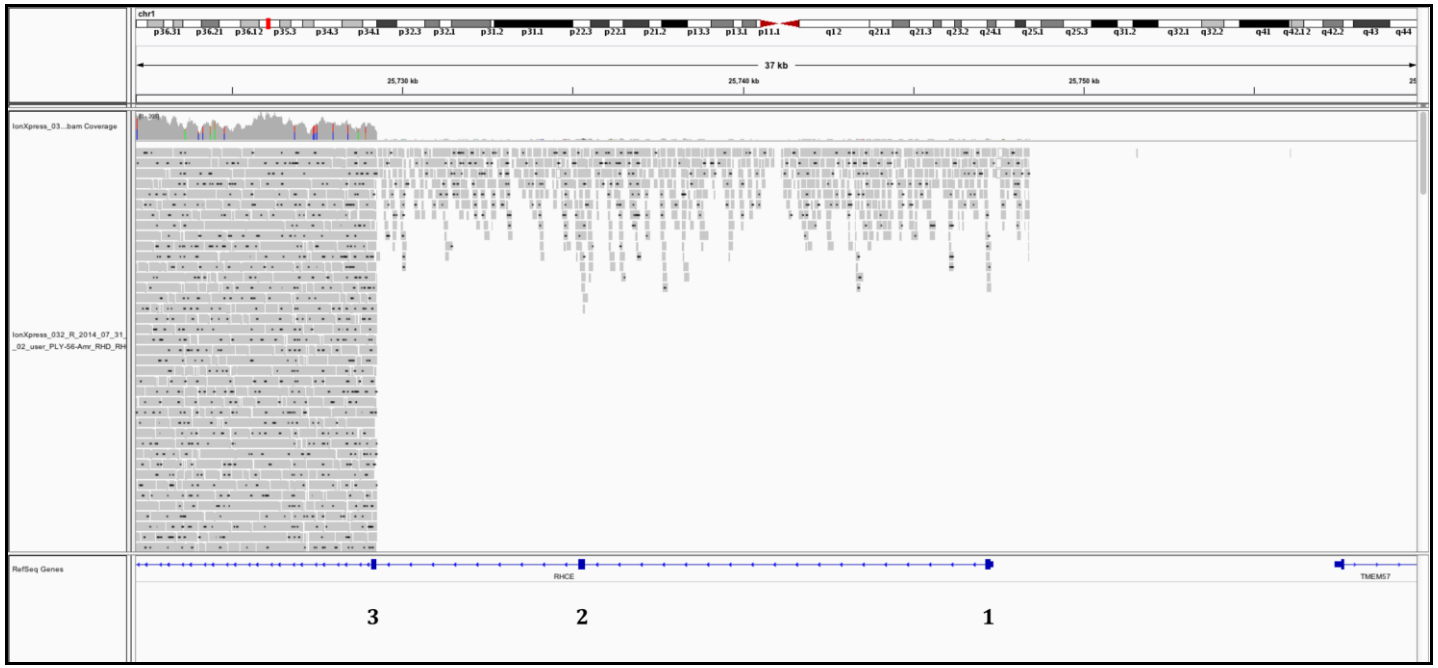


Figure 6.15 Regions with low depth of coverage for the first amplicon of the *RHCE* gene.

The covered area include (exon 1, intron 1, exon 2 and intron 2). The sequencing reads express a very low depth of coverage of around 8-10 sequencing reads as well as some missed regions in that area. This sample is a weak D sample with R_1r phenotype. Schematic diagram for the amplicons that covered both *RHD* and *RHCE* genes is demonstrated in Figure 6.1. Output from IGV software demonstrates the visualisation of the sequencing data. Chromosomal locations are shown on the top and the reference gene in blue at the bottom.

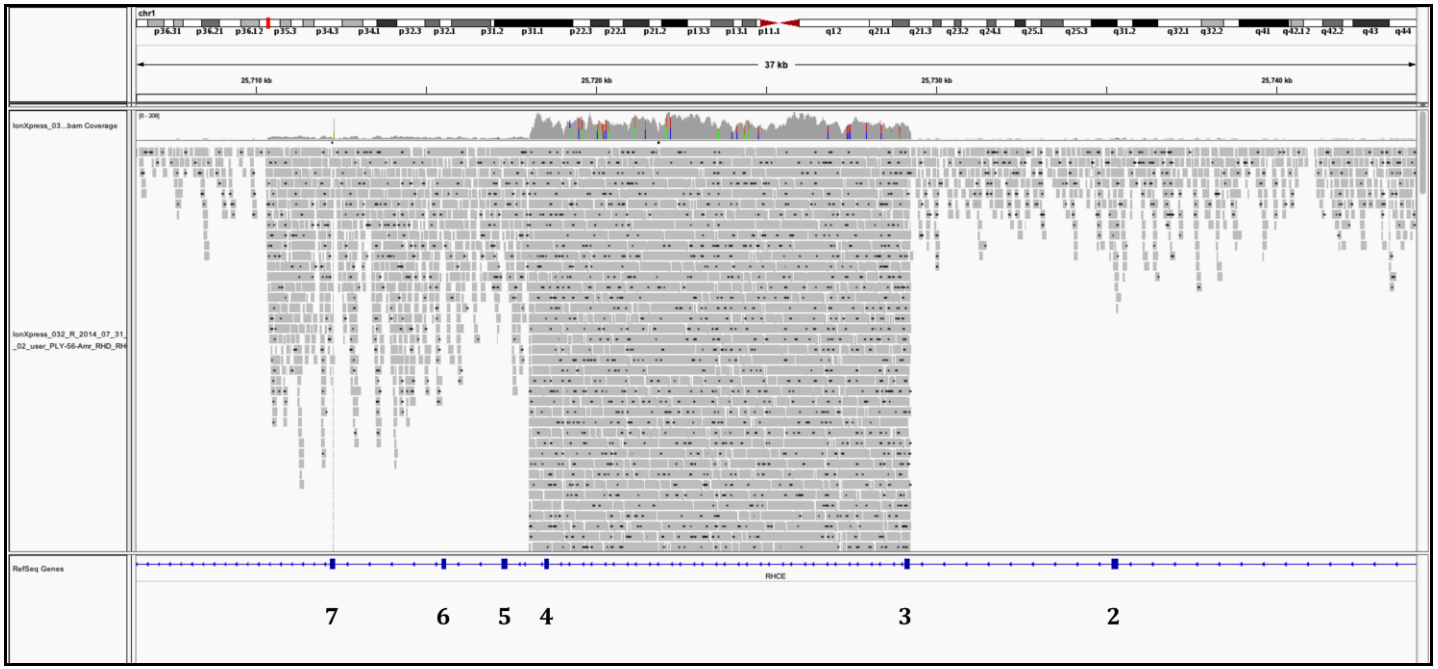


Figure 6.16 High depth of coverage for the regions that covered exon 3 and exon 4 of the *RHCE* region by the second amplicon.

This sample is a weak D sample with R_1r phenotype. Schematic diagram for the amplicons that covered both *RHD* and *RHCE* genes is demonstrated in Figure 6.1. Output from IGV software demonstrates the visualisation of the sequencing data. Chromosomal locations are shown on the top and the reference gene in blue at the bottom.

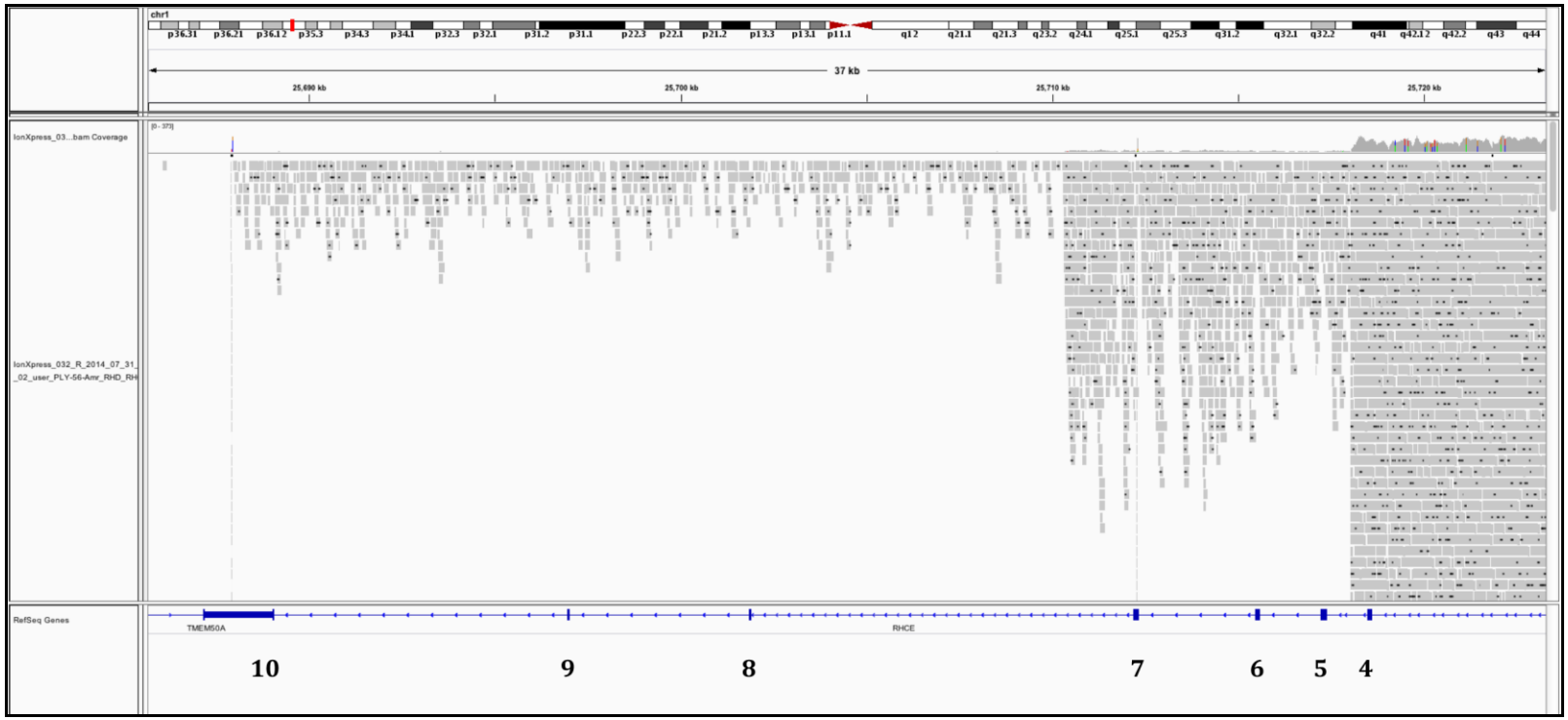


Figure 6.17 Very low depth of coverage in the *RHCE* gene in the regions between exon 5 and exon 10.

These regions were amplified by the third and fourth amplicons. This sample is a weak D sample with R_1r phenotype. Schematic diagram for the amplicons that covered both *RHD* and *RHCE* genes is demonstrated in Figure 6.1. Output from IGV software demonstrates the visualisation of the sequencing data. Chromosomal locations are shown on the top and the reference gene in blue at the bottom.

6.3.5.1 Obstacles of the visualisation

RhD-positive samples should observe no mutations in any of the exons of the *RHD* gene. Interestingly, all the samples were found to have the mutation of *RHD*DAU* allele in exon 8 1136C>T (Thr379Met). In fact, the reference gene used by the IGV software was the transcript variant 2 [NCI Reference Sequence: NM_001127691]. This reference gene shows the 1136C>T (Thr379Met) instead of 1136T>C (Met397Thr). Figure 6.17 shows the misalignment of the *RHD*DAU* mutation in exon 8 of the *RHD* gene due to the different reference allele used by the IGV software. Therefore, all the RhD-positive samples did not observe any mutation in any of the exons of the *RHD* gene.

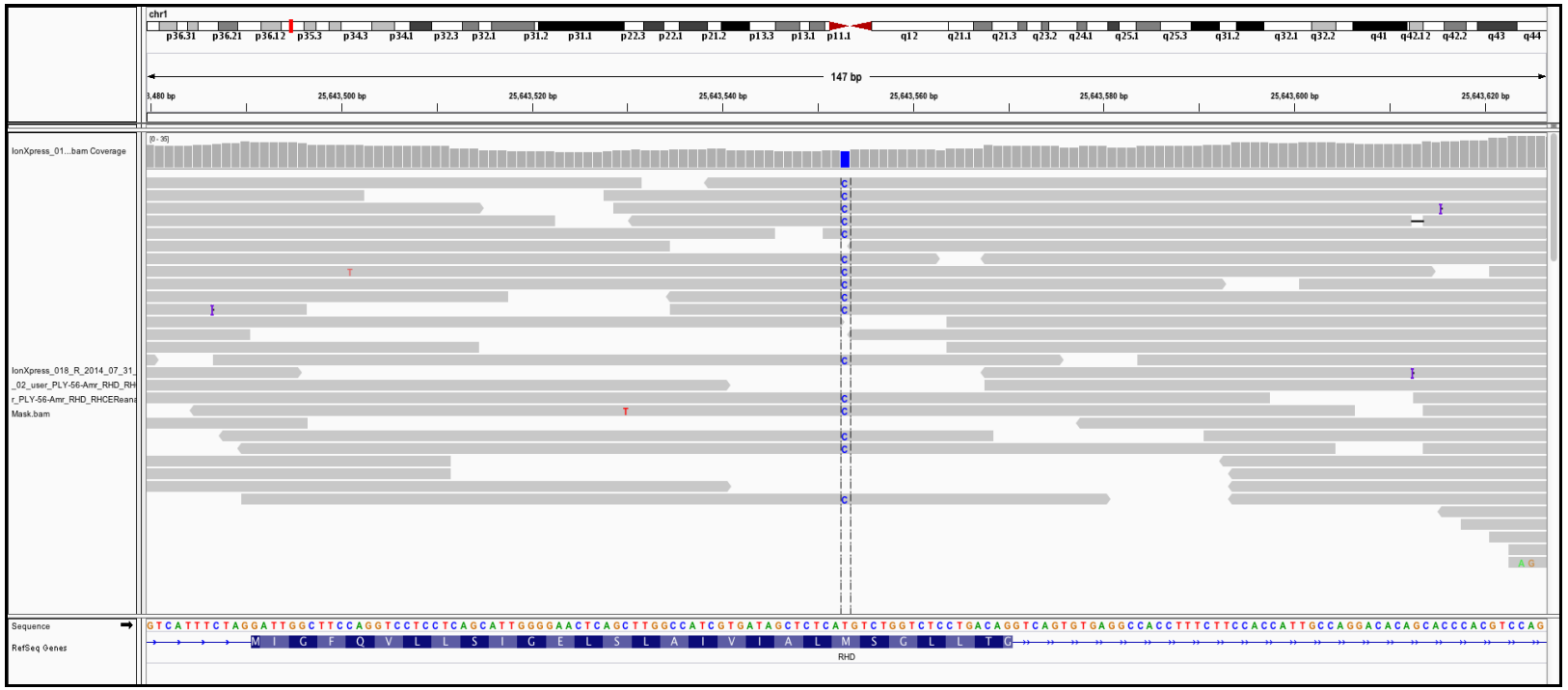


Figure 6.18 A misalignment in data visualisation of exon 8 of the *RHD* gene due to the different reference allele used by the IGV software.

The reference gene used by the IGV software was the transcript variant 2 in the human genome (hg19) [NCI Reference Sequence: NM_001127691]. This reference gene shows the 1136C>T (Thr379Met) which is the *RHD*DAU* allele instead of 1136T>C (Met397Thr). This sample is the R_1R_2 phenotype. Output from IGV software demonstrates the visualisation of the sequencing data. Chromosomal locations are shown on the top and the reference gene in blue at the bottom.

6.3.6 Variant analysis

VCF files were produced to annotate the sequencing data against the reference gene using the VariantCaller plugin in the Torrent Suite Software. Then, the VCF files were uploaded for annotation using the SeattleSeq Annotation 141 website (SeattleSeq Annotation Tool 141, 2014). The results of the genotyping are presented in the next section.

6.3.7 Genotyping of the Rh blood group system

6.3.7.1 Genotyping of the RHD gene

Ten samples were fully sequenced on the Ion PGM™ following amplification by LR-PCR. Five samples were typed by serology as RhD-positive samples, while the other five were typed as weak D samples. Regarding the RhD-positive samples, there were no mutations detected in any of the 10 exons of the *RHD* gene.

With respect to the weak D samples, two samples were observed to be weak D type 1 (*RHD*01W.01)* 809T>G Val270Gly. Furthermore, two samples were detected to be weak D type 2 *(RHD*01W.02)* 1154G>C at exon 9 (Gly385Ala). Figure 6.19 shows the SNP of the weak D type 2 on the IGV software. The last sample was shown to be DAR3.1 weak partial D 4.0 *(RHD*DAR3.01*). The last sample had three altered nucleotides: 602C>G (Thr201Arg) in exon 4, 667T>G (Phe223Val) in exon 5 and 819G>A (silent) in exon 6. Table 6.2 lists the genotyping results of the weak D samples found by NGS.

Sixty-four intronic SNPs in the 10 samples were found in the *RHD* gene. The hemizygous SNPs were associated with weak D samples. On the other hand, homozygous and heterozygous SNPs were found in the RhD-positive samples. Table 6.3 demonstrates the intronic SNPs found in the *RHD* gene with different phenotypes.

6.3.7.2 Genotyping of the RHCE gene

Table 6.4 lists the genotyping results of the *RHCE* gene. The reference gene used in the human genome assembly (hg19) is *RHCE*ce* allele. The alleles of *RHCE*c* and *RHCE*e* matched the reference gene while the variants of the other alleles, *RHCE*C* and *RHCE*E*, were called. Two samples of R_1R_1 were found to be homozygous for 307C>T (Pro103Ser) in exon 2 of the *RHCE* gene. This confirmed the serology and was genotyped as *RHCE*C* allele. The samples were predefined by serology as R_2R_2 and were found to be homozygous for the *RHCE*E* allele and had a homozygous SNP 676G>C (Ala226Pro) in exon 5 of the *RHCE* gene. Two samples of R_1R_2 were identified by conventional serology. The first sample was found to be heterozygous for 307C>T (Pro103Ser) in exon 2 and had a heterozygous SNP for 676G>C (Ala226Pro). The second sample had three SNPs for the *RHCE*C* allele. These include 178C>A (Leu60Ile), 203A>G (Asn68Ser) and 307C>T (Pro103Ser) and all were in exon 2 of the *RHCE* gene. In addition, it was found to have a heterozygous SNP for 676G>C (Ala226Pro).

