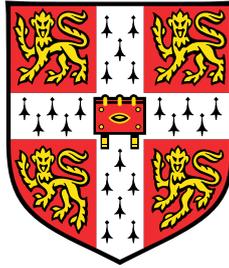


Blood Donor Genotyping



Nicholas S. Gleadall

Department of Haematology
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Gonville & Caius College

September 2020

I would like to dedicate this thesis to the 813,212 donors, patients and healthy individuals whose data this work has used. Their decision to share personal medical and genetic information so that it may be used for the benefit of others is laudable. I hope that the work presented here meets the expectations they had when making such a valuable donation.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Nicholas S. Gleadall
September 2020

Blood Donor Genotyping

Nicholas S. Gleadall

Abstract

Transfusion of blood is one of the oldest and most widely used clinical interventions. In 2020 the World Health Organisation reported that globally 118.5 million blood donations had been collected worldwide. This blood will be used to provide life-saving transfusion support for millions of individuals with a wide range of medical conditions.

To ensure the safety of each blood transfusion it is common policy to identify and ensure compatibility between the ABO and RhD antigens of both donor and recipient. Although this policy prevents the majority of adverse haemolytic transfusion reaction's (HTR), approximately 3% of recipients become sensitised following an immune reaction to a non-self blood group antigen after a single transfusion episode. This proportion can rise dramatically in patients requiring frequent transfusion support, with immunisation rates as high as 60% being reported for some haemoglobinopathy patients.

Sensitisation to non-self red blood cell (RBC) antigens confers a lifetime risk of HTRs, which from 2013 through 2017 were responsible for 17% (32 of 185) and 6% (7 of 110) of transfusion-related deaths reported to the US Food and Drug Administration and Serious Hazards of Transfusion UK, respectively. Furthermore, sensitisation can render transfusion-dependent patients non-transfusable and cause haemolytic disease in pregnancy which is potentially life-threatening to the fetus. A more precise blood matching policy will reduce sensitisation rates, however, adoption of this is resisted because of perceived logistical challenges, donor typing costs, and the lack of evidence from large scale clinical trials that reducing sensitisation rates results in health gains.

Antibody-based haemagglutination tests are the current gold standard for RBC antigen typing; however, reliable reagents and high-throughput techniques are not available for all clinically relevant antigens. DNA-based tests have been used to overcome these limitations, and a range of in-house and commercial assays have been developed for donor genotyping. However, due to the limited number of antigens typed for and the low throughput capacity of these assays, they have not been widely applied to blood donor typing.

Advances in the technologies used for genome-wide genotyping and sequencing have substantially reduced the cost of generating genetic variation data at population scale. Multiple studies have demonstrated that it is possible to extract antigen typing information from the data produced by these technologies, indicating that they could be used to deliver

genomics-based precision transfusion medicine to the patient bedside. However, the same studies also highlight a series of challenges that would have to be overcome before these technologies could be safely integrated into the clinical laboratory.

In this thesis, I will present the work that has been done to overcome some of the issues surrounding the interpretation of blood cell antigen typing from genomic data and the development of a universal donor genotyping platform which can be used by blood supply organisations worldwide to implement a policy of precision transfusion medicine.

Acknowledgements

The work presented in this thesis has been made possible through the continued support of many individuals.

Firstly, I would like to thank Karen, Stephen and Sam - my family - for supporting me throughout my studies. I would like to thank all friends, loungers, crewmates, and ghost house residents who have kept me entertained throughout my PhD, I have been incredibly fortunate to meet you all.

Secondly, from the NIHR BioResource, INTERVAL, COMPARE and DIS-III study teams I would like to thank Emanuele Di Angelantonio, John Bradley, Sarah Fahle, John M Jongerius, Nathalie Kingston, Jyoti Khadake, Amy McMahon, Susan Mehenny, Carmel Moore, Jennifer Sambrook, Olga Shamardina, Kathy Stirrups, Tiffany C. Timmer, Katja van den Hurk, Matt Walker, and Neil Walker for the genotyping and patient data they provided that was so crucial to this study.

Thirdly, from the ISBT, NHS Blood and Transplant, and Sanquin I would like to thank Christoph Gassner, Alan Grey, Shane Grimsley, Peter Ligthart, and Nicole Thornton for taking the time to give me an education in Immunohaematology and Judith Chapman, Gail Mifflin, Dave Roberts, and Jill Storry who have publicly supported this work and helped to bring it to the attention of the wider transfusion medicine community.

A special thanks must go to William Astle, Kim Brugger, Karyn Megy, and Christopher Penkett who have done me hundreds of favours and put up with continual questions for four years. I would also like to give special thanks to Debra Fletcher and Lindsay Walker who have without fail managed to keep me organised, on-time, and generally alive throughout.

From the Blood transfusion Genomics Consortium and my extensive panel of co-supervisors, I would like to thank Adam Butterworth, Jeremy Gollub, William Lane, John Ord, Ellen van der Schoot, Barbera Veldhuisen, Nicholas Watkins, and Connie Westhoff who have committed countless hours of their time to this project and to my professional development.