Regarding the R_0r sample, it had four heterozygous SNPs. The first SNP was in exon 1 48G>C (Trp16Cys). The second SNP was a heterozygous 733C>G (Leu245Val) in exon 5. The third SNP was a heterozygous 1025C>T (Thr342Ile) in exon 7. The outcome of the three SNPs gives *RHCE*ceVS.04* allele, which encodes partial e V+ VS+ antigens. The fourth SNP was in exon 5 as heterozygous 744T>C and was found to be a silent mutation.

The R_1r sample was identified by serology. It shows the three heterozygous SNPs of the *RHCE*C* allele 178C>A (Leu60Ile), 203A>G (Asn68Ser) and 307C>T (Pro103Ser) and all were in exon 2 of the *RHCE* gene. In addition, there was a silent mutation found 201A>G (Serine at amino acid 67) in exon 2. Furthermore, the novel allele that was identified by the HEA and HPA Panel [Chapter 4] was found and aligned properly to

the *RHCE* gene instead of the *RHD* gene. The SNP is in exon 2 of the *RHCE* and it was heterozygous 208C>T (Arg70Trp). Figure 6.20 shows all five SNPs found in the R_1r sample.

A sample was typed by serology as R_0r and due to the low depth of coverage of two sequencing reads, only *RHCE*e* could be identified by the NGS by matching the nucleotide of the reference gene. Figure 6.21 demonstrates the low depth of coverage of this sample with two sequencing reads around amino acid position 226. The last sample was identified by serology as R_2r and has been confirmed by NGS by finding the heterozygous SNP 676G>S (Ala226Pro) in exon 5. Due to the low depth of coverage, not all the intronic SNPs of the *RHCE* gene could not be identified. Introns 3-6 with high depth of coverage were identified [Table 6.5].

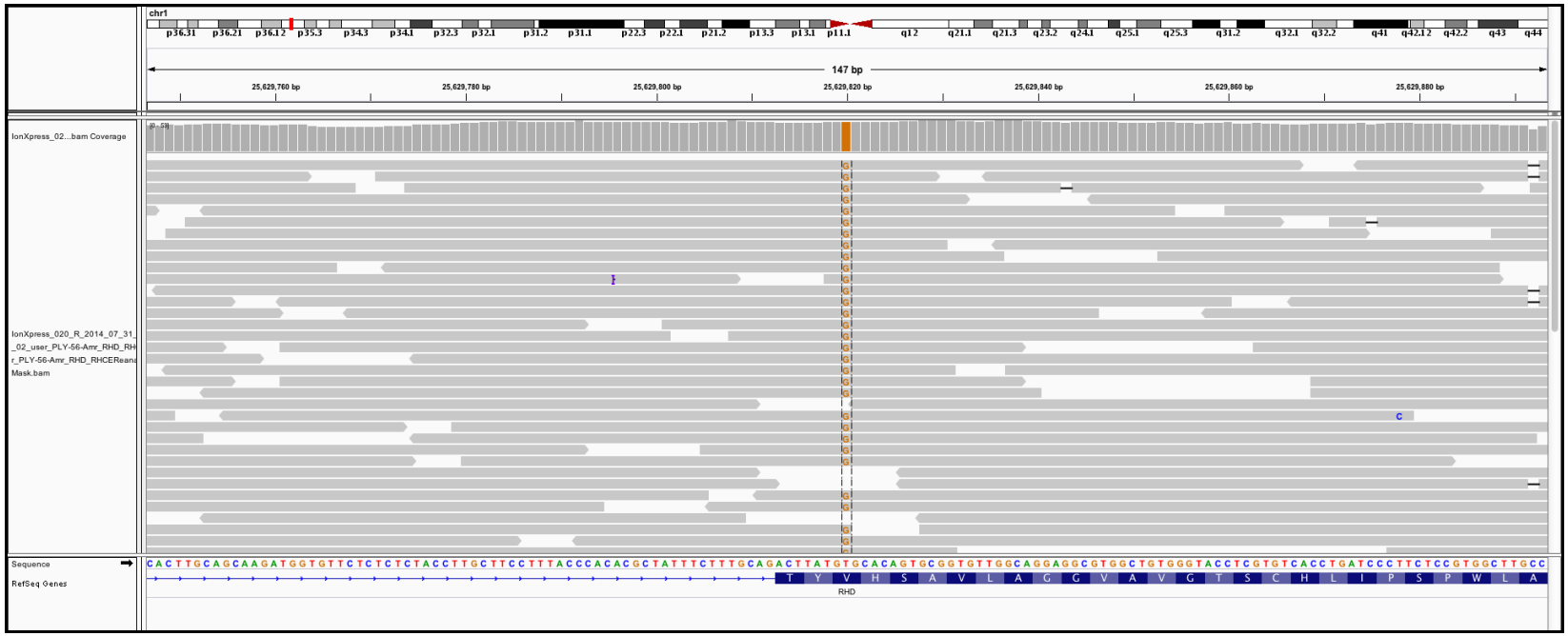


Figure 6.19 A hemizygous SNP in exon 9 of the *RHD* gene indicating a weak D type 2 sample.

The sample R_2r shows a SNP in exon 9 1154G>C (Gly385Ala) of the *RHD* gene. Output from IGV software demonstrates the visualisation of the sequencing data. Chromosomal locations are shown on the top and the reference gene in blue at the bottom.

Table 6.2 The genotyping results of five weak D samples using LR-PCR amplification followed by sequencing on the Ion PGM™.

Samples	Quantity	Exon	Nucleotides	Amino acid	Prediction from NGS Genotyping
6	1	4 5 6	602C>G 667T>G 819G>A	Thr201Arg Phe223Val Silent	DAR3.1 weak partial D 4.0 *(RHD*DAR3.01*)
9 and 10	2	6	809T>G	Val270Gly	Weak D type 1 (*RHD*01W.01)*
7 and 8	2	9	1154G>C	Gly385Ala	Weak D type 2 *(RHD*01W.02)*

Table 6.3 Intronic SNPs found in the *RHD* gene by NGS.

The serology of the samples is listed in Table 6.1. The depth of coverage used for homozygous SNP was 15, while 33 for heterozygous SNPs, respectively (if applicable). The areas with low depth of coverage could not be analysed. The highlighted SNPs with green shows are R_1R_2 phenotypes sharing some of the SNPs of the R_2 allele. Zygosity as defined by Ms. Kelly Sillence. Note all R_1R_2 samples defined by serology are in fact hemizygous for *RHD* gene.

#	Intronic SNPs	Zygosity	Intron	Phenotype	Sample number
1	25,611,035 G>C	Homozygous	1	R_1R_1, R_1r, R_0r, R_1R_2	1, 2, 3, 6, 9, 10
2	25,611,580 G>A	Homozygous	2	R_2R_2, R_2r, R_1R_2	4, 5, 7, 8
3	25,611,635 T>C	Homozygous	2	R_1r	10
4	25,614,235 C>T	Homozygous in R_1r Heterozygous & Homozygous in R_1R_1	2	R_1R_1, R_1r	2, 3, 6
5	25,614,400 C>G	Homozygous	2	R_2R_2, R_2r, R_1R_2	4, 5, 7, 8
6	25,618,869 T>G	Homozygous	3	R_0r	6
7	25,619,444 G>A	Homozygous	3	R_1R_1, R_1r, R_0r, R_1R_2	1, 2, 3, 6, 9, 10
8	25,621,489 T>C	Homozygous	3	R_0r	6
9	25,621,894 G>A	Homozygous	3	R_0r	6
10	25,621,980 C>T	Homozygous	3	R_2R_2, R_2r, R_1R_2	4, 5, 7, 8
11	25,622,244 A>G	Homozygous	3	R_2R_2	4
12	25,622,288 T>A	Homozygous	3	R_1R_1, R_1r, R_0r	1, 2, 3, 6, 9, 10
13	25,622,291 G>A	Homozygous	3	R_1R_1, R_1r, R_0r, R_1R_2	1, 2, 3, 6, 9, 10
14	25,622,341 A>G	Homozygous	3	R_1R_1, R_0r, R_2r	1, 2, 3, 6, 8, 9
15	25,623,105 C>A	Heterozygous	3	R_1R_1	3
16	25,623,118 T>G	Homozygous	3	R_0r	6
17	25,623,620 A>G	Homozygous	3	R_0r	6
18	25,623,622 C>T	Homozygous	3	R_0r	6
19	25,623,750 T>A	Homozygous	3	R_0r	6
20	25,623,967 A>G	Homozygous	3	R_1R_1, R_1r, R_0r, R_1R_2	1, 2, 3, 9, 10
21	25,624,252 G>C	Homozygous	3	R_0r	6
22	25,625,318 G>A	Homozygous	3	R_0r	6
23	25,625,471 T>C	Homozygous	3	R_2R_2, R_2r, R_1R_2	4, 5, 7, 8
24	25,626,327 G>A	Homozygous	3	R_0r	6
25	25,627,023 A>G	Homozygous	3	R_0r	6
26	25,627,066 C>G	Homozygous	3	R_2R_2, R_2r, R_1R_2	4, 5, 7, 8
27	25,627,121 T>G	Homozygous	3	R_0r	6
28	25,627,323 A>G	Heterozygous	3	R_2R_2	5
29	25,628,396 T>G	Homozygous Heterozygous in R_2R_2	5	R_1R_1, R_2R_2, R_1r, R_0r, R_1R_2,	1, 2, 3, 4, 6, 9, 10
30	25,628,572 G>A	Homozygous	5	R_0r	6
31	25,628,814 A>C	Homozygous	5	R_0r	6
32	25,629,797 C>T	Homozygous	5	R_0r	6
33	25,630,950 A>G	Homozygous	5	R_0r	6
34	25,631,188 A>C	Homozygous	5	R_0r	6
35	25,631,436 A>T	Homozygous	5	R_1R_1, R_1r, R_0r, R_1R_2,	1, 2, 3, 9, 10
36	25,631,534 C>T	Homozygous	5	R_0r	6
37	25,631,730 T>G	Homozygous	5	R_0r	6
38	25,631,983 G>A	Homozygous	5	R_0r	6
39	25,632,068 A>G	Homozygous	5	R_0r	6
40	25,632,389 A>G	Homozygous	5	R_2R_2, R_2r, R_0r	4, 5, 6, 7, 8
41	25,632,639 T>G	Homozygous	5	R_0r	6
42	25,632,654 C>T	Homozygous	5	R_2R_2	4
43	25,632,738 A>G	Homozygous	5	R_0r	6
44	25,633,531 G>C	Homozygous	7	R1R1, R1r, R_0r, R_1R_2	1, 2, 3, 6, 9, 10
45	25,633,982 A>G	Homozygous	7	R_0r	6
46	25,634,205 G>A	Homozygous	7	R_2R_2, R_1r R_2r, R_1R_2	4, 5, 6, 10
47	25,635,134 A>C	Heterozygous	7	R_1R_1	3
48	25,635,336 G>C	Homozygous	7	R_2R_2, R_0r	4, 5, 6, 7, 8
49	25,637,536 G>A	Homozygous	7	R_0r	6
50	25,638,011 G>A	Homozygous	7	R_1R_1, R_1r, R_0r, R_1R_2	1, 2, 3, 9, 10
51	25,638920 G>A	Homozygous	7	R_0r	1
52	25,642,132 C>T	Homozygous	7	R_0r	6
53	25,644,654 A>G	Heterozygous in R_1R_1 Homozygous in R_1R_2	8	R_1R_1, R_1R_2	3, 9
54	25,645,175 C>T	Heterozygous	8	R_1R_1	3
55	25,654,529 C>T	Homozygous	8	R_0r	6
56	25,646,258 C>A	Homozygous	8	R_0r	6
57	25,646,634 C>A	Homozygous	8	R_0r	6
58	25,646,748 A>C	Homozygous	8	R_1R_1, R_1r, R_0r, R_1R_2	1, 2, 3, 9, 10
59	25,646,763 A>G	Homozygous	8	R_0r	6
60	25,646,933 T>G	Homozygous	8	R_2R_2, R_2r, R_1R_2	4, 5, 7, 8
61	25,648,349 T>C	Homozygous	8	R_2R_2, R_2r, R_1R_2	4, 5, 7, 8
62	25,648,782 C>G	Homozygous	9	R_0r	6
63	25,648,885 T>C	Homozygous	9	R_0r	6
64	25,651,301 G>A	Homozygous	9	R_1R_1, R_1r, R_1R_2	3, 9, 10

Table 6.4 The genotyping results of the *RHCE* alleles.

Sample	Serology	Quantity	Exon	Nucleotides	Amino acid	NGS Genotyping
1	R_0r	1	5	676C>G Homozygous	Pro226Ala	*RHCE*e*
2 and 3	R_1R_1	2	2	307C>T Homozygous	Pro103Ser	*RHCE*C*
4 and 5	R_2R_2	2	5	676G>C Homozygous	Ala226Pro	*RHCE*E*
6	R_0r	1	1 2 5 5 5 7	48G>C Heterozygous 307T>C Homozygous 733C>G Heterozygous 744T>C Heterozygous 676C>G Homozygous 1025C>T Heterozygous	Trp16Cys Ser103Pro Leu245Val Silent Pro226Ala Thr342Ile	*RHCE*ceVS.04* *RHCE*c* *RHCE*e*
7	R_1R_2	1	2 5	307C>T Heterozygous 676G>C Heterozygous	Pro103Ser Ala226Pro	*RHCE*C/RHCE*c* *RHCE*E/RHCE*e*
8	R_2r	1	5	676G>C Heterozygous	Ala226Pro	*RHCE*E/RHCE*e*
9	R_1R_2	1	1 2 2 2 5	48G>C 177C>A§ 202A>G§ 307C>T§ All Heterozygous 676G>C Heterozygous	Trp16Cys Leu60Ile Asn68Ser Pro103Ser Ala226Pro	*RHCE*C/RHCE*c* *RHCE*E/RHCE*e*
10	R_1r	1	1 2 2 2 2 2	48G>C§ 177C>A§ 202A>G§ 307C>T§ 200A>G All Heterozygous 208C>T Heterozygous	Trp16Cys Leu60Ile Asn68Ser Pro103Ser Silent Arg70Trp	*RHCE*C* *Novel*

§= These SNPs are corresponding to the *RHCE*C* allele are not observed in sample 7. These were matching the reference gene in which G at 48, C at 177, A at 202 and C at 307.

Table 6.5: Intronic SNPs found in the *RHCE* gene by NGS.

#	Intronic SNPs	Zygosity	Intron	Phenotype	Sample number
1	25,728,930C>A	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_2R_2, R_2r, R_0r	3	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
2	25,728,612 Gdel	Heterozygous	3	R_1R_1, R_2R_2, R_2r, R_1R_2, R_1r	2, 3, 8, 9, 10
3	25,728,396 A>G	Heterozygous	3	R_1r, R_1R_2	9, 10
4	25,727,9(	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_2R_2, R_2r, R_0r, R_1R_2	3	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
5	25,727,486 A>G	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous R_2R_2, R_2r, R_0r, R_1R_2,	3	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
6	25,727,438G>A	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_2R_2, R_2r, R_0r, R_1R_2	3	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
7	25,727,386G>A	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_2R_2 R_2r, R_0r, R_1R_2	3	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
8	25,726,916 Cdel	Heterozygous	3	R_1R_1, R_2R_2, R_1r, R_0r, R_1R_2	2, 3, 4, 6, 7, 8, 9, 10
9	25,626,836 G>A	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_2R_2, R_2r	3	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
10	25,724,783A>G	Homozygous in R_2R_2, R_2r, R_1R_2 Heterozygous in R_1r, R_0r, R_1R_2	3	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	4, 5, 6, 7, 8, 9, 10
11	25,724,524C>A	Heterozygous	3	R_0r	6
12	25,724,501T>C	Heterozygous	3	R_1R_1, R_1r, R_1R_2	3, 9, 10
13	25,724,366C>T	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_2R_2, R_2r, R_1R_2	3	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
14	25,724,306G>T	Heterozygous	3	R_0r	8
15	25,724,234C>T	Heterozygous	3	R_1R_1	3
16	25,724,146A>G	Heterozygous in R_1r, R_0r, R_1R_2, Homozygous in R_2R_2, R_2r, R_0r	3	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
17	25,724,005G>A	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_2R_2, R_2r, R_0r	3	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
18	25,722,334A>G	Heterozygous	3	R_1R_1	2
19	25,722,206A>G	Heterozygous in R_0r, R_1r, R_1R_2 Homozygous in R_2R_2, R_2r, R_0r	3	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
20	25,722,090T>C	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_2R_2, R_2r, R_0r	3	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
21	25,721,457C>G	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_2R_2, R_0r, R_2r	3	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
22	25,721,154T>C	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_2R_2, R_2r	3	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
23	25,720,404G>A	Heterozygous	3	R_0r	6
24	25,720,370T>C	Homozygous in R_2R_2, R_2r, R_0r, R_1R_2 Heterozygous in R_1R_2, R_1r	3	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	4, 5, 6, 7, 8, 9, 10
25	25,720,301G>A	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_2R_2, R_2r	3	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
26	25,720,248C>G	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_2R_2, R_2r	3	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
27	25,720,204C>A	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_2R_2, R_2r	3	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
28	25,720,101C>G	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_1R_1, R_2R_2, R_2r	3	R_1R_1, R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 2, 4, 5, 6, 7, 8, 9, 10
29	25,720,088C>T	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_1R_1, R_2R_2, R_2r	3	R_1R_1, R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 2, 4, 5, 6, 7, 8, 9, 10
30	25,720,045C>G	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_2R_2, R_2r	3	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
31	25,719,677A>G	Heterozygous	3	R_0r	6
32	25,719,599A>T	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_1R_1, R_2R_2, R_2r	3	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
33	25,719,677A>G	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_2R_2, R_2r	3	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
34	25,719,235T>G	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_1R_1, R_2R_2, R_2r	3	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
35	25,717,841C>T	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_1R_1, R_2R_2, R_2r	4	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
36	25,717,139G>A	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_1R_1, R_2R_2, R_2r	5	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
37	25,715,885G>T	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_2R_2, R_2r	5	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10
38	25,714,249A>C	Heterozygous	6	R_1R_2	9
39	25,713,234T>A	Heterozygous in R_1r, R_0r, R_1R_2 Homozygous in R_2R_2, R_2r	6	R_2R_2, R_1r, R_2r, R_0r, R_1R_2	1, 4, 5, 6, 7, 8, 9, 10