Last but certainly not least, I would like to thank Willem H Ouwehand who has been consistently supportive, generous with his time and ideas, and an inspiration throughout. Being his student has been a privilege and a joy, I could not have asked for a better supervisor.

Because of him, I have had the opportunity to complete 1 PhD, visit 9 countries, give 32 presentations, win 6 prizes, receive 1 Lancet rejection, receive 1 Blood Advances acceptance, meet 1 billionaire, climb in the Rocky Mountains, sail on Lake Ontario, save William Lane from drowning in the Rhine, and make no less than 25 Circos plots. I will never be able to thank him enough for his decision to give me an opportunity to study for my PhD at the University of Cambridge.

Table of contents

List of figures	xv
List of tables	xvii
Nomenclature	xix
1 Introduction	1
1.1 A <i>brief</i> history of blood transfusion	1
1.2 Blood group antigens	3
1.3 Pre-transfusion testing and donor screening	4
1.4 The molecular basis of blood group antigens	7
1.4.1 The ABO blood group system	7
1.4.2 The Kell blood group system	11
1.5 Donor Genotyping	11
1.6 Population Scale Genotyping	13
1.7 Aims	17
2 Materials and Methods	19
2.1 Overview	19
2.2 An overview of Axiom array genotyping	22
2.3 Next Generation Sequencing	26
2.4 Bioinformatics	30
2.4.1 bloodTyper	30
2.4.2 HLA impute	32
2.4.3 BWA, SAMtools, and BCFtools	34
2.4.4 Genome Analysis Toolkit, Manta and Canvas	34
2.4.5 Axiom Analysis Suite	34
2.5 Summary of external data sources	35
2.6 Blood Donor DNA samples	35

2.6.1	INTERVAL panel (n=1,257)	35
2.6.2	NIHR BioResource panel (n=507)	36
2.6.3	COMPARE and DIS3 panel (n=7,477)	36
2.7	Clinical RBC and PLT antigen typing data	37
2.8	RBC and HPA Antigen typing concordance analysis	39
2.9	HLA antigen typing concordance analysis	39
2.10	Data availability	40
2.11	My contribution to each chapter	40
3	Mapping blood group antigens to the human genome	43
3.1	Abstract	43
3.2	Introduction	44
3.3	Chapter Workflow	46
3.4	The Blood transfusion Genomics Consortium	48
3.5	Data driven review of ISBT reference transcripts	49
3.6	Exploring whole-genome sequencing data	53
3.6.1	Further validation of the bloodTyper WGS algorithm	53
3.6.2	Antigen typing 13,037 whole genomes	54
3.6.3	Variation in the Kell blood group system	60
3.7	Discussion	65
4	A universal donor genotyping platform	69
4.1	Abstract	69
4.2	Introduction	70
4.3	Chapter workflow	72
4.4	Stage 1 - Proof of principle	74
4.5	Stage 2 - Designing an array for donor genotyping	77
4.6	Stage 3: Large scale blinded trial of the UKBBv2 array	83
4.6.1	Selection of trial set samples and initial genotyping quality control	86
4.6.2	Allele frequency validation of the UKBBv2 array	90
4.6.3	Overall antigen typing performance	92
4.6.4	RBC antigen typing performance	93
4.6.5	HPA antigen typing	93
4.6.6	HLA antigen typing	93
4.7	Impact of an extensively genotyped donor panel	94
4.7.1	Identification of individuals with rare phenotypes	96
4.7.2	Immediate clinical benefit of extended typing data	96

4.7.3	Donor and patient health information	98
4.8	Discussion	98
5	Typing genetically complex antigens	103
5.1	Abstract	103
5.2	Introduction	104
5.3	Chapter workflow	109
5.4	Assay design	111
5.5	Validation of BGC targeted sequencing assay	113
5.5.1	Comparison to WGS data	113
5.5.2	Comparison to clinical antibody typing	116
5.6	Investigation of discordant array samples	120
5.6.1	Algorithmic discordances	121
5.6.2	Array discordances	122
5.6.3	Error in donor record antigen typing data	123
5.6.4	Unresolved discordances	123
5.7	Discussion	124
6	Discussion	127
6.1	Translating this work	130
	References	133