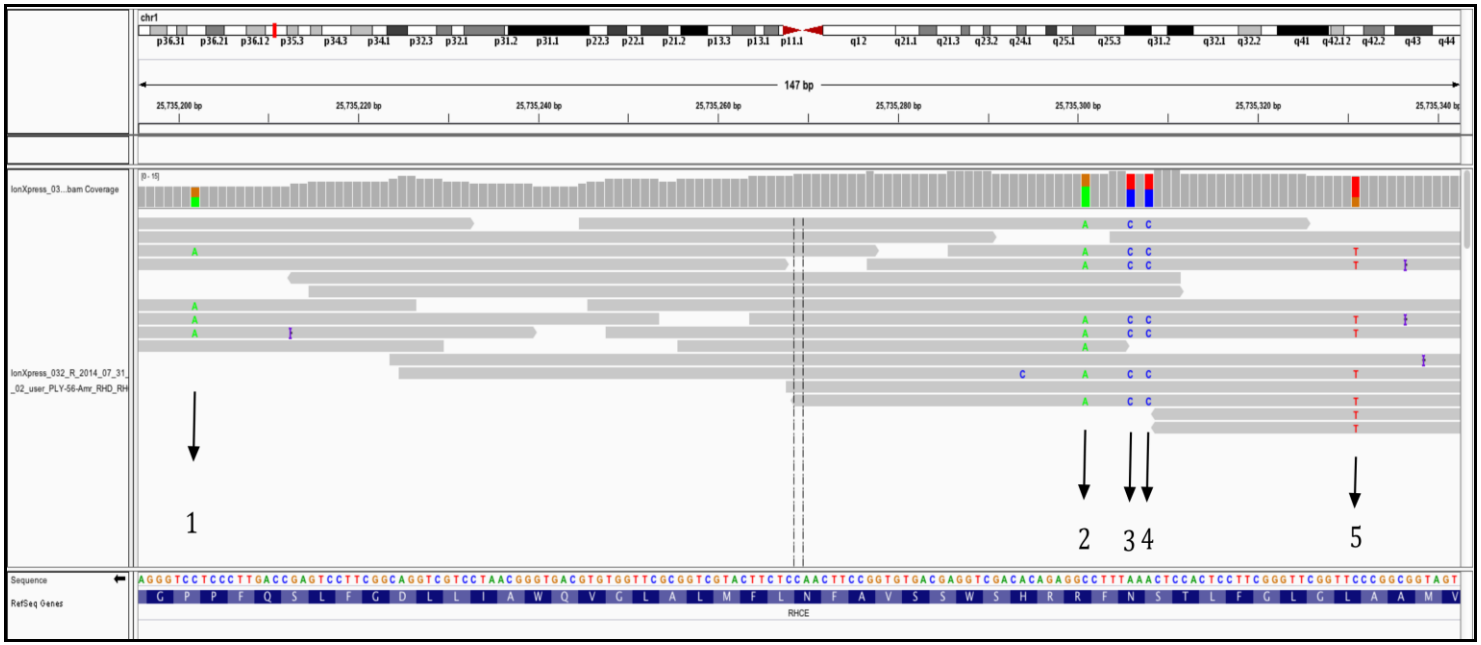


Figure 6.20 Five SNPs in the R_1r sample.

All five SNPs are located in exon 2 of the *RHCE* gene and listed below in order as annotated in the figure:
1. 307C>T (Pro103Ser).
2. The novel SNP is 208C>T (Arg70Trp).
3. 202A>G (Asn68Ser).
4. Unreported silent mutation found 200A>G (Serine at amino acid 67).
5. 177C>A (Leu60Ile).

SNP numbers 1, 3 and 5 are associated with the *RHCE*C* allele. Output from IGV software demonstrates the visualisation of the sequencing data. Chromosomal locations are shown on the top and the reference gene in blue at the bottom.

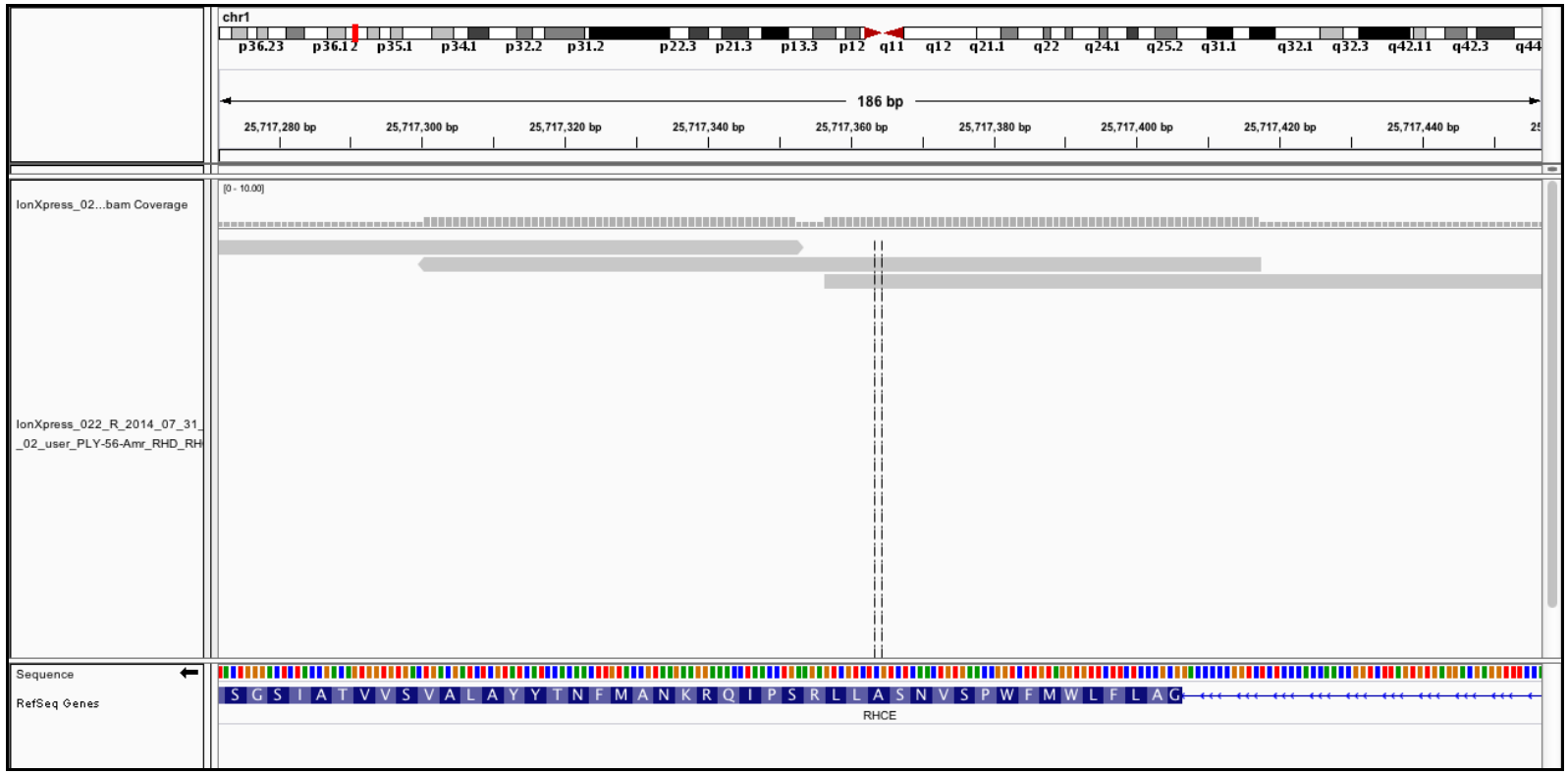


Figure 6.21 Low depth of coverage for *RHCE*e* allele.

The sample serotyped with R_0r phenotype. The photo only shows two sequencing reads in amino acid position 226 to genotype in exon 5 of the *RHCE* gene. The depth of coverage was insufficient for genotyping. Output from IGV software demonstrates the visualisation of the sequencing data. Chromosomal locations are shown on the top and the reference gene in blue at the bottom.

6.4 Discussion

6.4.1 Primer design

In this project, both genes of the Rh blood group system were amplified by LR-PCR and were fully sequenced based on NGS using the Ion PGM™ platform. The first step in designing a specific primer was to find the differences in sequencing nucleotides between the *RHD* and *RHCE* genes.

The primer design then took place to obtain products for both genes with consideration of the specificity according to the human reference genome (hg19). The primer pairs were substituted numerous times and in various combinations and different primer pairs were performed in order to obtain the required and specific amplicons (data not shown).

There were no co-amplification issues observed in the *RHD* and *RHCE* data regarding the specificity of the primer sequences of the LR-PCR approach in comparison with the HEA and HPA Panel [see Chapter 4]. Fichou and co-workers (2014) observed the same issue, in which those two genes were co-amplified using Ion Ampliseq™ Custom Panel. The primers of the LR-PCR were designed bearing in mind the homologous genes of the Rh blood group system. Interestingly, the *RHCE*C* allele successfully and its related SNPs on exon 2 was genotyped by the LR-PCR method. Moreover, the novel allele that was discovered by the HEA and HPA Panel and misaligned to the *RHD* gene was genotyped and perfectly aligned to the *RHCE* gene using the LR-PCR approach [Figure 6.20]. Furthermore, the three SNPs of the *RHCE*C* allele were identified in that sample containing the novel allele [Table 6.4]. These SNPs had not been identified by the HEA and HPA Panel. It can be summarised that co-amplification of exon 2 of *RHD* and *RHCE* took place by the HEA and HPA Panel. Obviously, this is due to the non-specific primers that were provided for that panel which did not take into consideration the homologous genes of the Rh blood group system.

Stabentheiner and co-workers (2011) successfully achieved sequencing of the 10 exons of the *RHD* by taking into consideration the designing of specific primers. Here, in this study the LR-PCR approach was used, not only for the 10 exons of the *RHD*, but also for both entire genes including exons and introns. Therefore, this method may provide high resolution and increase the depth of knowledge regarding the Rh blood group genes.

6.4.2 LR-PCR Polymerase

Many versions of long-range polymerase kits were used, starting by using the LongAmp Hot Start Taq 2X Master Mix and then changed using the PrimeSTAR GXL Polymerase with the requirement of little optimisation. Jia and co-workers (2014) found that the PrimeSTAR GXL Polymerase successfully achieved the best amplification of the *BRCA1* and *BRCA2* genes among a total of different six polymerases used without alteration of the PCR conditions, although the difference in melting temperatures between the primer sequences (Jia et al., 2014). The same results were obtained in the LR-PCR experiment in which all the amplicons of the *RHD* and *RHCE* genes were amplified using the same condition although there were differences between the T_m of the primers. Finally, the last PCR products obtained, three products for the *RHD* gene and four for the *RHCE* gene **[Figure 6.2]**, were sequenced due to the time factor although they were not as good PCR bands as for sequencing.

Another extraction kit Gentra® Puregene® Blood kit was used to purify the genomic DNA [section 2.2.2]. This kit was designed to obtain extremely high molecular weight templates from the human genomic DNA, which approximately 100-200 Kb. Unfortunately, the same outcomes were obtained regarding the PCR products. This is suggested that the design primer sequences with smaller range may overcome the sequencing issues related to the depth of coverage. Moreover, it is highly recommended that to design the new primer sequences according to the new release of human genome

assembly (GRCh38/hg38). This new release could not be used because it has been allowed for the public in December 2013, which following the finishing the design of the current primer sequences in this study (Rosenbloom et al., 2015).

6.4.3 The depth of coverage

According to Bentley et al. (2008), they stated that a depth of coverage of 15× and 33× could be achievable for the homozygous and heterozygous SNPs, respectively. The sequencing on the Ion PGM™ platform was achieved but with low depth of coverage. The depth of coverage for the *RHD* gene ranged from around 16 to 35 sequencing reads. Although these numbers were sufficient for the hemizygous SNPs represented in the weak D samples (Bentley et al., 2008), the first amplicon of the *RHD* gene expressed the lowest level of depth of coverage, which ranged from one to seven sequencing reads and some missed regions were not covered [Figure 6.14]. All the *RHD* amplicons were around 22 Kb.

Regarding the *RHCE* gene, the depth of coverage of the first amplicon was very low which was around 8-10 sequencing reads and showed missed regions [Figure 6.15]. The size of the first product was 19,426 bp and covered the regions of exons 1-3. The second amplicon was 11,215 bp, covered exons 3 and 4 and had an extremely high depth of coverage. The depth of coverage for that region was around 130-191 sequencing reads and no missed regions were observed [Figure 6.16].

The depth of coverage was then decreased again in the areas covered by exons 5-7 to around 4-21 sequencing reads. The third amplicon was designed to cover the region of exon 5 to exon 7, which was 17,856 bp. The fourth amplicon covered from exon 7 to exon 10 expressed an extremely low depth of coverage, which was about four to eight sequencing reads [Figure 6.17].

Obviously, the high depth of coverage was achieved with the smallest product of the LR-PCR as can be seen in the second product of the *RHCE* gene. That product of

11,215 bp obtained more than a sufficient depth of coverage and is adequate for the genotyping analysis. Moreover, the two products of the *KEL* gene were around 12 Kb and achieved a high depth of coverage [see Chapter 5]. Consequently, smaller products for both the *RHD* and *RHCE* genes are required with a re-design of the primers' sequences. It is highly recommended that the primer sequences of the *RHD* and *RHCE* genes should be redesigned in order to obtain shorter amplicons ranging between 10-12 Kb. This is in order to accomplish a proper amplification and to obtain a high depth of coverage.

The method used to quantify the sequencing library was based on Agilent® 2100 Bioanalyzer. Fichou et al. (2014) found that using of Ion Library Equalizer™ Kit observed 10-folds of higher depth of coverage in comparison to Ion Library Quantitation Kit (Fichou et al., 2014). The Ion Library Equalizer™ Kit is based on library amplification using unique primers, capturing the sequencing library using special beads and heat elution to normalise the library to 100 pM to prevent dilution of the library. This might be used in the future if the low depth of coverage obtained although this kit can be prone to polymerase error because the amplification procedure.

6.4.4 Genotyping results of the Rh blood group system

6.4.4.1 Genotyping results of the RHD gene

Ten RhD-positive samples were genotyped by NGS. Out of the five normal RhD-positive samples, no mutations were observed in the entire 10 exons of the *RHD* gene. This confirmed the predefined phenotypes by conventional serology. All the weak D samples observed homozygous SNPs (i.e. hemizygous) and there were two samples of weak D type 1 (*RHD*01W.01)*, two samples of weak D type 2 *(RHD*01W.02)* and the fifth sample was DAR3.1 weak partial D 4.0 *(RHD*DAR3.01*) [Table 6.2].

6.4.4.2 Genotyping results of the RHCE gene

With respect to the *RHCE* gene, the alleles of *RHCE*C, RHCE*c, RHCE*E* and *RHCE*e* were genotyped easily by NGS in both forms of zygosity, i.e. heterozygous or homozygous. The reference gene used in the human genome assembly (hg19) is *RHCE*ce* allele. The alleles of *RHCE*c* and *RHCE*e* matched the reference gene while the variants of the other alleles, *RHCE*C* and *RHCE*E*, were called. The genotyping of the *RHCE* alleles regarding to the C, c, E, and e antigens were achieved.

The allele *RHCE*ceVS.04,* was genotyped successfully by the LR-PCR in the R_0r sample. This allele observed a silent mutation which it did not previously report and it might be considered as a new allele of *RHCE*ceVS*. The individual was mentioned in the serology sheet as having an 'other White' background. The genotyping of the novel *RHCE* allele 208C>T (Arg70Trp), identified by HEA and HPA Panel, was achieved despite the issue of low depth of coverage. This novel allele, which was misaligned to the *RHD* gene instead of the *RHCE* gene by the HEA and HPA Panel, was properly genotyped in the *RHCE* gene by the LR-PCR approach.

6.4.4.3 Genotyping results of the intronic SNPs

There were numerous intronic SNPs observed in the *RHD* and *RHCE* genes whether they were hemizygous in the case of the weak D, homozygous, or heterozygous. However, because of the low depth of coverage, Only 64 and 39 intronic SNPs could be identified in the *RHD* and *RHCE*, respectively.

In this experiment, because a low depth of coverage was observed, most of the SNPs detected regarding the *RHD* gene were homozygous or hemizygous [see Table 6.3]. The depth of coverage for the *RHCE* gene was high especially in intron 3 and 4. Therefore, the heterozygous SNPs in the *RHCE* gene were detected easily [see Table 6.5].

The locations of the intronic SNPs need to be avoided when designing new primers. These reported SNPs here should be taken into consideration when the new primer pairs

would be designed. This is in order to prevent the allele dropout during the annealing of the primers pairs to the target of interest due to the mismatching nucleotide of that SNP. Consequently, no amplification will be achieved as some of experiments could not achieve any amplification (data not shown). Mullins and co-workers (2007) stated that 12 primer pairs were changed in order to prevent allele dropout when amplify the cadherin 1 (*CDH1*) gene. They recommend assessing the primer-binding site whether having any intronic SNPs that have been reported in the databases (Mullins et al., 2007). The databases were used by Mullins and co-workers (2007) include University of California Santa Cruz genomic bioinformatics browser (UCSC Genome Bioinformatics, 2015), the Ensembl Genome browser (Ensembl Genome Browser, 2015) and the Single Nucleotide Polymorphism Database [dbSNP] (Single Nucleotide Polymorphisms Database, 2014). Therefore, the assessment of the reported intronic SNPs needs to be taken into account when the new sets of primers are designed for the *RHD* and *RHCE* genes.

6.4.5 Zygosity results

Samples 7 and 9 were typed by serology as weak D but presumed serotype was given as R_1R_2. Regarding weak D samples, these types of samples are always hemizygous in which one copy of the *RHD* gene observed in such samples. Zygosity testing, which was performed by Ms. Kelly Sillence, confirmed that it was only a single copy of the *RHD* gene was found in those two samples (unpublished data). Accordingly, this is suggested that, along with the results of the zygosity test, the phenotypes of samples 7 and 9 may be one of the following phenotypes (R_1r'', R_2r', R_zr or R_0r^y).

Regarding sample 7, it carries the intronic SNPs that are associated with R_2R_2 and R_2r phenotypes. These intronic SNPs include 2, 5, 10, 23, 26, 60 and 61 [see the highlighted SNPs in Table 6.3]. R_0r^y is extremely rare in English population approximately <0.01%, therefore sample 7 is extremely unlikely to be this phenotype (Daniels, 2013a). The

percentage of these phenotypes R_1r'', R_2r', R_zr in English population is as follows; 1%, 0.28% and 0.19%. Therefore, sample 7 is possibly R_2r' phenotype.

Sample 9 and 10 (R_1r) observed four SNPs in exon 2 of the *RHCE* gene, which are corresponding to *RHCE*C* allele. These SNPs were 48G>C (Trp16Cys), 177C>A (Leu60Ile), 202A>G (Asn68Ser) and 307C>T (Pro103Ser). This data suggested that sample 9 is probably R_1r'' phenotype.

6.4.6 The cost of the NGS for the Rh blood group system

The cost of the NGS library preparation for one sample is around £100.79 for both the *RHD* and *RHCE* genes. The number of samples can be scaled depending on the required load of the sequencing. For example, the Ion 314™ chip can be used for up to three samples with a 100× depth of coverage. The number can be raised to match the requirement of the laboratory to 35 samples and 70 samples for Ion 316™ chip and Ion 318™ chip, respectively. Table 6.5 lists the different types of chips, their capacities, the number of samples that can be obtained and the cost per run.

Table 6.5 The number of samples of the *RHD* and *RHCE* genes that can be obtained with a depth of coverage of 100× in addition to the cost of the sequencing run.

Chip type	Capacity	Number of samples	Cost per run
Ion 314™	30-50 Mb	3	£271.55
Ion 316™	300-500 Mb	35	£444.80
Ion 318™	600 Mb – 1 Gb	70	£619.67

The price per sequencing run on the Ion PGM™ is given.

In conclusion, the LR-PCR approach for the Rh blood group system contributes a high-resolution genotyping based on NGS. The different alleles especially the weak and rare variants of both the *RHD* and *RHCE* genes can be easily identified. This approach provides a crucial method to identify new alleles as well as to facilitate the investigation of the hybrid alleles. This can be performed better if the issues of the small products are designed to obtain as high a depth of coverage as possible.

Chapter 7 : General Discussion and Conclusion

Alloimmunisation occurs following transfusion of mismatched serotyped units to the patients. These transfusion reactions have been increased in patients requiring multiple units, such as those with SCD (Chou and Westhoff, 2011). Accordingly, the demand of the serological reagents has become higher in cost in some of special cases [see section 7.6]. Moreover, conventional serology cannot identify the phenotype in recently transfused patients. Using BGG can preclude these issues. Currently, many techniques and platforms of BGG are available. However, none of these platforms, including the microarray platforms, are able to define the new alleles. The array-based platforms have issues including the requirement of updating for new alleles as well as inability to investigate the weakening and silencing alleles. Furthermore, an issue of non-valid samples was reported [Chapter 3]. This is may be due to allelic droput in which the amplification did not take place or due to a presence of novel allele in the primer-binding sites. In addition, it could be the hybridisation of the microarray beads to the sample did not occur.

Interestingly, NGS offers the high-throughput upon these platforms as well as in identification of the new variants, whilst simultaneously describing all known alleles (Avent et al., 2015). In this PhD project, two approaches were used to assess whether NGS is capable of genotyping the blood groups. This will be discussed in the next sections.

7.1 The HEA and HPA Panel

The first approach was amplicon-based target selection using Ion Ampliseq™ Custom Panel for capturing the exons of the blood group genes and the SNPs encoding the HPAs (HEA and HPA Panel) [see Chapter 4]. The HEA and HPA Panel is a rapid technique in comparison to other sequencing library construction methods in which no

fragmentation, size selection or multiple steps of purification are required. The sequencing library takes up to 5 hours to prepare and the clonal amplification as well as the sequencing can be carried out in the same day. Therefore, the results can be obtained by the next day. It is more comprehensive to genotype the blood groups and HPAs in a single assay. Therefore, there may avoid using multiple kits for various genes for blood groups or a separate kit for HPAs.

Indeed, out of 28 samples four novel alleles were investigated [see Chapter 4]. The first novel allele was investigated in the *RHCE* gene (208C>T Arg70Trp). The location of the amino acid 70 is extracellular to the *RHCE* protein and this is suggested that such a SNP may be weakening the expression of the antigen on the RhCE membrane. This SNP was misaligned to *RHD* gene instead of the *RHCE* gene. This issue was resolved by LR-PCR approach for the Rh blood group genes, which have the specific primer sequences [see Chapter 6]. The misalignment was due to the primer sequences provided by the panel not being specific and did not take the homologous genes into account. Consequently, co-amplification of homologous genes occurred. Fichou et al. (2014) reported this issue when they used Ampliseq™ Custom Panel for BGG. The remaining three alleles were investigated in the *KEL* gene. The first SNP was in exon 4 331G>A (Ala111Thr). The second SNP was in exon 17 1907C>T (Ala636Val) and the third SNP was in exon 19 2165T>C (Leu722Pro). All these amino acid substitutions occur on the external domain of the Kell glycoprotein. Accordingly, this might change the protein conformation and therefore it might affect the antigenicity of the Kell glycoprotein.

Apart from the co-amplification issue, all the advantages mentioned above give the HEA and HPA Panel to be the method of choice in the near future if the issue of co-amplification of homologous genes is resolved. Liu et al (2014) stressed that it is impossible to genotype all the blood group antigens by the previous molecular techniques, such as the microarray, in a single assay. Therefore, this can be feasible

using a broad panel encompassing all of the genes encoding the blood groups (Liu et al., 2014).

Because the Ion Ampliseq™ Custom Panel has the capacity to amplify 6144 amplicons per pool, the panel can be enlarged to encompass many targets. In fact, this gives the advantage to Ion Ampliseq™ Custom Panel over the microarray, which can have the multiplex up to 500 targets. In the United States, it has been recommended that using 17 blood group systems for genotyping as a molecular reference standard by the American Association of Blood Banks (Denomme et al., 2010). The other blood groups to be included in the HEA and HPA Panel are Lutheran, Scianna, Landsteiner-Wiener, Knops, Cromer, Indian, Ok. Regarding the HEA and HPA Panel, more blood group genes can be added for an extremely comprehensive assay. In addition, human leukocyte antigens (HLA) and human neutrophil antigens could be added in order to provide matching units to preclude complications regarding stem cell transplantation and transfusion related acute lung injury.

Moreover, the designed panel can be expanded more to obtain all the human blood group genes. The total human genome contains 3.2×10^9 bp while the exome, the exons exclusive only to the coding areas, comprises 1-2%. The customised panel reduces the undesired DNA sequences. All the blood group genes comprise an approximate of 150 Kb (Tilley and Grimsley, 2014).

7.2 LR-PCR approach

The second approach was the LR-PCR for the *KEL*, *RHD* and *RHCE* genes. LR-PCR followed by deep sequencing on the Ion PGM™ approach provides a higher resolution analysis. Kell blood group system along with the LR-PCR approach was used as a model for the Rh blood group system. The sequencing of the entire *KEL* gene helped to genotype all the alleles and predict the related antigens in particular the high prevalence ones. Moreover, the approach will help to resolve the discrepancies obtained by

serology. An interesting finding was that genotyping of a sample showed it to be heterozygous for the SNP 841C>T (Arg281Trp) in exon 8 indicating the *KEL*02.03* allele encoding Kp^a antigen. This sample was not typed by serology and some cases were reported involvement of the Kp^a antigen causing HDFN (Smoleniec et al., 1994; Costamagna et al., 1997). Some rare Kell samples, such as K_{null} or K_{mod}, can be investigated to assess whether any intronic mutations can be involved in the genetic backgrounds of such cases.

Following the Kell NGS sequencing, the approach of LR-PCR has been applied on *RHD* and *RHCE* genes. The outcomes of the sequencing reactions were satisfactory especially in differentiating between RhD-positive samples and the weak D apart from the missing regions found during the analysis. The Rh LR-PCR approach perfectly distinguished between RhD-positive, weak D and partial D samples. In five samples that serotyped as weak D, two samples were weak D type 1, two samples weak D type 2 and one DAR3.1 weak partial D 4.0. Furthermore, the *RHCE*ceVS.04* allele was observed in the *RHCE* gene in one of the samples. Moreover, this may be able to resolve the complex situations in particular to the hybrid genes of the Rh blood group system. Unquestionably, this gives the advantage to NGS in identification of the unprecedented alleles.

In fact, the LR-PCR for the *RHD* and *RHCE* genes resolved the misalignment, which occurred using the HEA and HPA Panel for the novel allele of the *RHCE* (208C>T Arg70Trp) when it incorrectly mapped to the *RHD* gene. This SNP was reported recently in French Guiana case with weak E antigen phenotype and with further cloning the allele revealed to be *RHCE*cE* (Vrignaud et al., 2014). In contrast, in our study the serotyping of this sample was (C+c+e+) and genotyped as *RHCE*Ce/RHCE*ce*. Further cloning needs to be carried out to determine in which the novel allele belong

either to *RHCE*Ce or RHCE*ce*. This could not be performed due to the time remaining for PhD completion.

7.3 Intronic SNPs

Intronic SNPs may help to distinguish the allele of interest. In other words, it can be stated that each phenotype has a pattern of its own intronic SNPs. For instance, the R_0r phenotypes had many unique intronic SNPs in comparison to the other sequenced phenotypes [Table 6.3]. Moreover, regarding the *RHD* gene the intronic SNPs may help to determine the specific allele with the assistance of zygosity testing [see Chapter 6]. Two samples were investigated by the zygosity test to be having a single copy of the *RHD* gene, while they were typed by serology as R_1R_2. One sample shared intronic SNPs with samples of R_2R_2 and R_2r phenotypes. Consequently, it might be possible for this sample to be R_2r' phenotype. The second sample observed the four SNPs associated with the *RHCE*C* allele in exon 2 of the *RHCE* gene, therefore this suggests that this sample may be R_1r'' phenotype.

Non-amplification of some of the Rh amplicons was observed (data not shown). This may be due to intronic SNPs, which may hamper amplification and cause allelic dropout. Hence, the knowledge regarding the intronic SNPs is crucial to preclude any SNPs located within the primer-binding sites, as many of the intronic SNPs are available in the public databases.

At least seven patterns have been found during the analysis of the intronic SNPs of the Kidd blood group system (Avent et al., 2015). Avent et al. (2015) stated that it might be worthwhile to establish and confirm the genetic background of certain weakening or silencing for Kidd blood group alleles. In fact, this will assess the genetic background for all the blood groups especially the hybrid genes of the Rh. In a thorough study of the *BRCA1* gene for the breast cancer, Ratanaphan et al. (2011) found novel intronic variants related to the Thai families with inherited breast cancer. Out of five breast

cancer patients, a single sample observed an unprecedented intronic mutation, which was a deletion of 14 nucleotides in intron 7 of the *BRCA1* gene (Ratanaphan et al., 2011). This deletion shows the importance of the analysis of the intronic SNPs and how it may help in clinical investigations of patients.

Furthermore, the intronic SNPs are useful to determine the zygosity. For example, the father can be genotyped to the *RHD* introns when the fetus at risk of HDFN. The intronic SNPs will appear as hemizygous or heterozygous depending on how many *RHD* copies the father has.

Currently, the limitations of this technique include the low depth of coverage and some missed regions. Indeed, the smallest product of 11,215 bp among the seven amplicons for the entire Rh blood group system observed the highest depth of coverage. This may be because the product was small in comparison to other amplicons. In addition, the products of the *KEL* gene were approximately 12 Kb and achieved a high depth of coverage. Ozcelik et al. (2012) used amplicons ranging from approximately 5 Kb to 14 Kb to amplify the entire genes of *BRCA1* and *BRCA2* for breast cancer. The same finding of successfully sequencing products for HLA genes using LR-PCR was obtained when PCR amplicon size ranged approximately from 4 Kb to 11 Kb. It should be noted that they were using the same polymerase kit used for the Rh blood group genes and the Ion PGM™ was used for sequencing (Shiina et al., 2012). Consequently, redesigning of new primers to give amplicons not exceeding 14 Kb is highly recommended to obtain a proper coverage in sequencing. By then, numerous samples at least 100 serotyped blood donors can be investigated. The LR-PCR approach will easily sequence the null alleles and hybrid genes and resolve the complicated cases in Rh and Kell blood group systems.

7.4 Small read length and cloning

Due to small reads of the NGS platforms, in the Ion PGM™ platform 200-400 bp, it is unfeasible to determine whether the SNPs of a particular allele are located in *cis* or *trans* arrangement. It might be possible to perform cloning following by the sequencing on the Ion PGM™. In fact, single molecule sequencing offers long read length of approximately 98 Kb using a MiniION platform from Oxford Nanopore technologies (Laver et al., 2015). This number exceeds the size of most of the blood group genes if is not all. It might help to demonstrate if the SNPs are inherited together or separately.

7.5 Data analysis and storage

The genotyping process does not require a deep knowledge of bioinformatics to analyse the data, as many tools are easy to use. The user needs to be aware about the knowledge regarding the different alleles encoding the blood group antigens. Most importantly, special software was required to convert the VCF files and digitalise that to predicted phenotypes in order to facilitate these approaches and to be adopted for the blood bank. Therefore, software is needed to automatically score blood group alleles, but none are as yet commercially available.

Screening of donors and patients at blood bank on a routine basis by the NGS has challenges of keeping record of the sequencing information. This is due to the amount of the generated sequences from NGS platforms are extremely huge, therefore sufficient storage is needed. Cloud storage could resolve the issue of the extreme data generated from the NGS (Shanker, 2012).

7.6 The cost

NGS will replace the current array-based platforms due to the cost factor. The price of the HEA and HPA Panel for genotyping 11 blood group systems and 16 HPAs is approximately £121.22 per sample. Regarding the high-resolution sequencing of the

LR-PCR products of the *KEL* gene costs around £50.12 and regarding the *RHD* and *RHCE* genes £115.54 per sample. On the other hand, the serology has become extremely expensive in particular some cases such as autoimmune haemolytic anaemia can be reached to $1490 [around £1048.20] (Mazonson et al., 2014). In future, WGS or WES may apply for every individual when they are born, thus no further cost will be required for the BGG.

7.7 The future of the NGS technology

As the technology of NGS is rapidly evolving, lots of optimization occurs nowadays to develop and speed up library construction, clonal amplification and the bioinformatics tools. For example, Ion Chef™ system from Thermo Fisher Scientific has been released recently to generate sequencing template and perform loading onto the chip rapidly and automation of the procedure. The future of NGS may be will develop WGS by using the single molecule sequencing for every individual when they are born and can get all the information needed. Therefore, no additional cost will be added to burden the blood bank. Moreover, this will increase the knowledge regarding the genetic basis of the blood groups. Interestingly, the genetic basis of the Vel blood group system was performed by whole exome sequencing of five individuals of Vel negative (Cvejic et al., 2013). This is shown the importance of NGS, which will pave the way to resolve the genetic basis of orphan antigens and those not currently allocated to blood group systems.

7.8 Future work

Future work can be carried out including an extensive customised panel of blood groups and HPAs for BGG and non-invasive foetal genotyping from maternal plasma to prevent HDFN. Optimisation of Sanger sequencing is required to validate the SNPs in particular the novel ones. High resolution genotyping using LR-PCR approach followed

by sequencing can resolve complicated cases in particular the samples with hybrid genes. This approach can be performed for other blood group systems such as MNS, Diego, Colton and Dombrock as some serological reagents are not available such as Dombrock blood group (Reid, 2009). MNS blood group system has hybrid genes because it has three genes with high homology. The amplification of the entire genes of the MNS blood group, and other blood groups, will broaden our knowledge about the variants causes and intronic SNPs.

Regarding the high resolution sequencing for Rh, redesigning of the primers is essential to obtain a satisfactory level of sequencing. Short PCR products should be considered in order to obtain a better depth of coverage of sequencing. This type of sequencing would allow studying the Rh variants in depth.

The knowledge of intronic SNPs regarding those variants can be expanded and illustrate how such samples could behave and may assist to determine the specific haplotype. Furthermore, chromosomal locations of the defined intronic SNPs must be avoided when redesign the Rh primers to preclude any allelic dropout. Interestingly, intronic SNPs may assist to determine the Rh zygosity.

It is highly recommended to establish an integrated system based on laboratory information management system (LIMS) for NGS for the BGG. The system should include tracking for different projects or studies, sample information registry, tracking form for NGS library preparation including sample input quality and barcoded submission form, workflow for NGS data analysis, data reporting system with high accessibility and storage of the sequencing data (Scholtalbers et al., 2013). In addition, comprehensive traceability must be ensured for the accomplished analysis (Bianchi et al., 2016). This may provide a suitable and reliable laboratory automation environment to save time and reduce errors.

In conclusion, two approaches have been developed in this project. The first one is a comprehensive assay for blood groups and HPAs. Such an approach may help the community to investigate both donors and patients. Furthermore, it will definitely assist regarding the identification of novel alleles. The second approach of LR-PCR can be used to resolve complicated cases as well as to broaden our knowledge about the blood groups in particular to the intronic SNPs and primer design. Although of some issues aligned with the both approaches, NGS will likely replace conventional serology in the near future.

References

Agilent Technologies. (2013). *Agilent High Sensitivity DNA Kit Guide* [Online]. Available at: http://www.agilent.com/cs/library/usermanuals/Public/G2938-90321_SensitivityDNA_KG_EN.pdf [Accessed 23 April 2013].

Ajay, S. S., Parker, S. C. J., Ozel Abaan, H., Fuentes Fajardo, K. V. & Margulies, E. H. (2011). Accurate and comprehensive sequencing of personal genomes. *Genome Research*, **21,** 1498-1505.

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data* [Online]. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc [Accessed 6 June 2013].

Anstee, D. J. (2009). Red cell genotyping and the future of pretransfusion testing. *Blood*, **114,** 248-56.

Avent, N. D. (1997). Human erythrocyte antigen expression: its molecular bases. *British Journal of Biomedical Science*, **54,** 16-37.

Avent, N. D. (1998). Antenatal genotyping of the blood groups of the fetus. *Vox Sanguinis*, **74 Suppl 2,** 365-74.

Avent, N. D. (2009). Large-scale blood group genotyping: clinical implications. *British Journal of Haematology*, **144,** 3-13.