List of figures

1.1	Types of blood group antigen	3
1.2	Clinical significance of each blood group	5
1.3	Biosynthesis of ABO group antigens	8
1.4	Molecular configuration of common ABO antigens	10
1.5	A basic ABO genotyping algorithm	11
1.6	gnomAD variation in coding space of RBC antigen genes	15
2.1	Thesis overview	21
2.2	Overview of Axiom array genotyping	23
2.3	Example genotype call plots	25
2.4	Overview of Illumina NGS sequencing	27
2.5	NGS sequence alignment at <i>SMIMI</i> locus	28
2.6	Overview of the bloodTyper analysis pipeline	31
2.7	Overview of the HLA*IMP:02 model	33
3.1	Chapter 3 workflow	47
3.2	Membership of the Blood transfusion Genomics Consortium	48
3.3	BLUEPRINT expression data for the <i>KEL</i> gene	50
3.4	INTS genotyping report of Rh _{null} patient	56
3.5	WGS sequence alignment at <i>RHAG</i> locus in Rh _{null} patient	57
3.6	Pedigree of Rh _{null} patient	59
3.7	Importance of phasing	61
3.8	ISBT <i>KEL</i> haplotypes identified in the WGS data of 13,037 individuals	63
3.9	Complete variation in the <i>KEL</i> gene of 13,037 individuals	64
4.1	Chapter 4 workflow	73
4.2	Original UK Biobank Array antigen typing performance	75
4.3	384 Blood Typing Array antigen typing performance	79
4.4	Copy number aware genotyping of Rh variants	82

4.5	Genetic variation in antigen controlling genes compared to array content . . .	85
4.6	Quality assessment of UKBBv2 array using HapMap samples	87
4.7	Comparison of European ancestry allele frequencies measured by the UKBBv2 array and Illumina short read whole genome sequencing (WGS)	91
4.8	Concordance between clinical and UKBBv2 array Red Cell and Platelet antigen typing	92
4.9	Concordance between clinical and UKBBv2 array HLA antigen typing results	94
4.10	Comparison of clinical and UKBBv2 array antigen typing availability . . .	95
4.11	UKBBv2 array genotyping significantly increases the number of available units for patients with multiple RBC antibodies	97
5.1	Loss of genome structure information in the array genotyping process . . .	105
5.2	Coverage of the <i>ABO</i> gene by different sequencing technologies	108
5.3	Chapter 5 workflow	110
5.4	Sequencing coverage of antigen encoding target regions for BGC capture assay	114
5.5	Concordance between BGC capture assay inferred and donor record antigen types	117
5.6	Differences in read depth profile for different C/c antigen phenotypes	118
5.7	Screenshot of the BGC capture bloodTyper report for a Jk ^b discordant sample	119
5.8	Breakdown of UKBBv2 discordant trial samples	121

List of tables

2.1	Data sources	35
2.2	Antigen typing data available in electronic donor record	38
3.1	ISBT transcripts compared with BLUEPRINT most expressed transcripts .	51
3.2	Patients in the NIHR BioResource WGS project with rare antigen negative phenotypes	55
4.1	Antigen typing variants for which no working probeset could be identified .	78
4.2	Genes relevant to antigen expression selected for coding variant enrichment on UKBBv2 array	84
4.3	Trial set samples removed from further analysis	89
4.4	Examples of rare blood group phenotypes and number of homozygous individuals identified	97
5.1	ISBT alleles which cannot be typed using current array technologies	107
5.2	BGC capture target genes	112
5.3	Genotype to genotype comparison between BGC capture and WGS data . .	116

Nomenclature

Roman Symbols

100KGP 100,000 genomes project

1KGP 1000 genomes project

384 Blood Typing Array 384 High Throughput Axiom Blood Typing Single Nucleotide Polymorphism Screen Array

BGC Blood Transfusion Genomics Consortium

EBV Epstein Barr Virus

GATK Genome Analysis Toolkit

GWAs Genome Wide Association Study

indel Short insertion or deletion variant usually under 50 base pairs

ISBT International Society of Blood Transfusion

ISBT WP ISBT Working Party for Red Cell Immunogenetics and Blood Group Terminology

LD linkage disequilibrium

LRG Locus Reference Genomic

MAF minor allele frequency

MALDI-TOF Matrix-assisted laser desorption/ionization time of flight spectrometry

NGS Next Generation Sequencing

NHSBT NHS Blood and Transplant

PCR Polymerase Chain Reaction

PCR-SSP PCR with Sequence Specific Primers

PLT Platelet

RBC Red Blood Cell

RFLP Restriction Fragment Length Polymorphism

ROI Region of interest

RT-PCR Real Time PCR

SNP Single Nucleotide Polymorphism

SV Structural variation and/or genetic rearrangement

TFS Thermo Fisher Scientific

UKBB UK BioBank

UKBBv1 array Original UK Biobank Axiom Array

UKBBv2 array Version 2 UK Biobank Axiom Array - validated for blood typing

WES Whole Exome Sequencing

WGS Whole Genome Sequencing

WTCCC Wellcome Trust Case Control Consortium

Chapter 1

Introduction

1.1 A *brief* history of blood transfusion

In 1825 James Blundell, an obstetrician at Guys and St. Thomas' Hospital in London, infused a woman suffering severe postpartum haemorrhage with four ounces of her husband's blood. Amazingly she survived, and thus the first successful human-to-human blood transfusion was documented. In the following years, Blundell repeated the procedure with varying degrees of success leading him to acknowledge that there were serious risks associated with "the operation", and that with the current level of knowledge it seemed right to "confine transfusion to the first class of cases only, namely those in which there seems to be no hope for the patient unless blood can be thrown into the veins".[1] Due to unpredictable and unsafe results, blood transfusions were generally avoided for the remainder of the 19th century.