Avent, N. D., Madgett, T. E., Halawani, A. J., Altayar, M. A., Kiernan, M., Reynolds, A. J. & Li, X. (2015). Next-generation sequencing: academic overkill or high-resolution routine blood group genotyping? *ISBT Science Series*, **10,** 250-256.

Avent, N. D., Madgett, T. E., Lee, Z. E., Head, D. J., Maddocks, D. G. & Skinner, L. H. (2006). Molecular biology of Rh proteins and relevance to molecular medicine. *Expert Reviews in Molecular Medicine*, **8,** 1-20.

Avent, N. D. & Martin, P. G. (1996). Kell typing by allele-specific PCR (ASP). *British Journal of Haematology*, **93,** 728-730.

Avent, N. D., Martinez, A., Flegel, W. A., Olsson, M. L., Scott, M. L., Nogues, N., Pisacka, M., Daniels, G., Van Der Schoot, E., Muniz-Diaz, E., Madgett, T. E., Storry, J. R., Beiboer, S. H., Maaskant-Van Wijk, P. A., Von Zabern, I., Jimenez, E., Tejedor, D., Lopez, M., Camacho, E., Cheroutre, G., Hacker, A.,

Jinoch, P., Svobodova, I. & De Haas, M. (2007). The BloodGen project: toward mass-scale comprehensive genotyping of blood donors in the European Union and beyond. *Transfusion,* **47,** 40S-6S.

Avent, N. D., Martinez, A., Flegel, W. A., Olsson, M. L., Scott, M. L., Nogues, N., Pisacka, M., Daniels, G. L., Muniz-Diaz, E., Madgett, T. E., Storry, J. R., Beiboer, S., Maaskant-Van Wijk, P. M., Von Zabern, I., Jimenez, E., Tejedor, D., Lopez, M., Camacho, E., Cheroutre, G., Hacker, A., Jinoch, P., Svobodova, I., Van Der Schoot, E. & De Haas, M. (2009). The Bloodgen Project of the European Union, 2003-2009. *Transfusion Medicine & Hemotherapy,* **36,** 162-167.

Avent, N. D. & Reid, M. E. (2000). The Rh blood group system: a review. *Blood,* **95,** 375-87.

Avent, N. D., Ridgwell, K., Tanner, M. J. & Anstee, D. J. (1990). cDNA cloning of a 30 kDa erythrocyte membrane protein associated with Rh (Rhesus)-blood-group-antigen expression. *Biochemical Journal,* **271,** 821-5.

Aygun, B., Padmanabhan, S., Paley, C. & Chandrasekaran, V. (2002). Clinical significance of RBC alloantibodies and autoantibodies in sickle cell patients who received transfusions. *Transfusion,* **42,** 37-43.

Babinszki, A. & Berkowitz, R. L. (1999). Haemolytic disease of the newborn caused by anti-c, anti-E and anti-Fya antibodies: report of five cases. *Prenatal Diagnosis,* **19,** 533-6.

Bedtools. (2015). *bedtools: a powerful toolset for genome arithmetic* [Online]. Available at: http://bedtools.readthedocs.org/ [Accessed 12 December 2014].

Beiboer, S. H., Wieringa-Jelsma, T., Maaskant-Van Wijk, P. A., Van Der Schoot, C. E., Van Zwieten, R., Roos, D., Den Dunnen, J. T. & De Haas, M. (2005). Rapid genotyping of blood group antigens by multiplex polymerase chain reaction and DNA microarray hybridization. *Transfusion,* **45,** 667-79.

Bentley, D. R. (2000). The Human Genome Project—An Overview. *Medicinal Research Reviews,* **20,** 189-196.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush,

M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E. Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature,* **456,** 53-59.

Bianchi, V., Ceol, A., Ogier, A. G. E., De Pretis, S., Galeota, E., Kishore, K., Bora, P., Croci, O., Campaner, S., Amati, B., Morelli, M. J. & Pelizzola, M. (2016). Integrated Systems for NGS Data Management and Analysis: Open Issues and Available Solutions. *Front Genet,* **7**.

Bowman, J. M. (1998). RhD hemolytic disease of the newborn. *New England Journal of Medicine,* **339,** 1775-7.

Camara-Clayette, V., Rahuel, C., Lopez, C., Hattab, C., Verkarre, V., Bertrand, O. & Cartron, J. P. (2001). Transcriptional regulation of the KEL gene and Kell protein expression in erythroid and non-erythroid cells. *Biochemical Journal,* **356,** 171-80.

Cartron, J. P. (1994). Defining the Rh blood group antigens: Biochemistry and molecular genetics. *Blood Reviews,* **8,** 199-212.

Cartron, J. P., Bailly, P., Le Van Kim, C., Cherif-Zahar, B., Matassi, G., Bertrand, O. & Colin, Y. (1998). Insights into the structure and function of membrane polypeptides carrying blood group antigens. *Vox Sanguinis,* **74 Suppl 2,** 29-64.

Chang, J. G., Wang, J. C., Yang, T. Y., Tsan, K. W., Shih, M. C., Peng, C. T. & Tsai, C. H. (1998). Human RhDel is caused by a deletion of 1,013 bp between introns 8 and 9 including exon 9 of RHD gene. *Blood,* **92,** 2602-4.

Chen, Y. X., Peng, J., Novaretti, M., Reid, M. E. & Huang, C. H. (2004). Deletion of arginine codon 229 in the Rhce gene alters e and f but not c antigen expression. *Transfusion,* **44,** 391-8.

Cherif-Zahar, B., Bloy, C., Le Van Kim, C., Blanchard, D., Bailly, P., Hermand, P., Salmon, C., Cartron, J. P. & Colin, Y. (1990). Molecular cloning and protein structure of a human blood group Rh polypeptide. *Proceedings of the National Academy of Sciences USA,* **87,** 6243-7.

Cherif-Zahar, B., Matassi, G., Raynal, V., Gane, P., Mempel, W., Perez, C. & Cartron, J. P. (1998). Molecular defects of the RHCE gene in Rh-deficient individuals of the amorph type. *Blood,* **92,** 639-46.

Chou, S. T. & Westhoff, C. M. (2010). The Rh and RhAG blood group systems. *Immunohematology,* **26,** 178-86.

Chou, S. T. & Westhoff, C. M. (2011). The role of molecular immunohematology in sickle cell disease. *Transfusion & Apheresis Science,* **44,** 73-9.

Colin, Y., Cherif-Zahar, B., Le Van Kim, C., Raynal, V., Van Huffel, V. & Cartron, J. P. (1991). Genetic basis of the RhD-positive and RhD-negative blood group polymorphism as determined by Southern analysis. *Blood,* **78,** 2747-52.

Conroy, M. J., Bullough, P. A., Merrick, M. & Avent, N. D. (2005). Modelling the human rhesus proteins: implications for structure and function. *British Journal of Haematology,* **131,** 543-51.

Coombs, R. R., Mourant, A. E. & Race, R. R. (1946). In-vivo isosensitisation of red cells in babies with haemolytic disease. *Lancet,* **1,** 264-6.

Coombs, R. R. & Roberts, F. (1959). The antiglobulin reaction. *British Medical Bulletin,* **15,** 113-8.

Costamagna, L., Barbarini, M., Viarengo, G. L., Pagani, A., Isernia, D. & Salvaneschi, L. (1997). A case of hemolytic disease of the newborn due to anti-Kpa. *Immunohematology,* **13,** 61-2.

Curtis, B. R. (2008). Genotyping for human platelet alloantigen polymorphisms: applications in the diagnosis of alloimmune platelet disorders. *Seminars in Thrombosis & Hemostasis* **34,** 539-48.

Curtis, B. R. & Mcfarland, J. G. (2014). Human platelet antigens - 2013. *Vox Sanguinis,* **106,** 93-102.

Cvejic, A., Haer-Wigman, L., Stephens, J. C., Kostadima, M., Smethurst, P. A., Frontini, M., Van Den Akker, E., Bertone, P., Bielczyk-Maczynska, E., Farrow, S., Fehrmann, R. S., Gray, A., De Haas, M., Haver, V. G., Jordan, G., Karjalainen, J., Kerstens, H. H., Kiddle, G., Lloyd-Jones, H., Needs, M., Poole, J., Soussan, A. A., Rendon, A., Rieneck, K., Sambrook, J. G., Schepers, H., Sillje, H. H., Sipos, B., Swinkels, D., Tamuri, A. U., Verweij, N., Watkins, N. A., Westra, H. J., Stemple, D., Franke, L., Soranzo, N., Stunnenberg, H. G., Goldman, N., Van Der Harst, P., Van Der Schoot, C. E., Ouwehand, W. H. & Albers, C. A. (2013). SMIM1 underlies the Vel blood group and influences red blood cell traits. *Nature Genetics,* **45,** 542-5.

Danek, A., Rubio, J. P., Rampoldi, L., Ho, M., Dobson-Stone, C., Tison, F., Symmans, W. A., Oechsner, M., Kalckreuth, W., Watt, J. M., Corbett, A. J., Hamdalla, H. H., Marshall, A. G., Sutton, I., Dotti, M. T., Malandrini, A., Walker, R. H., Daniels, G. & Monaco, A. P. (2001). McLeod neuroacanthocytosis: genotype and phenotype. *Annals of Neurology,* **50,** 755-64.

Daniels, G. (2005). The molecular genetics of blood group polymorphism. *Transplant Immunology,* **14,** 143-53.

Daniels, G. (2009). The molecular genetics of blood group polymorphism. *Human Genetics,* **126,** 729-742.

Daniels, G. (2013a). *Human Blood Groups, 3rd ed*, Wiley-Blackwell.

Daniels, G. (2013b). Variants of RhD--current testing and clinical consequences. *British Journal of Haematology,* **161,** 461-70.

Daniels, G. & Bromilow, I. (2013). *Essential Guide to Blood Groups, 3rd ed*, Wiley-Blackwell.

Daniels, G. & Green, C. (2000). Expression of red cell surface antigens during erythropoiesis. *Vox Sanguinis,* **78 Suppl 2,** 149-53.

Daniels, G. L., Faas, B. H., Green, C. A., Smart, E., Maaskant-Van Wijk, P. A., Avent, N. D., Zondervan, H. A., Von Dem Borne, A. E. & Van Der Schoot, C. E. (1998). The VS and V blood group polymorphisms in Africans: a serologic and molecular analysis. *Transfusion,* **38,** 951-8.

Daniels, G. L., Fletcher, A., Garratty, G., Henry, S., Jorgensen, J., Judd, W. J., Levene, C., Lomas-Francis, C., Moulds, J. J., Moulds, J. M., Moulds, M., Overbeeke, M., Reid, M. E., Rouger, P., Scott, M., Sistonen, P., Smart, E., Tani, Y., Wendel, S. & Zelinski, T. (2004). Blood group terminology 2004: from the International Society of Blood Transfusion committee on terminology for red cell surface antigens. *Vox Sanguinis,* **87,** 304-16.

Denomme, G. A. & Van Oene, M. (2005). High-throughput multiplex single-nucleotide polymorphism analysis for red cell and platelet antigen genotypes. *Transfusion,* **45,** 660-6.

Denomme, G. A., Westhoff, C. M., Castilho, L. M., St-Louis, M., Castro, V. & Reid, M. E. (2010). Consortium for Blood Group Genes (CBGG): 2009 report. *Immunohematology,* **26,** 47-50.

Drago, F., Karpasitou, K. & Poli, F. (2009). Microarray Beads for Identifying Blood Group Single Nucleotide Polymorphisms. *Transfusion Medicine & Hemotherapy,* **36,** 157-160.

Drago, F., Karpasitou, K., Spinardi, L., Crespiatico, L., Scalamogna, M. & Poli, F. (2010). A Microsphere-Based Suspension Array for Blood Group Molecular Typing: An Update. *Transfusion Medicine & Hemotherapy,* **37,** 336-338.

Ensembl Genome Browser. (2015). *Ensembl Genome Browser 82: Homo sapiens* [Online]. Available at: http://www.ensembl.org [Accessed 12 September 2015].

Ewing, B., Hillier, L., Wendl, M. C. & Green, P. (1998). Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment. *Genome Research,* **8,** 175-185.

Eyers, S. A., Ridgwell, K., Mawby, W. J. & Tanner, M. J. (1994). Topology and organization of human Rh (rhesus) blood group-related polypeptides. *The Journal of Biological Chemistry,* **269,** 6417-23.

Faas, B. H., Beckers, E. A., Wildoer, P., Ligthart, P. C., Overbeeke, M. A., Zondervan, H. A., Von Dem Borne, A. E. & Van Der Schoot, C. E. (1997). Molecular background of VS and weak C expression in blacks. *Transfusion,* **37,** 38-44.

Fichou, Y., Audrezet, M. P., Gueguen, P., Le Marechal, C. & Ferec, C. (2014). Next-generation sequencing is a credible strategy for blood group genotyping. *British Journal of Haematology,* **167,** 554-62.

Flegel, W. A., Von Zabern, I. & Wagner, F. F. (2009). Six years' experience performing RHD genotyping to confirm D- red blood cell units in Germany for preventing anti-D immunizations. *Transfusion,* **49,** 465-71.

Flegel, W. A. & Wagner, F. F. (2002). Molecular biology of partial D and weak D: implications for blood bank practice. *Clinical Laboratory,* **48,** 53-9.

Freda, V. J., Gorman, J. G. & Pollack, W. (1966). Rh factor: prevention of isoimmunization and clinical trial on mothers. *Science,* **151,** 828-30.

Fukumori, Y., Ohnoki, S., Shibata, H., Yamaguchi, H. & Nishimukai, H. (1995). Genotyping of ABO blood groups by PCR and RFLP analysis of 5 nucleotide positions. *International Journal of Legal Medicine,* **107,** 179-82.

Gabriel, C., Stabentheiner, S., Danzer, M. & Proll, J. (2011). What Next? The Next Transit from Biology to Diagnostics: Next Generation Sequencing for Immunogenetics. *Transfusion Medicine & Hemotherapy,* **38,** 308-317.

Gonzalez, C. E. & Pengetze, Y. M. (2005). Post-transfusion purpura. *Current Hematology Reports,* **4,** 154-9.

Grobel, R. K. & Cardy, J. D. (1971). Hemolytic disease of the newborn due to anti-EW. A fourth example of the Rh antigen, EW. *Transfusion,* **11,** 77-8.

Halawani, A. J., Altayar, M. A., Kiernan, M., Reynolds, A. J., Madgett, T. E. & Avent, N. D. (2014). Human Erythrocyte Antigens and Human Platelet Antigens Panel: A Genotyping Protocol Based on Next-generation Sequencing. *Transfusion Medicine,* **24, suppl. 2,** 1-32.

Hashmi, G. (2007). Red blood cell antigen phenotype by DNA analysis. *Transfusion,* **47,** 60S-3S.

Hashmi, G., Shariff, T., Seul, M., Vissavajjhala, P., Hue-Roye, K., Charles-Pierre, D., Lomas-Francis, C., Chaudhuri, A. & Reid, M. E. (2005). A flexible array format for large-scale, rapid blood group DNA typing. *Transfusion,* **45,** 680-8.

Hashmi, G., Shariff, T., Zhang, Y., Cristobal, J., Chau, C., Seul, M., Vissavajjhala, P., Baldwin, C., Hue-Roye, K., Charles-Pierre, D., Lomas-Francis, C. & Reid, M. E. (2007). Determination of 24 minor red blood cell antigens for more than 2000 blood donors by high-throughput DNA analysis. *Transfusion,* **47,** 736-47.

Hert, D. G., Fredlake, C. P. & Barron, A. E. (2008). Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis,* **29,** 4618-26.

Ho, M., Chelly, J., Carter, N., Danek, A., Crocker, P. & Monaco, A. P. (1994). Isolation of the gene for McLeod syndrome that encodes a novel membrane transport protein. *Cell,* **77,** 869-80.

Hodkinson, B. P. & Grice, E. A. (2015). Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. *Advances in Wound Care,* **4,** 50-58.

Hpa Sequence Database. (2015). *Immuno Polymorphism Database: All HPA - alloantigen/protein data* [Online]. Available at: https://www.ebi.ac.uk/ipd/hpa/table1.html [Accessed 17 October 2015].

Huang, C. H., Chen, Y., Reid, M. E. & Seidl, C. (1998). Rhnull disease: the amorph type results from a novel double mutation in RhCe gene on D-negative background. *Blood,* **92,** 664-71.

Hurd, C. M., Cavanagh, G., Schuh, A., Ouwehand, W. H. & Metcalfe, P. (2002). Genotyping for platelet-specific antigens: techniques for the detection of single nucleotide polymorphisms. *Vox Sanguinis,* **83,** 1-12.

Hurd, P. J. & Nelson, C. J. (2009). Advantages of next-generation sequencing versus the microarray in epigenetic research. *Briefings in Functional Genomics & Proteomics,* **8,** 174-183.

International Society of Blood Transfusion. (2015). *Red Cell Immunogenetics and Blood Group Terminology* [Online]. Available at: http://www.isbtweb.org [Accessed 14 October 2015].

Ion Ampliseq Designer. (2014). Available at: http://www.ampliseq.com [Accessed 2 January 2014].

Jia, H., Guo, Y., Zhao, W. & Wang, K. (2014). Long-range PCR in next-generation sequencing: comparison of six enzymes and evaluation on the MiSeq sequencer. *Scientific Reports,* **4,** 5737.

Jones, J., Scott, M. L. & Voak, D. (1995). Monoclonal anti-D specificity and Rh D structure: criteria for selection of monoclonal anti-D reagents for routine typing of patients and donors. *Transfusion Medicine,* **5,** 171-84.

Jung, H. H., Danek, A. & Frey, B. M. (2007). McLeod syndrome: a neurohaematological disorder. *Vox Sang,* **93,** 112-21.

Jungbauer, C., Hobel, C. M., Schwartz, D. W. & Mayr, W. R. (2012). High-throughput multiplex PCR genotyping for 35 red blood cell antigens in blood donors. *Vox Sanguinis,* **102,** 234-42.

Karpasitou, K., Drago, F., Crespiatico, L., Paccapelo, C., Truglio, F., Frison, S., Scalamogna, M. & Poli, F. (2008). Blood group genotyping for Jk(a)/Jk(b), Fy(a)/Fy(b), S/s, K/k, Kp(a)/Kp(b), Js(a)/Js(b), Co(a)/Co(b), and Lu(a)/Lu(b) with microarray beads. *Transfusion,* **48,** 505-12.