Things began to change in 1901 when Karl Landsteiner discovered the ABO blood group by mixing sera and red cells from 22 different individuals.[2] He observed that some combinations would agglutinate, and deduced a pattern which allowed his subjects to be divided into three groups; A, B and C (later O). A year later his students performed a follow-up study with 155 individuals, confirming the earlier findings and also identifying four individuals who did not fall into any of the three groups, later these people were classified as group AB. Landsteiner had identified that most people had "naturally occurring" antibodies in their sera which could react to antigens present on the red cells of others, causing the donor red cells to be broken down once transfused (haemolysis).

Although Landsteiner's discovery explains why early transfusions had been so unpredictable, it was not until 1907 that an American surgeon, Reuben Ottenberg, suggested that it might be a good idea if both donor and recipient were ABO grouped before transfusion and their blood mixed in the laboratory (cross-matched) to ensure compatibility. Pre-transfusion

blood group testing dramatically increased the safety of human blood transfusions and as a result, they became steadily more popular in clinical practice.

Despite matching for the ABO blood group system, adverse reactions following blood transfusions were still reported for some patients and several individuals continued the search for undiscovered blood group antigens using antibody-mediated agglutination. In 1927 Landsteiner and Levine discovered the M, N and P blood group antigens in rabbits immunised with human red cells, and in 1939 Landsteiner and Wiener identified the Rhesus antigen (now known as D) by injecting rhesus monkey red cells into rabbits and guinea pigs.[3, 4]

The early search for human blood groups was impeded by the fact that only antibodies of the IgM class which directly agglutinated red cells could be studied. The barrier to discovery was overcome in 1945 by Coombs, Mourant, and Race who developed the indirect anti-globulin test.[5] This new method, based on the property of anti-human antibodies to bind human antibodies, enabled the detection of antibodies of the IgG class which do not directly agglutinate red cells. This important development led to an explosion in the discovery rate of new blood groups and has underpinned the field of transfusion medicine to date.

1.2 Blood group antigens

The term 'blood group' usually refers to blood cell surface antigens, predominantly those on the surface of red blood cells (RBC). Antigens can be proteins, glycoproteins or glycolipids, which form part of the red cell membrane (see Fig. 1.1). They have numerous functions such as; membrane transporters (Diego, Kidd), receptor and adhesion molecules (Duffy, Lutheran), complement regulatory glycoproteins (Cromer, Knops), enzymes (Yt, Kell, Dombrock), structural components (Diego, Gerbich) or components of the glycocalyx (MNS).

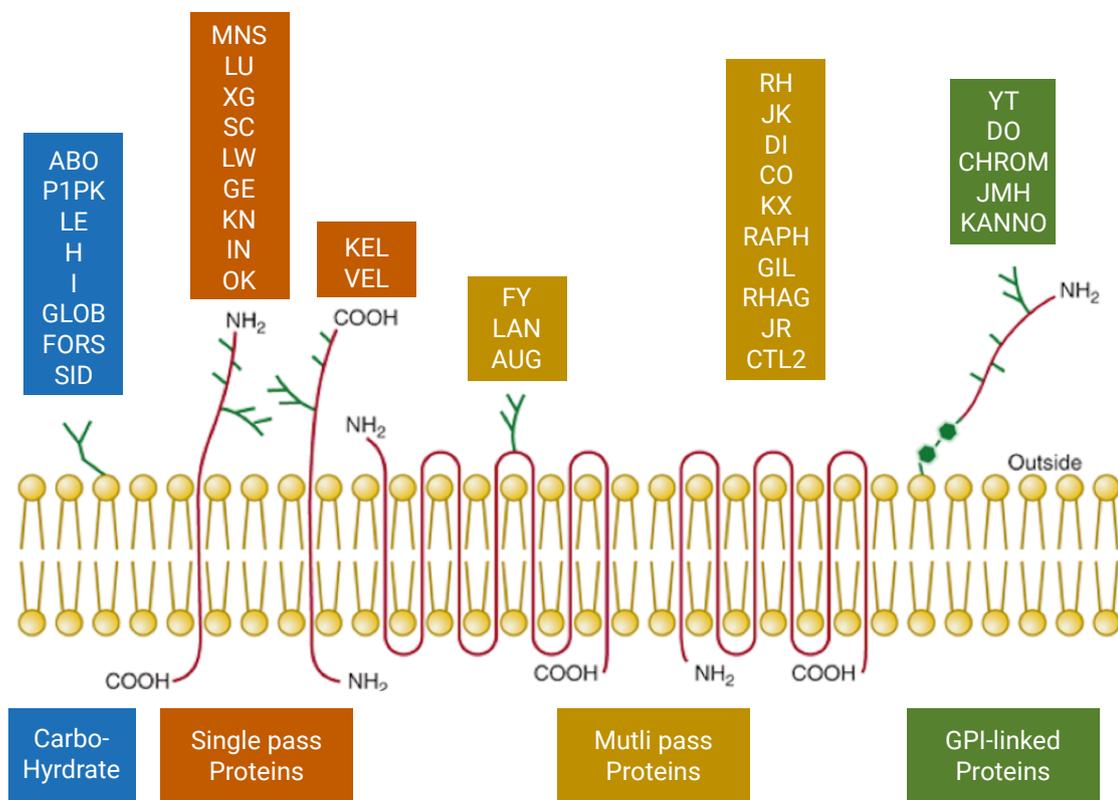


Fig. 1.1 Cartoon showing the different types of red cell antigens. The erythrocyte membrane is represented in the centre of the image, above this represents the outside of the cell and below inside. Figure modified from [6].

It is the presence and absence due to inherited variation of red cell surface antigens that defines the blood group of an individual. Antigens are defined by antibodies that occur either 'naturally' due to encountering antigens ubiquitous in the environment (e.g. anti-A which is specific to the A antigen) or are formed as a result of active immunisation to non-self

RBC antigens following exposure to human RBCs from another individual (e.g. formation of anti-D, which is specific to the D antigen, if a D-negative patient is transfused with blood from a D-positive donor).

Blood group systems are officially defined as 'systems of one or more antigens governed by a single gene or complex of two or more closely linked homologous genes with virtually no recombination between them'. Each system is genetically discrete from every other blood group system. In order for a blood group system to be established the underlying gene must be identified and sequenced. The International Society of Blood Transfusion (ISBT) Working Party for Red Cell Immunogenetics and Blood Group Terminology (ISBT WP) maintains an official record of all currently recognised blood group systems. There are currently 36 blood group systems containing 326 red cell antigens.[7]

The ISBT also maintain three categories for antigens that have not yet been linked to blood group systems. Collections were designed to group antigens that are biochemically, genetically or serologically similar where the genetic basis has not yet been discovered; there are currently 5 collections containing 14 antigens. There are also two antigen series; the 700 series contains antigens that do not fit into any system or collection which have an incidence of <1% across all human ethnic populations, and the 901 series contains antigens that have a frequency >99% across populations of different ethnic ancestry. There are currently 16 and 4 antigens in the 700 and 901 series, respectively.[7]

The work in this thesis will predominately focus on the RBC antigens in blood group systems.

1.3 Pre-transfusion testing and donor screening

To ensure the safety of a blood transfusion it is important to accurately identify the blood groups of both recipient and donor. Not all blood group antigens stimulate clinically significant antibodies.(see Fig. 1.2).

For general purposes there are three requirements for the prevention of alloimmunisation following RBC transfusion; 1) The RBCs of the donor must be ABO compatible, 2) RBCs from a D-positive donor should ideally not be given to individuals who are D-negative and must not be given to D-negative women of child-bearing age, and 3) the RBCs should be obtained from a donor who lacks the antigens that the recipient has been previously sensitised to. Patients who are transfusion dependant, such as those with Haemoglobinopathies, require more careful blood matching as the likeliness of an alloimmunisation event is increased because of the repeated transfusions.