Khamlichi, S., Bailly, P., Blanchard, D., Goossens, D., Cartron, J. P. & Bertrand, O. (1995). Purification and partial characterization of the erythrocyte Kx protein deficient in McLeod patients. *European Journal of Biochemistry,* **228,** 931-4.

Kiefel, V., Santoso, S., Katzmann, B. & Mueller-Eckhardt, C. (1989). The Bra/Brb alloantigen system on human platelets. *Blood,* **73,** 2219-2223.

Koelewijn, J. M., Vrijkotte, T. G., Van Der Schoot, C. E., Bonsel, G. J. & De Haas, M. (2008). Effect of screening for red cell antibodies, other than anti-D, to detect hemolytic disease of the fetus and newborn: a population study in the Netherlands. *Transfusion,* **48,** 941-52.

Körmöczi, G. F., Gassner, C., Shao, C.-P., Uchikawa, M. & Legler, T. J. (2005). A comprehensive analysis of DEL types: partial DEL individuals are prone to anti-D alloimmunization. *Transfusion,* **45,** 1561-1567.

Koshy, R., Patel, B. & Harrison, J. S. (2009). Anti-Kpa-induced severe delayed hemolytic transfusion reaction. *Immunohematology,* **25,** 44-7.

Kudo, T., Iwasaki, H., Nishihara, S., Shinya, N., Ando, T., Narimatsu, I. & Narimatsu, H. (1996). Molecular genetic analysis of the human Lewis histo-blood group system. II. Secretor gene inactivation by a novel single missense mutation

A385T in Japanese nonsecretor individuals. *The Journal of Biological Chemistry,* **271,** 9830-7.

Kumpel, B. M. (2008). Lessons learnt from many years of experience using anti-D in humans for prevention of RhD immunization and haemolytic disease of the fetus and newborn. *Clinical and Experimental Immunology,* **154,** 1-5.

Kumpel, B. M. & Elson, C. J. (2001). Mechanism of anti-D-mediated immune suppression--a paradox awaiting resolution? *Trends in Immunology,* **22,** 26-31.

Landsteiner, K. (1961). On agglutination of normal human blood. *Transfusion,* **1,** 5-8.

Laver, T., Harrison, J., O'neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K. & Studholme, D. J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection & Quantification,* **3,** 1-8.

Le Van Kim, C., Mouro, I., Cherif-Zahar, B., Raynal, V., Cherrier, C., Cartron, J. P. & Colin, Y. (1992). Molecular cloning and primary structure of the human blood group RhD polypeptide. *Proceedings of the National Academy of Sciences USA,* **89,** 10925-9.

Lee, S. (1997). Molecular basis of Kell blood group phenotypes. *Vox Sanguinis,* **73,** 1-11.

Lee, S., Debnath, A. K. & Redman, C. M. (2003a). Active amino acids of the Kell blood group protein and model of the ectodomain based on the structure of neutral endopeptidase 24.11. *Blood,* **102,** 3028-34.

Lee, S., Lin, M., Mele, A., Cao, Y., Farmar, J., Russo, D. & Redman, C. (1999). Proteolytic processing of big endothelin-3 by the kell blood group protein. *Blood,* **94,** 1440-50.

Lee, S., Naime, D. S., Reid, M. E. & Redman, C. M. (1997). Molecular basis for the high-incidence antigens of the Kell blood group system. *Transfusion,* **37,** 1117-22.

Lee, S., Russo, D. C., Reid, M. E. & Redman, C. M. (2003b). Mutations that diminish expression of Kell surface protein and lead to the Kmod RBC phenotype. *Transfusion,* **43,** 1121-5.

Lee, S., Russo, D. C., Reiner, A. P., Lee, J. H., Sy, M. Y., Telen, M. J., Judd, W. J., Simon, P., Rodrigues, M. J., Chabert, T., Poole, J., Jovanovic-Srzentic, S., Levene, C., Yahalom, V. & Redman, C. M. (2001). Molecular defects

underlying the Kell null phenotype. *The Journal of Biological Chemistry,* **276,** 27281-9.

Lee, S., Wu, X., Reid, M., Zelinski, T. & Redman, C. (1995). Molecular basis of the Kell (K1) phenotype. *Blood,* **85,** 912-6.

Lee, S., Zambas, E. D., Marsh, W. L. & Redman, C. M. (1991). Molecular cloning and primary structure of Kell blood group protein. *Proceedings of the National Academy of Sciences USA,* **88,** 6353-7.

Lee, S., Zambas, E. D., Marsh, W. L. & Redman, C. M. (1993). The human Kell blood group gene maps to chromosome 7q33 and its expression is restricted to erythroid cells. *Blood,* **81,** 2804-9.

Legler, T. J., Maas, J. H., Köhler, M., Wagner, T., Daniels, G. L., Perco, P. & Panzer, S. (2001). RHD sequencing: a new tool for decision making on transfusion therapy and provision of Rh prophylaxis. *Transfusion Medicine,* **11,** 383-388.

Levine, P., Burnham, L., Katzin, E. M. & Vogel, P. (1941). The role of iso-immunization in the pathogenesis of erythroblastosis fetalis. *American Journal of Obstetrics & Gynecology,* **42,** 925-937.

Liu, Z., Liu, M., Mercado, T., Illoh, O. & Davey, R. (2014). Extended blood group molecular typing and next-generation sequencing. *Transfusion Medicine Reviews,* **28,** 177-86.

López Marínez, M., Chinnapapagari, S. K. R., Olsson, M. L., Nogués, N., Scott, M. L., Písacka, M., Daniels, G., Van Der Schoot, E., Muniz-Diaz, E., Madgett, T. E., Storry, J. R., Beiboer, S. H., Maaskant -Vanwijk, P. A., Von Zabern, I., Jiménez, E., Tejedor, J., Azkarate, M., Vesga, M. A., Camacho, E., Cheroutre, G., Link, A., Jinoch, P., Svobodova, I., Martinez, A., De Haas, M., Flegel, W. A. & Avent, N. D. 2009. MASS-SCALE EXTENSIVE GENOTYPING OF 3000 RBCSAMPLES BY BLOODCHIP- V 1.0. *Vox Sanguinis.* Blackwell Publishing Ltd.

Lucas & Metcalfe (2000). Platelet and granulocyte glycoprotein polymorphisms. *Transfusion Medicine,* **10,** 157-174.

Mardis, E. R. (2008). Next-Generation DNA Sequencing Methods. *Annual Review of Genomics & Human Genetics,* **9,** 387-402.

Mardis, E. R. (2013). Next-Generation Sequencing Platforms. *Annual Review of Analytical Chemistry,* **6,** 287-303.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., Mcdade, K. E., Mckenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F. & Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* **437,** 376-80.

Marini, A. M., Urrestarazu, A., Beauwens, R. & Andre, B. (1997). The Rh (rhesus) blood group polypeptides are related to NH4+ transporters. *Trends in Biochemical Sciences,* **22,** 460-1.

Masouredis, S. P., Sudora, E., Mahan, L. C. & Victoria, E. J. (1980). Immunoelectron microscopy of Kell and Cellano antigens on red cell ghosts. *Haematologia* **13,** 59-64.

Mazonson, P., Efrusy, M., Santas, C., Ziman, A., Burner, J., Roseff, S., Vijayaraghavan, A. & Kaufman, R. (2014). The HI-STAR study: resource utilization and costs associated with serologic testing for antibody-positive patients at four United States medical centers. *Transfusion,* **54,** 271-277.

Mcbean, R. S., Hyland, C. A. & Flower, R. L. (2014). Approaches to determination of a full profile of blood group genotypes: single nucleotide variant mapping and massively parallel sequencing. *Computational & Structural Biotechnology Journal,* **11,** 147-51.

Mcgann, H. & Wenk, R. E. (2010). Alloimmunization to the D antigen by a patient with weak D type 21. *Immunohematology,* **26,** 27-9.

Metcalfe, P., Watkins, N. A., Ouwehand, W. H., Kaplan, C., Newman, P., Kekomaki, R., De Haas, M., Aster, R., Shibata, Y., Smith, J., Kiefel, V. & Santoso, S. (2003). Nomenclature of human platelet antigens. *Vox Sanguinis,* **85,** 240-245.

Mouro, I., Colin, Y., Cherif-Zahar, B., Cartron, J.-P. & Van Kim, C. L. (1993). Molecular genetic basis of the human Rhesus blood group system. *Nature Genetics,* **5,** 62-65.

Mouro, I., Colin, Y., Sistonen, P., Le Pennec, P. Y., Cartron, J. P. & Le Van Kim, C. (1995). Molecular basis of the RhCW (Rh8) and RhCX (Rh9) blood group specificities. *Blood,* **86,** 1196-201.

Mueller-Eckhardt, C., Kiefel, V., Grubert, A., Kroll, H., Weisheit, M., Schmidt, S., Mueller-Eckhardt, G. & Santoso, S. (1989). 348 cases of suspected neonatal alloimmune thrombocytopenia. *Lancet,* **1,** 363-6.

Mullins, F. M., Dietz, L., Lay, M., Zehnder, J. L., Ford, J., Chun, N. & Schrijver, I. (2007). Identification of an intronic single nucleotide polymorphism leading to allele dropout during validation of a CDH1 sequencing assay: implications for designing polymerase chain reaction-based assays. *Genetics in Medicine,* **9,** 752-760.

Mullis, K. B., Erlich, H. A., Arnheim, N., Horn, G. T., Saiki, R. K. & Scharf, S. J. 1987. Process for amplifying, detecting, and/or-cloning nucleic acid sequences. Google Patents.

National Centre for Biotechnology Information Primer Blast. (2015). *Primer Designing Tool* [Online]. Available at: http://www.ncbi.nlm.nih.gov/tools/primer-blast/ [Accessed 3 Febraury 2013].

Newman, P. J., Derbes, R. S. & Aster, R. H. (1989). The human platelet alloantigens, PlA1 and PlA2, are associated with a leucine33/proline33 amino acid polymorphism in membrane glycoprotein IIIa, and are distinguishable by DNA typing. *The Journal of Clinical Investigation,* **83,** 1778-81.

Nishihara, S., Narimatsu, H., Iwasaki, H., Yazawa, S., Akamatsu, S., Ando, T., Seno, T. & Narimatsu, I. (1994). Molecular genetic analysis of the human Lewis histo-blood group system. *The Journal of Biological Chemistry,* **269,** 29271-8.

Noizat-Pirenne, F., Lee, K., Pennec, P. Y., Simon, P., Kazup, P., Bachir, D., Rouzaud, A. M., Roussel, M., Juszczak, G., Menanteau, C., Rouger, P., Kotb, R., Cartron, J. P. & Ansart-Pirenne, H. (2002). Rare RHCE phenotypes in black individuals of Afro-Caribbean origin: identification and transfusion safety. *Blood,* **100,** 4223-31.

Novotny, V. M. (1999). Prevention and management of platelet transfusion refractoriness. *Vox Sanguinis,* **76,** 1-13.

O'keefe, D. S. & Dobrovic, A. (1996). Decreased stability of the O allele mRNA transcript of the ABO gene. *Blood,* **87,** 3061-2.

Pamphilon, D. H. & Scott, M. L. (2007). Robin Coombs: his life and contribution to haematology and transfusion medicine. *British Journal of Haematology,* **137,** 401-8.

Paris, S., Rigal, D., Barlet, V., Verdier, M., Coudurier, N., Bailly, P. & Bres, J. C. (2014). Flexible automated platform for blood group genotyping on DNA microarrays. *The journal of Molecular Diagnostics,* **16,** 335-42.

Peng, C. T., Shih, M. C., Liu, T. C., Lin, I. L., Jaung, S. J. & Chang, J. G. (2003). Molecular basis for the RhD negative phenotype in Chinese. *International Journal of Molecular Medicine,* **11,** 515-21.

Polin, H., Danzer, M., Proll, J., Hofer, K., Heilinger, U., Zopf, A. & Gabriel, C. (2008). Introduction of a real-time-based blood-group genotyping approach. *Vox Sanguinis,* **95,** 125-30.

Poole, J. & Daniels, G. (2007). Blood group antibodies and their significance in transfusion medicine. *Transfusion Medicine Reviews,* **21,** 58-71.

Prager, M. (2007). Molecular genetic blood group typing by the use of PCR-SSP technique. *Transfusion,* **47,** 54S-59S.

Prober, J. M., Trainor, G. L., Dam, R. J., Hobbs, F. W., Robertson, C. W., Zagursky, R. J., Cocuzza, A. J., Jensen, M. A. & Baumeister, K. (1987). A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science,* **238,** 336-41.

Ratanaphan, A., Panomwan, P., Canyuk, B. & Maipang, T. (2011). Identification of novel intronic BRCA1 variants of uncertain significance in a Thai hereditary breast cancer family. *Journal of Genetics,* **90,** 327-31.

Reid, M. E. (2003). Applications of DNA-based assays in blood group antigen and antibody identification. *Transfusion,* **43,** 1748-1757.

Reid, M. E. (2009). Transfusion in the age of molecular diagnostics. *Hematology / the Education Program of the American Society of Hematology.***,** 171-7.

Reid, M. E., Lomas-Francis, C. & Olsson, M. L. (2012). *The Blood Group Antigen FactsBook, 3rd ed,* Boston, Academic Press.

Reid, M. E., Rios, M., Powell, V. I., Charles-Pierre, D. & Malavade, V. (2000). DNA from blood samples can be used to genotype patients who have recently received a transfusion. *Transfusion,* **40,** 48-53.

Rieneck, K., Bak, M., Jonson, L., Clausen, F. B., Krog, G. R., Tommerup, N., Nielsen, L. K., Hedegaard, M. & Dziegiel, M. H. (2013). Next-generation sequencing: proof of concept for antenatal prediction of the fetal Kell blood group phenotype from cell-free fetal DNA in maternal plasma. *Transfusion,* **53,** 2892-8.

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology,* **29,** 24-26.

Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M., Harte, R. A., Heitner, S., Hickey, G., Hinrichs, A. S., Hubley, R., Karolchik, D., Learned, K., Lee, B. T., Li, C. H., Miga, K. H., Nguyen, N., Paten, B., Raney, B. J., Smit, A. F., Speir, M. L., Zweig, A. S., Haussler, D., Kuhn, R. M. & Kent, W. J. (2015). The UCSC Genome Browser database: 2015 update. *Nucleic Acids Research,* **43,** D670-81.

Rossi, K. Q., Scrape, S., Lang, C. & O'shaughnessy, R. (2013). Severe hemolytic disease of the fetus due to anti-Kpa antibody. *International Journal of Blood Transfusion & Immunohematology,* **3,** 18-20.

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W., Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., Clark, T. A., Huber, M., Branciforte, J. T., Stoner, I. B., Cawley, S. E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J. A., Namsaraev, E., Mckernan, K. J., Williams, A., Roth, G. T. & Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature,* **475,** 348-52.

Russo, D., Redman, C. & Lee, S. (1998). Association of XK and Kell blood group proteins. *The Journal of Biological Chemistry,* **273,** 13950-6.

Russo, D., Wu, X., Redman, C. M. & Lee, S. (2000). Expression of Kell blood group protein in nonerythroid tissues. *Blood,* **96,** 340-6.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences USA,* **74,** 5463-5467.

Scholtalbers, J., Rößler, J., Sorn, P., De Graaf, J., Boisguérin, V., Castle, J. & Sahin, U. (2013). Galaxy LIMS for next-generation sequencing. *Bioinformatics,* **29,** 1233-1234.

Scott, M. L. (2001). Monoclonal anti-D for immunoprophylaxis. *Vox Sanguinis,* **81,** 213-8.

Scott, M. L., Voak, D., Jones, J. W., Avent, N. D., Liu, W., Hughes-Jones, N. & Sonneborn, H. (1996). A structural model for 30 Rh D epitopes based on serological and DNA sequence data from partial D phenotypes. *Transfusion Clinique et Biologique,* **3,** 391-6.

Scott, M. L., Voak, D., Liu, W., Jones, J. W. & Avent, N. D. (2000). Epitopes on Rh proteins. *Vox Sanguinis,* **78 Suppl 2,** 117-20.

Seattleseq Annotation Tool 141. (2014). *SeattleSeq Variation Annotation* [Online]. Available at: http://snp.gs.washington.edu/.

Shanker, A. (2012). Genome research in the cloud. *Omics,* **16,** 422-8.

Shao, C. P., Maas, J. H., Su, Y. Q., Kohler, M. & Legler, T. J. (2002). Molecular background of Rh D-positive, D-negative, D(el) and weak D phenotypes in Chinese. *Vox Sanguinis,* **83,** 156-61.

Shendure, J. & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology,* **26,** 1135-45.

Shiina, T., Suzuki, S., Ozaki, Y., Taira, H., Kikkawa, E., Shigenari, A., Oka, A., Umemura, T., Joshita, S., Takahashi, O., Hayashi, Y., Paumen, M., Katsuyama, Y., Mitsunaga, S., Ota, M., Kulski, J. K. & Inoko, H. (2012). Super high resolution for single molecule-sequence-based typing of classical HLA loci at the 8-digit level using next generation sequencers. *Tissue Antigens,* **80,** 305-16.

Siebert, P. D. & Fukuda, M. (1986). Isolation and characterization of human glycophorin A cDNA clones by a synthetic oligonucleotide approach: nucleotide sequence and mRNA structure. *Proc Natl Acad Sci U S A,* **83,** 1665-9.

Simsek, S., De Jong, C. A., Cuijpers, H. T., Bleeker, P. M., Westers, T. M., Overbeeke, M. A., Goldschmeding, R., Van Der Schoot, C. E. & Von Dem Borne, A. E. (1994). Sequence analysis of cDNA derived from reticulocyte mRNAs coding for Rh polypeptides and demonstration of E/e and C/c polymorphisms. *Vox Sanguinis,* **67,** 203-9.

Single Nucleotide Polymorphisms Database. (2014). *dbSNP: Short Genetic Variations* [Online]. Available at: http://www.ncbi.nlm.nih.gov/projects/SNP/] [Accessed 21 May 2014].

Singleton, B. K., Green, C. A., Avent, N. D., Martin, P. G., Smart, E., Daka, A., Narter-Olaga, E. G., Hawthorne, L. M. & Daniels, G. (2000). The presence of an RHD pseudogene containing a 37 base pair duplication and a nonsense mutation in africans with the Rh D-negative blood group phenotype. *Blood,* **95,** 12-8.

Smoleniec, J., Anderson, N. & Poole, G. (1994). Hydrops fetalis caused by a blood group antibody usually undetected in routine screening. *Archives of Disease in Childhood Fetal & Neonatal edition,* **71,** F216-F217.

Southcott, M. J., Tanner, M. J. & Anstee, D. J. (1999). The expression of human blood group antigens during erythropoiesis in a cell culture system. *Blood,* **93,** 4425-35.

Stanfield, G. M. & Horvitz, H. R. (2000). The ced-8 gene controls the timing of programmed cell deaths in C. elegans. *Molecular Cell,* **5,** 423-33.

Storry, J. R., Joud, M., Christophersen, M. K., Thuresson, B., Akerstrom, B., Sojka, B. N., Nilsson, B. & Olsson, M. L. (2013). Homozygosity for a null allele of SMIM1 defines the Vel-negative blood group phenotype. *Nature Genetics,* **45,** 537-41.

Storry, J. R. & Olsson, M. L. (2004). Genetic basis of blood group diversity. *British Journal of Haematology,* **126,** 759-71.

Strobel, E., Noizat-Pirenne, F., Hofmann, S., Cartron, J. P. & Bauer, M. F. (2004). The molecular basis of the Rhesus antigen Ew. *Transfusion,* **44,** 407-9.

Sun, C. F., Chou, C. S., Lai, N. C. & Wang, W. T. (1998). RHD gene polymorphisms among RhD-negative Chinese in Taiwan. *Vox Sanguinis,* **75,** 52-7.

Tax, M. G., Van Der Schoot, C. E., Van Doorn, R., Douglas-Berger, L., Van Rhenen, D. J. & Maaskant-Vanwijk, P. A. (2002). RHC and RHc genotyping in different ethnic groups. *Transfusion,* **42,** 634-44.

The Rhesusbase. (2015). *The RHD Mutation Database* [Online]. Available at: http://www.rhesusbase.info/ [Accessed 21 September 2015].

Thermo Fisher Scientific. (2012). *Preparing Long Amplicon (>400 bp) Libraries Using the Ion XpressTM Plus Fragment Library Kit Publication No. MAN0007044, Revision 3* [Online]. Available at: https://ioncommunity.thermofisher.com/community/protocols-home/pgm-protocols [Accessed 13 March 2013].

Thermo Fisher Scientific. (2014a). *The Ion Proton System* [Online]. Available at: https://tools.thermofisher.com/content/sfs/brochures/CO06326_Proton_Spec_Sheet_FHR.pdf [Accessed 25 October 2012].

Thermo Fisher Scientific. (2014b). *Ion PGM Template OT2 200 Kit User Guide Catalog No. 4480974 Publication No. MAN0007220 Rev. A.0.* [Online]. Available at: https://ioncommunity.thermofisher.com/community/protocols-home/pgm-protocols [Accessed 7 August 2014].

Thermo Fisher Scientific. (2015a). *The Ion PGM System* [Online]. Available at: https://tools.thermofisher.com/content/sfs/brochures/PGM-Specification-Sheet.pdf [Accessed 25 October 2015].

Thermo Fisher Scientific. (2015b). *Ion Hi-Q™ Chemistry for the Ion PGM™ System* [Online]. Available at: https://tools.thermofisher.com/content/sfs/brochures/Ion-PGM-Hi-Q-Sequencing-Kit-Flyer.pdf [Accessed 26 October 2015].

Tilley, L. & Grimsley, S. (2014). Is Next Generation Sequencing the future of blood group testing? *Transfusion & Apheresis Science,* **50,** 183-8.

Tuson, M., Hue-Roye, K., Koval, K., Imlay, S., Desai, R., Garg, G., Kazem, E., Stockman, D., Hamilton, J. & Reid, M. E. (2011). Possible suppression of fetal erythropoiesis by the Kell blood group antibody anti-Kp(a). *Immunohematology,* **27,** 58-60.

Ucsc Genome Bioinformatics. (2015). *Human (Homo sampiens) Genome Browser Gateway* [Online]. Available at: http://genome.ucsc.edu/cgi-bin/hgGateway [Accessed 13 September 2015].

Urbaniak, S. J. & Greiss, M. A. (2000). RhD haemolytic disease of the fetus and the newborn. *Blood Reviews,* **14,** 44-61.

Vaughan, J. I., Manning, M., Warwick, R. M., Letsky, E. A., Murray, N. A. & Roberts, I. a. G. (1998). Inhibition of Erythroid Progenitor Cells by Anti-Kell Antibodies in Fetal Alloimmune Anemia. *New England Journal of Medicine,* **338,** 798-803.

Veldhuisen, B., Van Der Schoot, C. E. & De Haas, M. (2009). Blood group genotyping: from patient to high-throughput donor screening. *Vox Sanguinis,* **97,** 198-206.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., Mckusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., et al. (2001). The sequence of the human genome. *Science,* **291,** 1304-51.

Von Dem Borne, A. E. G. & Décary, F. (1990). ICSH/ISBT Working Party on Platelet Serology. *Vox Sanguinis,* **58,** 176-176.

Vrignaud, C., Ramelet, S., Pecquet, F., Hennion, M., Narboux, C., Braun, F., Nataf, J. & Peyrard, T. (2014). Characterization of 11 Novel RHCE Alleles In French Blood Donors *Vox Sanguinis* **107 (Suppl. 1),** 187.

Wagner, F. F. & Flegel, W. A. (2000). RHD gene deletion occurred in the Rhesus box. *Blood,* **95,** 3662-8.

Wagner, F. F. & Flegel, W. A. (2004). Review: the molecular basis of the Rh blood group phenotypes. *Immunohematology,* **20,** 23-36.

Wagner, F. F., Frohmajer, A., Ladewig, B., Eicher, N. I., Lonicer, C. B., Muller, T. H., Siegel, M. H. & Flegel, W. A. (2000). Weak D alleles express distinct phenotypes. *Blood,* **95,** 2699-708.

Wagner, F. F., Gassner, C., Muller, T. H., Schonitzer, D., Schunter, F. & Flegel, W. A. (1999). Molecular basis of weak D phenotypes. *Blood,* **93,** 385-93.

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., Guo, Y., Feng, B., Li, H., Lu, Y., Fang, X., Liang, H., Du, Z., Li, D., Zhao, Y., Hu, Y., Yang, Z., Zheng, H., Hellmann, I., Inouye, M., Pool, J., Yi, X., Zhao, J., Duan, J., Zhou, Y., Qin, J., Ma, L., Li, G., Yang, Z., Zhang, G., Yang, B., Yu, C., Liang, F., Li, W., Li, S., Li, D., Ni, P., Ruan, J., Li, Q., Zhu, H., Liu, D., Lu, Z., Li, N., Guo, G., Zhang, J., Ye, J., Fang, L., Hao, Q., Chen, Q., Liang, Y., Su, Y., San, A., Ping, C., Yang, S., Chen, F., Li, L., Zhou, K., Zheng, H., Ren, Y., Yang, L., Gao, Y., Yang, G., Li, Z., Feng, X., Kristiansen, K., Wong, G. K.-S., Nielsen, R., Durbin, R., Bolund, L., Zhang, X., Li, S., Yang, H. & Wang, J. (2008). The diploid genome sequence of an Asian individual. *Nature,* **456,** 60-65.

Weiner, C. P. & Widness, J. A. (1996). Decreased fetal erythropoiesis and hemolysis in Kell hemolytic anemia. *American Journal of Obstetrics & Gynecology,* **174,** 547-51.

Westhoff, C. M. (2006). Molecular testing for transfusion medicine. *Current Opinion in Hematology,* **13,** 471-475.

Westhoff, C. M. & Reid, M. E. (2004). Review: the Kell, Duffy, and Kidd blood group systems. *Immunohematology,* **20,** 37-49.

Westhoff, C. M., Silberstein, L. E., Wylie, D. E., Skavdahl, M. & Reid, M. E. (2001). 16Cys encoded by the RHce gene is associated with altered expression of the e antigen and is frequent in the R0 haplotype. *British Journal of Haematology,* **113,** 666-71.

Wienzek-Lischka, S., Krautwurst, A., Fröhner, V., Hackstein, H., Gattenlöhner, S., Bräuninger, A., Axt-Fliedner, R., Degenhardt, J., Deisting, C., Santoso, S., Sachs, U. J. & Bein, G. (2015). Noninvasive fetal genotyping of human platelet

antigen-1a using targeted massively parallel sequencing. *Transfusion,* **55,** 1538-1544.

Xu, Q., Grootkerk-Tax, M. G., Maaskant-Van Wijk, P. A. & Van Der Schoot, C. E. (2005). Systemic analysis and zygosity determination of the RHD gene in a D-negative Chinese Han population reveals a novel D-negative RHD gene. *Vox Sanguinis,* **88,** 35-40.

Yamamoto, F., Marken, J., Tsuji, T., White, T., Clausen, H. & Hakomori, S. (1990). Cloning and characterization of DNA complementary to human UDP-GalNAc: Fuc alpha 1----2Gal alpha 1----3GalNAc transferase (histo-blood group A transferase) mRNA. *The Journal of Biological Chemistry,* **265,** 1146-51.

Yamamoto, F., Mcneill, P. D. & Hakomori, S. (1992). Human histo-blood group A2 transferase coded by A2 allele, one of the A subtypes, is characterized by a single base deletion in the coding sequence, which results in an additional domain at the carboxyl terminal. *Biochemical & Biophysical Research Communications,* **187,** 366-74.

Appendices

Appendix A

Margaret Kenwright Young Scientist Award 2014 from British Blood Transfusion

Society (BBTS) for the following work:

Halawani, A. J., Altayar, M. A., Kiernan, M., Reynolds, A. J., Madgett, T. E. & Avent, N. D. 2014. Human Erythrocyte Antigens and Human Platelet Antigens Panel: A Genotyping Protocol Based on Next-Generation Sequencing. *Transfusion Medicine,* **24,** suppl. 2**,** 1-32.

Alloimmunisation occurs when there is disparity between the antigens of donor red cells and the patient during blood transfusion. Therefore it is important to have comprehensive information regarding the antigens on the donor and patient red cells. Previous platforms of genotyping techniques have several limitations such as cost, requirement of previous knowledge of single nucleotide polymorphisms (SNPs) and identification of hybrid genes. Next-generation sequencing (NGS) is able to provide high-throughput sequencing and capture many targets within the human genome. A protocol has been designed to genotype human erythrocyte antigens (HEA) and human platelet antigens (HPA) based on NGS. The protocol is for 11 blood group systems (ABO, RH, Kell, Kidd, Duffy, MNS, Diego, Dombrock, Yt, Colton and Vel) and HPA (1–16). The method is based on Ion Ampliseq™ Custom Panel provided by Life Technologies. Four weak D donor blood samples were amplified using ultra- high multiplex PCR. The primers were then partially digested and ligated to barcoded adaptors. After that, the sequencing libraries were added to the beads for clonal amplification and enrichment. The sequencing templates were finally loaded onto a 314 Chip and have been sequenced on an Ion Torrent Personal Genome Machine™. The data were analysed using CLC Genomic Workbench Version 7.0. The output results generated more than 400 K sequencing reads with an average coverage depth of 700. Out of the four samples the following variants were found: Weak D Type 1 (*RHD*01W.01*), Weak D Type 2 (*RHD*01W.02*), Weak D Type 4.0 (*RHD*DAR3.01*), *GYPB*S, DO*A, JK*B, FY*A, FY*B, JK*01W.01, HPA-1a/HPA-1b, HPA-3a/HPA-3b, HPA-5a/HPA-5b* and *HPA-15b/ HPA-15b*. The weak D sample genotyped as (809T>G, Val270Gly), which is a Weak D Type 1, also had a novel SNP on exon 2 (208C>T, Arg70Trp). Confirmation of this SNP will be done by long-range PCR of both *RHD* and *RHCE* genes. We have described here a rapid approach to sequence all clinically significant blood group and HPA alleles. The method would be readily applicable to blood donors and alloimmunised patients. Ultimately, NGS will pave the way to genotype blood group variants in the near future.

Conference proceedings

Halawani, A. J., Altayar, M. A., Kiernan, M., Kaushik, N., Reynolds, A. J., Madgett, T. E. & Avent, N. D. 2013. Comprehensive Genotyping for Kell and Rh Blood Group Systems by Next-generation DNA Sequencing. *Transfusion Medicine,* **23,** suppl. 2**,** 30-71.

Blood group genotyping (BGG) has emerged as a core technique in transfusion medicine and has impacted on the clinical management of multi-transfused patients. The vast majority of these technical platforms require previous knowledge of the blood group polymorphisms under investigation. Next-generation sequencing (NGS) has emerged as a powerful replacement technology to genotyping single nucleotide polymorphisms (SNP), insertion/deletion (indels) and gene rearrangements. We have used NGS to define Rh and Kell alleles by amplification of the entire genes (KEL RHD, and RHCE). DNA was extracted from 14 random blood donor samples. Two primer pairs were designed to amplify the entire KEL gene using long-range polymerase chain reaction (LR-PCR). The sequencing library was constructed by fragmenting the DNA, ligating into barcoded adaptors and size selected using SPRIselect magnetic beads. The sequencing template was then immobilised to sphere particles that clonally amplified using emulsion PCR, emulsion breaking and enrichment for positive sphere particles. Finally, the sequencing reaction was loaded into a chip and sequenced on the Ion Torrent Personal Genome Machine (PGM) that generating 2·6 million reads. Data was analysed using CLC Genomics workbench Version 6·0.4. Data from the serological testing was confirmed by NGS. Two samples were typed serologically as K antigen and four as Kp^b and this was confirmed by NGS as an initial genotype as [K antigen, Thr193Met (578C>T), heterozygous SNP 53·4% and 46·3%, respectively] and Kp^b (Arg281, 841C and 842G). Moreover, one sample was found to be Kp^a Arg281Trp (841C>T). Genotyping will be performed for all the antigens of Kell blood group system that encode the following antigens: (K/k, Kp^a/Kp^b/Kp^c, Js^a/Js^b, K11/K17, K12, K13, K14/K24, VLAN/VONG, K18, K19, Km, K22, TOU, RAZ, KALT, KTIM, KYO/KYOR, KUCI, KANT, KASH, KELP, KETI and KHUL). NGS using LR-PCR approach offers a powerful technique enabling users to investigate comprehensive screening of all the Kell and Rh antigens. Similar approaches are in progress to define Rh alleles and that will reveal their variants in respect of the hybrid genes.

Altayar, M. A., **Halawani, A. J**., Kiernan, M., Kaushik, N., Reynolds, A. J., Madgett, T. E. & Avent, N. D. 2013. Next Generation Sequencing of ABO, Duffy and Kidd Blood Group Genotyping. *Transfusion Medicine,* **23,** suppl. 2**,** 30-71.

Blood group Genotyping (BGG) has become well established in transfusion medicine. However, all current technologies are based on pre-existing knowledge of known polymorphisms. Next generation sequencing circumvents this requirement and adopts a discovery mode, which is important, as almost every new BGG project reveals new alleles. NGS has become high-throughput, rapid and accurate. Also, costs have significantly reduced in the past five years. In this pilot study, Ion Torrent Personal Genome Machine (PGM) sequencer was used to optimise and develop a reliable protocol for sequencing the entire Duffy, ABO and Kidd blood group genes including flanking regions. A DNA library was prepared from randomly selected DNA samples. Duffy, ABO and Kidd genes were targeted by long-range PCR, enzymatic amplicon fragmentation before ligation with barcoded adapters and size selection. Templates were immobilised on beads, then clonally amplified using emulsion PCR. Sequencing revealed millions of reads that were then aligned to the reference gene sequences. Variants were visualised with two software packages, CLC workbench and Integrative Genomics Viewer (IGV). Initial Bioinformatics analysis of Duffy gene samples revealed all genotypes matched phenotype, however a number of SNPs (single nucleotide polymorphisms) were identified in exons 1,2 and intron1 and under investigation. We also identified that a significant number of the FY genes sequenced (5/12) encoded the previously described Ala100Thr mutation. The samples concerned were FY (a-b+) (3 of 4) (1 homozygous and 2 heterozygous), FY (a+b-) (0 of 1) and FY (a+b+) (2/7) both heterozygous). We therefore conclude that the Ala100Thr mutation is much more frequent than previously described. We suggest that NGS will supplant other genotyping platforms in the near future and will potentially become the methodology of choice for genotyping patients and donors.

Halawani, A. J., Altayar, M. A., Kiernan, M., Reynolds, A. J., Kaushik, N., Madgett, T. E. & Avent, N. D. 2014. Can Next-generation DNA Sequencing Solve the RH Complexity for Genotyping? *Vox Sang,* **107,** suppl. 1**,** 57-248.

The Rh blood group system is considered as the most polymorphic blood group system and contains many variants. These variants may cause severe complications to patients, due to alloimmunisation from mismatching products of blood transfusion. Next-generation sequencing (NGS) is an intensely powerful technique capable of sequencing huge regions of the human genome. Here we sequenced the entire RHD and RHCE genes to genotype for the RhD and RhCE antigens, in order to provide a safer transfusion practice. DNA was extracted from random blood donors with known serology. By using long-range PCR (LR-PCR), four and three primer pairs were designed (giving PCR products in the range from 8 to 24 Kb) for RHD and RHCE, respectively. The sequencing libraries were then made by fragmenting the amplicons and ligating to adaptors using Ion Xpress™ Plus Fragment Library Kit. Size selection was performed by SPRIselect magnetic beads. After that, the sequencing template was immobilised to beads, which possess a complementary strand to the adaptors, for clonal amplification using emulsion PCR. Finally, the sequencing template was loaded onto a 316 chip and sequenced on the Ion Torrent Personal Genome Machine™ Sequencer. Millions of reads were generated and the data were analysed with CLC Genomics Workbench (Version 6.5). Sanger sequencing will be utilised in order to validate any variants from the sequencing data. NGS using the LR-PCR approach offers a crucial method assisting users to genotype many samples in a single run for detecting the Rh variants in depth. This will be extremely worthwhile to genotype the difficult alleles of Rh, especially hybrid genes, and will pave the way for the discovery of novel alleles. NGS for blood group alleles may represent a viable alternative to array and bead based platforms, and it is no more expensive and technically challenging on a per sample basis.