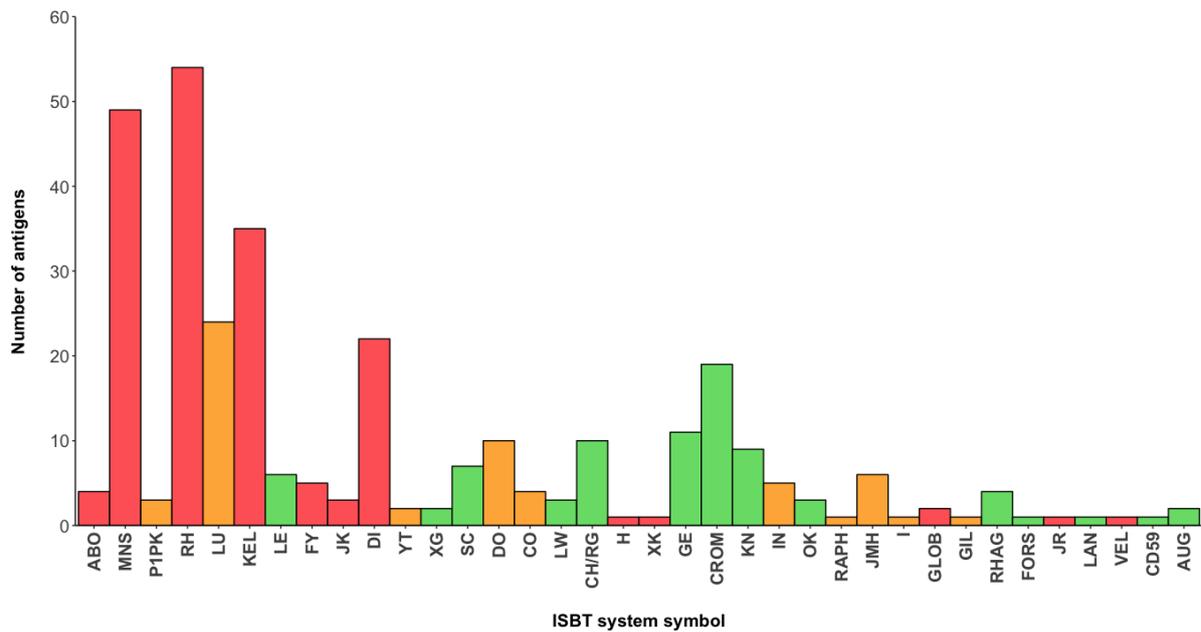


Fig. 1.2 Clinical significance of each blood group. The number of antigens in each system is shown. Bars are coloured to indicate the incidence of haemolytic transfusion reactions and haemolytic disease of the foetus and newborn for each system. Blood group systems with antigens that frequently stimulate clinically important alloantibodies are shown in red, those which occasionally stimulate clinically important alloantibodies are shown in orange, those which rarely stimulate clinically important alloantibodies are shown in green. Figure modified from [6].

Typing by antibody-mediated haemagglutination is the current gold standard approach for donors and patients; this method is also referred to as 'typing by serology'. These tests rely on antibody-mediated agglutination of RBCs and have been the foundation of pre-transfusion testing and donor typing since the discovery of human blood groups in 1901. Early testing was limited to antigens for which the cognate antibodies were of the IgM type and could directly agglutinate RBCs. In 1945 the field was revolutionised following the discovery that anti-human antibodies could be used to agglutinate RBCs which were sensitised with IgG antibodies against a RBC antigen and thereby cause indirect agglutination. The indirect antiglobulin test (also known as the Coombs test) allowed a large number of RBC antigens to be identified and new blood group systems to be defined.[5]

Early antibody typing reagents were of polyclonal origin, either obtained from patients sensitised by transfusion or during pregnancy. Such human antisera contain many different antibody specificities and such 'contaminating antibodies' were frequently the cause of erroneous false-positive antigen typing results. Furthermore, there was no batch-to-batch consistency for this generation of polyclonal typing reagents. In addition, antisera frequently contained anti-HLA antibodies and the RBCs of a fraction of donors do express HLA class I antigens, particularly from the HLA-B locus.

In 1975 a second revolution hit the field, and the world, when Köhler and Milstein published a method for obtaining murine monoclonal antibodies by fusing lymphocytes from the spleens of specifically immunised mice with mouse myeloma cells.[8] The method was used to produce murine monoclonal typing reagents that allowed specific typing for a wide range of blood group antigens (anti-A, -B, -P, P₁, -P^k, -I, -Le^a, -Le^b, -M, N, -k). The first human monoclonal antibodies with blood group specificity were produced in 1980 by the transformation of lymphocytes from D-immunised donors with Epstein Barr Virus (EBV) into antigen secreting lymphoblastoid cells.[9, 10] EBV-transformed cell lines proved too unstable for reliable industrial production of monoclonal anti-D, and the method was replaced by a technique in which human lymphocytes are fused with Murine myelomas to produce stable heteromyeloma cell lines.[11, 12] The latter method was used to produce many of the diagnostic monoclonal typing antibodies still in use today (anti-D, -C, -c, -E, -e, G, K, Jk^a, Jk^b, H, -Le^a, -Le^b, -IgG, -C3d).

Currently, automated antibody typing systems such as the Olympus PK7200 are used to type the RBC antigens of donors in major systems such as ABO and RH. Typing for less common and rare antigens is done via manual 96-well plate based antibody haemagglutination and antisera titration methods. Hospital blood banks predominantly use medium-throughput methods for ABO and D typing of patients such as Gel-cards.

There are several situations where serological typing of RBC antigens is impractical, unreliable or unavailable. As discussed, serological typing relies either on the availability of typing antibodies or human antisera. For some RBC antigens, these are simply not available or are not reliable, for example, there is no existing antibody for accurate Dombrock antigen typing. Finally, following transfusion mixed field haemagglutination may prevent accurate typing. Patients with warm-type RBC autoantibodies and a positive direct antiglobulin test cannot be reliably typed. Likewise, the administration of therapeutic monoclonal antibodies for the treatment of myeloma renders the indirect antiglobulin test unsuitable for pre-transfusion antibody detection.

It is for these reasons that immunohematologists have turned to the knowledge of blood group genetics to develop DNA based antigen genotyping tests.