Altayar, M. A., **Halawani, A. J.**, Kiernan, M., Reynolds, A. J., Kaushik, N., Madgett, T. E. & Avent, N. D. 2014 Extensive Genotyping of Blood Groups Duffy, Kidd and ABO by Next-generation Sequencing *Vox Sang*, **107,** suppl. 1**,** 57-248.

The determination of the blood groups present in an individual attains a great clinical importance for the purposes of blood transfusion and transplantation. Blood group genotyping (BGG) has become well established in transfusion medicine. However, all current technologies are based on predefined knowledge of known polymorphisms. The approach of Next Generation Sequencing (NGS) circumvents this requirement and adopts a discovery mode, which is important, as almost every new BGG project discovers new alleles. NGS is capable of producing high-throughput, rapid and accurate data that results in extensive and detailed genotyping. Also, costs have considerably reduced in the past years. In this pilot study, Ion Torrent Personal Genome Machine (PGM™) sequencer was used to optimise and develop a reliable protocol for sequencing the entire Duffy (DARC), Kidd (SLC14A1) and ABO blood group genes including flanking regions. First, a DNA library was prepared from 12 randomly selected DNA samples. DARC, SLC14A1 and ABO genes were targeted by long-range PCR, enzymatic amplicon fragmentation before ligation with barcoded adapters and size selection. Templates were immobilised on beads, then clonally amplified using emulsion PCR. Then, up to 20 samples from FY and JK samples were selected with certain serology for sequencing. The reason for this is to seek the possibility that those serologically typed as negative for a particular blood group are actually genotypically weak positive. Sequencing revealed millions of reads with great coverage depth that were then aligned to the reference gene sequences. Variants were analysed and visualised with software packages, Ion Torrent Suite™ plugins, CLC Genomic Workbench, Integrative Genomics Viewer (IGV) and SeattleSeq Annotation 138 website. Initial bioinformatics analysis of samples for the DARC and SLC14A1 genes revealed various single nucleotide polymorphisms (SNPs) in exons encoding for amino acid changes, such as (Gly42Asp and Ala100Thr in DARC) and (Asp280Asn, Ala270Ala and Glu44Lys in SLC14A1). In the FY samples sequenced, a significant number (5/12) encoded the Ala100Thr mutation. In addition, the Allele JK*01W.01 (associated with Jka+w) was found with a frequency of 8%. A great number of polymorphisms (SNPs and Indel) were found in introns in JK samples (ranging from 52 to122) and in FY samples (ranging from 2 to 4). An example of one of these SNPs is chromosomal position 43319274 in the SLC14A1 gene, close to the splice site region of exon 8. Interestingly, sets of intronic polymorphisms present differently among samples with same phenotype. Therefore, the intronic polymorphisms are possibly unique to individuals or families. The ABO sequencing is in progress. We suggest that NGS shows the capability of the comprehensive sequencing of blood group genes and will supplant other genotyping platforms in the near future, becoming the potential methodology of choice for genotyping patients and donors.

Avent, N. D., Madgett, T. E., **Halawani, A. J.**, Altayar, M. A., Kiernan, M. & Reynolds, A. J. 2014. Next Generation Sequencing: Academic Overkill or High-resolution Blood Group Genotyping? . *Vox Sang,* **107,** supple. 1**,** 1-56.

Background: Next generation sequencing (NGS) has emerged as a high throughput method to rapidly determine the genotype of an individual without prior knowledge of single nucleotide polymorphism sequences that govern the probe or primer sequences associated with many conventional genotyping systems. For blood group genotyping (BGG) array or bead based platforms have emerged as commercially available high resolution systems that are in routine use. However, they do not operate in discovery mode, and still many BGG studies are revealing new (and unknown) alleles. NGS may have greater capacity to deal with complex hybrid genes that are associated with ABO, RH and MNS polymorphic variation.

Aims: We aimed to develop a cheap viable alternative to bead/array based genotyping with much higher resolution being based on a commercially available NGS plat- form, the Ion Torrent PGM.

Methods: Using long-range PCR we have amplified the complete genomic sequences of ABO, RHCE, RHD, KEL, FY and JK with many other blood group genes also under development. These PCR products are fragmented, ligated to barcoded adaptors and sequenced on the Ion Torrent PGM machine using an appropriate chip.

Results: A high resolution (500–20009 coverage) NGS approach to blood group genotyping has been developed and complete sequences of blood group active genes can be rapidly sequenced and known and unknown blood group alleles have been defined quickly and cheaply. Bioinformatic approaches to filter data for known alleles have been done, but a considerable level of complexity has emerged especially in the intronic regions of KEL, RH and JK genes. We have defined the allele frequency of some JK*A (weak) is much higher than previously published and we are currently sequencing an extended cohort of phenotyped blood samples to assess the frequency of a number of these observed blood group alleles.

Summary/Conclusions: Next generation sequencing based blood group genotyping has significant value in the academic environment for unravelling the complex evolution of human blood group alleles. A level of complexity observed with intronic SNPs will make zygosity assignment (e.g. for RHD) extremely simple, and will help confirm the presence of rare mutations in many samples. NGS based genotyping is rapidly reducing in costs and the procedures and data interpretation is relatively simple for any competent molecular diagnostics laboratory. Whilst undoubtedly NGS is of great academic interest (as is true in our laboratory) the evolution into a system of complete utility where all data regarding a blood donor is stored as DNA sequence will undoubtedly replace current array based BGG systems. Bioinformatic assessment of NGS data can be simplified to extraction of data on a ‘as required’ basis. Key information regarding clinically significant blood group alleles are easy to process, and this is coupled with the added benefit of having the complete sequence of donor or patient available which can be interrogated on the discovery of a new blood group allele.

Halawani, A. J., Altayar, M. A., Kiernan, M., Li, X., Madgett, T. E. & Avent, N. D. 2015. High Resolution Genotyping of the Rh Blood Group System by Next-generation Sequencing *Vox Sang*, **109,** suppl. 1**,** 1-379.

The RH blood group system is the most complicated blood group system due to encoding by two highly homologous genes, RHD and RHCE. Typing the antigens of this system by conventional serology is not very appropriate to distinguish between D positive, weak D and partial D. It is only possible to assign weak D and partial D alleles accurately using blood group genotyping (BGG). Here, 10 samples of the RH system were genotyped, 5 D positive and 5 weak D samples using a long-range PCR (LR-PCR) approach coupled with next generation sequencing (NGS). For every sample, both genes, RHD and RHCE, were amplified by LR-PCR, with three amplicons for RHD and four amplicons for RHCE. Then the PCR products were fragmented, ligated to barcoded adaptors and sequenced using NGS on an Ion Torrent PGM^{TM} platform. We showed that LR-PCR for RHD and RHCE completely correlated with their corresponding genomic sequence. For the D positive samples, there were no obvious SNPs on the RHD exons. The 5 weak D samples have been identified as following; two weak D Type 1 (exon 6 809T>G Val270Gly), two weak D Type 2 (exon 9 1154G>C Gly385Ala) and one weak partial D 4.1 with [exon 1 48G>C (Trp16Cys), exon 4 602C>G (Thr201Arg), exon 5 667T>G (Phe223Val), exon 6 819G>A (silent)]. The LR-PCR method has confirmed that a novel heterozygous SNP, 208 C>T (Arg70Trp) in exon 2 is derived from the RHCE gene, although it had previously been identified by a Human Erythrocyte Antigen and Human Platelet Antigen panel as belonging to the RHD gene. More samples are currently being sequenced. Our approach, we believe, will facilitate the comprehensive genotyping of the antigens of the RH system, especially those with hybrid genes or insertions/deletions. Our method is able to demonstrate novel alleles by direct sequence analysis, a major drawback of current array-based BGG platforms.

Altayar, M. A., **Halawani, A. J.**, Kiernan, M., Madgett, T. E. & Avent, N. D. 2015. Complete Gene Sequencing of ABO Blood Group by Next-generation Sequencing *Vox Sang,* **109,** suppl. 1**,** 1-379.

The ABO blood group system is the most clinical significant in blood transfusion and transplantation medicine. Due to naturally occurring antibodies, mismatched transfusion of blood can cause rapid transfusion reactions. ABO is one of the most complex and polymorphic blood group genes, with an ever-increasing number of variant alleles. These variant alleles not only affect the specificity of the enzymes but also the activity of the enzymes, which might result in a weak phenotype. Therefore the determination of ABO alleles is important for the safety of blood transfusion and transplantation medicine. Although high-throughput platforms have revolutionised the approach towards blood group genotyping (BGG), they are based on pre-defined polymorphisms, which are not suitable for the discovery of new alleles. Next Generation Sequencing (NGS) circumvents this requirement and operates in discovery mode, which is critical for emerging alleles. NGS is capable of producing comprehensive, high-throughput, rapid and accurate data resulting in extensive genotyping. ABO genotyping has frequently only focused on exons 6 and 7, neglecting the rest of the gene. Following our successful NGS-genotyping of blood group genes Duffy (DARC) and Kidd (SLC14A1), here we have used the Ion Torrent Personal Genome Machine™ (PGM™) sequencer to optimise and develop a reliable protocol for sequencing the entire ABO blood group gene including flanking regions. In this pilot study, four long-range polymerase chain reactions (LR-PCR) were used to target the entire ABO gene plus over 5 kb upstream, regulatory regions (Promoter and CBF/NF-Y), and downstream in 16 randomly selected genomic DNA samples. DNA libraries were prepared by enzymatic fragmentation, ligation of barcoded adapters and size selection, before clonal amplification of templates was achieved using emulsion PCR and then the samples were loaded onto a 316 chip for sequencing. Millions of reads with great coverage depth (100–1800x) were generated, which were then aligned to the reference gene sequence (NG_006669.1). These data were analysed and visualised with multiple software packages, such as CLC Genomic Work- bench version 6.5. The serological phenotype data matched that of ABO genotyping. Bioinformatics analysis revealed a number of polymorphisms including single nucleotide polymorphisms (SNPs) and insertion/deletions (indels) distributed throughout exons, introns and the regulatory regions at around 4 kb upstream. Examples of amino acid changes due to the SNPs found in exons are those causing the differences between A and B alleles, previously described in exon 7 (Arg176Gly, Gly235Ser, Leu266Met and Gly268Ala), whilst other SNPs found in exons 3, 4 and 5 have been found to be of higher frequency in our samples than previously reported, including Arg63His (13/16 samples) and Ser74Pro (14/16 samples) and found in all ABO phenotypes. In addition, in two samples (of A and O phenotype), we showed the Trp181stop mutation, previously described only for the rare ABO*O.06 (O^6) allele. We suggest that NGS can provide a reliable approach to genotype ABO due to its powerful capabilities of comprehensive analysis and revealing novel alleles. NGS will supplant other genotyping platforms in the near future, becoming the potential methodology of choice for genotyping patients and donors for safe transfusion/ transplantation practice.

Appendix B

DNA concentrations and purity

Cohort A: Samples for Microarray assay (Chapter 3)

Sample ID	NanoVue	Purity	Qubit
1	93.6	1.95	111
2	239.6	1.95	256
3	11	1.94	50
4	540.9	1.90	230
5	289.8	1.92	233
6	252.2	1.93	301
7	146.8	1.93	356
8	103.7	1.92	504
9	84.2	1.94	200
10	55	1.95	140
11	26.4	1.95	90
12	87.4	1.95	80.5
13	232.5	1.87	233
14	116.8	1.92	103.6
15	147.8	1.92	344
16	96.2	1.90	423
17	107.1	1.96	256
18	75.6	1.90	226
19	78.6	1.97	202
20	122.6	1.88	235
21	111.7	1.90	322
22	60.1	1.88	109
23	136.4	1.94	190
24	88.9	1.90	112
25	260.6	1.94	270
26	175.7	1.89	118
27	207.4	1.95	119.4
28	322.2	1.94	510

Cohort B: Samples for HEA and HPA Panel (Chapter 4)

Sample ID	NanoVue	Purity	Qubit
1	83.5	1.95	102
2	104.6	1.95	99
3	4.7	1.75	50
4	129	1.95	90.4
5	81.9	1.93	112
6	122.9	1.94	234
7	77.1	1.94	224
8	160.5	1.92	340
9	147.4	1.94	120
10	303	1.95	543
11	212.5	1.91	433
12	28	1.87	67
13	45.5	1.70	88
14	45.5	1.72	91
15	89.5	1.83	88
16	35	1.51	55
17	142	1.82	241
18	56	1.75	40
19	297	1.84	441
20	44.5	1.78	98
21	506	1.82	236
22	239	1.87	340
23	231	1.99	356
24	281	1.92	200
25	181.5	1.85	180
26	140.5	1.82	340
27	79	1.71	101
28	120.5	1.84	140

Cohort C: Samples for Kell LR-PCR (Chapter 5)

Sample ID	NanoVue	Purity	Qubit
1	555	1.70	301
2	340	1.95	139
3	544	1.73	209
4	270	1.70	261
5	240	1.94	233
6	80	1.80	174
7	340	1.84	251
8	300	1.85	378
9	540	1.82	399
10	560	1.89	272
11	798	1.92	329
12	270	1.95	235
13	450	1.95	268
14	270	1.70	269
15	290	1.90	305
16	260	1.90	289
17	230	1.94	129
18	430	1.95	311
19	340	1.70	179
20	220	1.70	280

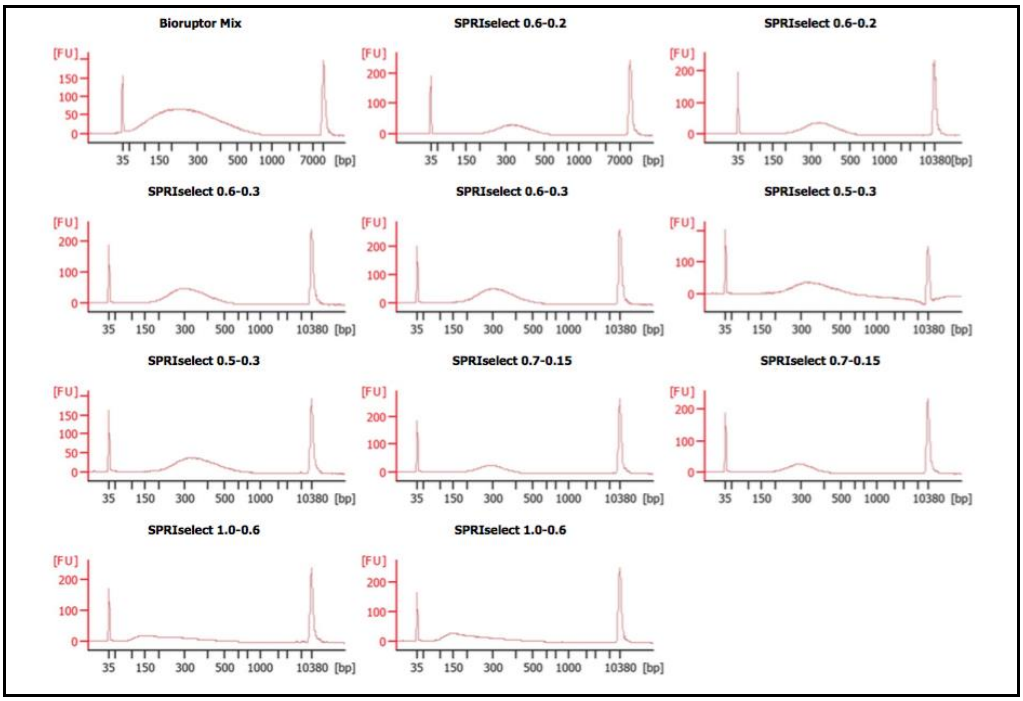
Cohort D: Samples for Rh LR-PCR (Chapter 6)

Sample ID	NanoVue	Purity	Qubit
1	301	1.95	340
2	230	1.95	115
3	222	1.89	138
4	332	1.87	267
5	240	1.88	121
6	255	1.89	209
7	111	1.95	68.5
8	229	1.70	123
9	232	1.89	89.9
10	243	1.77	108

Appendix C

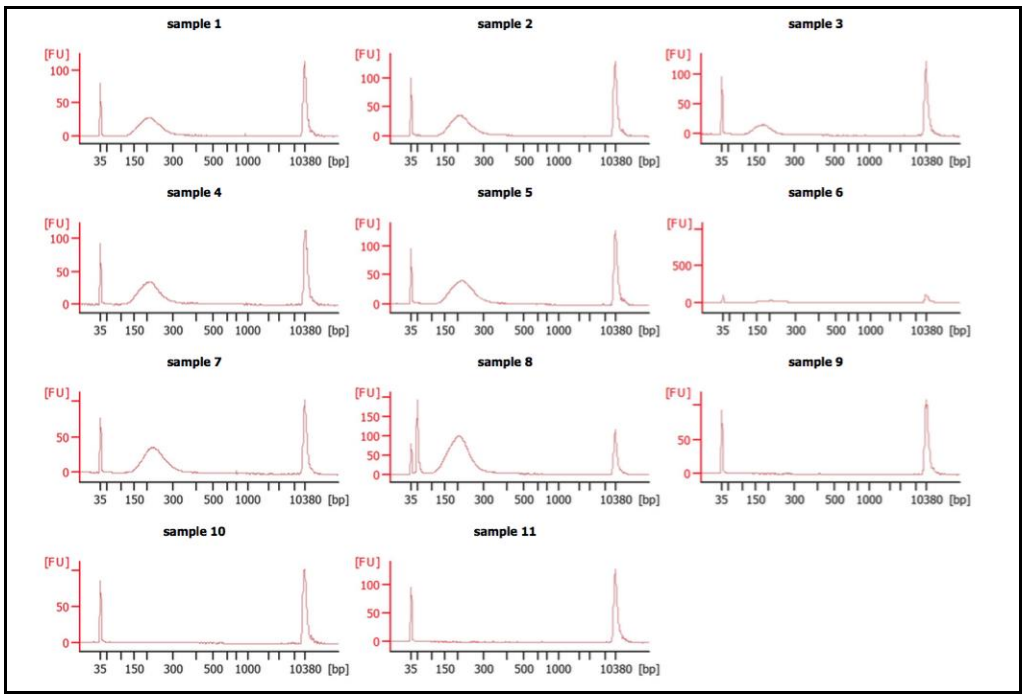
Optimisation of the size selection for NGS libraries

(a) Different SPRIselect beads ratio



Bioruptor mix was the sample, which had fragmentation by the physical sonication using Bioruptor™ UCD-200.

(b) 0.8X-0.7X ratio of SPRIselect



Sample 9,10 and 11 were just controls, which had no DNA.