1.4 The molecular basis of blood group antigens

Blood group antigen systems can be divided into two categories; 1) those where the primary product of the system gene(s) are glycosyltransferase enzymes which have a role in attachment of carbohydrates to existing RBC membrane structures (e.g. ABO) and 2) those where the antigen is the primary protein product of the system gene(s) (e.g. Kel).[6] Expression of blood group antigens can be controlled by single nucleotide polymorphisms (SNPs), small insertions or deletions (indels), larger-scale structural variations and genetic rearrangements (SVs), or combinations of these.[13, 14] It is easiest to understand the concepts of antigen expression with examples.

1.4.1 The ABO blood group system

The A and B antigens of the ABO system are carbohydrate determinants. Their expression is controlled by a combination of SNPs in the *ABO* gene of an individual which encodes a glycosyltransferase enzyme. They are distinguished from one another by an immunodominant terminal monosaccharide attached to a carbohydrate carrier structure, with N-acetylgalactosamine (GalNAc) or galactose (Gal) being added in group A and B, respectively (see Fig. 1.3).[15, 16]

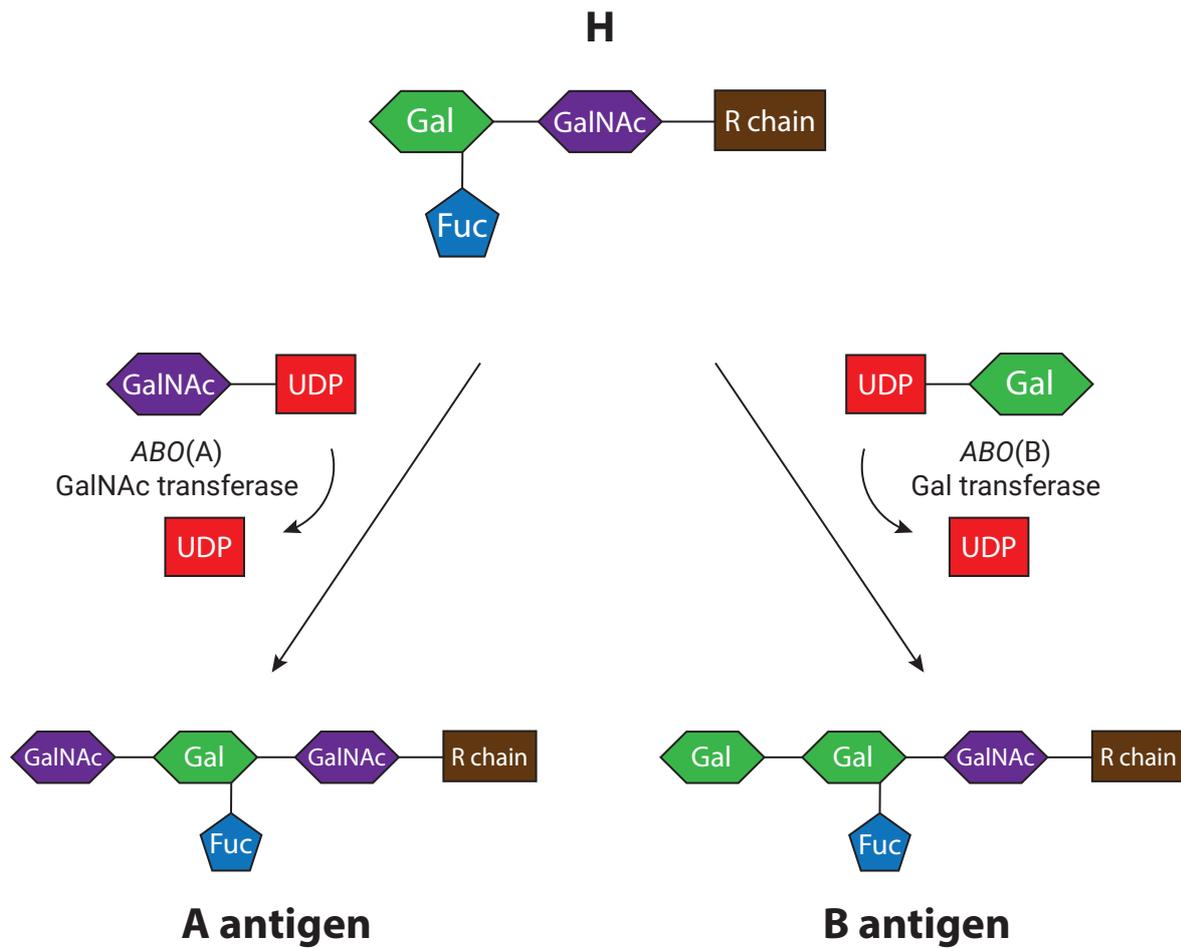


Fig. 1.3 Biosynthesis of A and B antigens from H antigen (Type 2 H shown).

A combination of SNPs at the *ABO* locus controls an individual's capacity to produce GalNAc-transferase or Gal-transferase which catalyse the biosynthesis of the A or B antigen, respectively. The A and B alleles of the *ABO* gene differ by 7 nucleotides, four of which are non-synonymous (ns) ones resulting in amino acid sequence changes that alter enzyme structure and function (see Fig. 1.4). The O phenotype, a lack of either A or B antigen, is due to a non-functional *ABO* gene. The most common O allele is identical to the *A1* allele but contains a deletion resulting in a frameshift and premature stop codon, producing a non-functional protein that lacks its catalytic domain. It is important to note that there are important differences in *ABO* allele frequencies between individuals of different ethnicities. *A1* (A101) is most common in Caucasians (90%), *A1* (A102) is most common in Asians (85%), the two alleles differ by one nsSNP, resulting in the amino acid polymorphism p.(Pro156Leu). For blood group O phenotypes, *O1* (O01) (LRG_792:c.261delG) is the most common genotype in Caucasian populations while *O2* (O03) (c.802GG>AG) is common in individuals of African ancestry.[17]

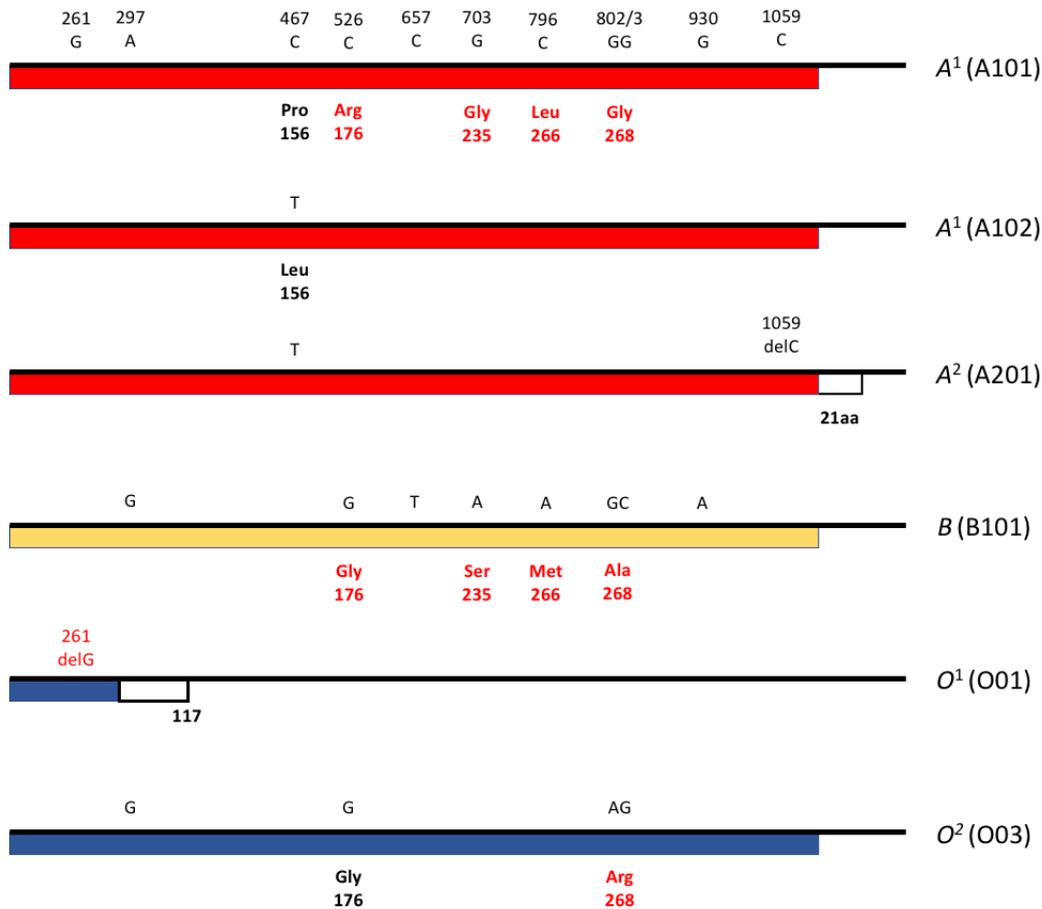


Fig. 1.4 Molecular configuration of 6 common *ABO* gene alleles. Black line represents cDNA sequence; coloured block represents protein product. cDNA changes are above each image; amino acid changes are below. Amino acid changes affecting the catalytic domain are shown in red. Changes are relative to A1 (A101). Figure modified from [6].

1.4.2 The Kell blood group system

In contrast to *ABO* the Kell blood group antigens are protein determinants. The Kell glycoprotein is the direct product of the *KEL* gene and carries 34 antigens.[18] In comparison to the *ABO* blood group system, the molecular genetics of the Kell system is relatively simple. Expression of the antithetical antigens, K (*KEL*01.01*) and k (*KEL*02*), is controlled by a single nsSNP (LRG_799:c.578>T). Kp^a and Kp^b are two other antithetical Kell system antigens which are also controlled by another SNP (LRG_799:c.841C>T). Other Kell phenotypes with more complex molecular genetics do exist, for example, there are Kell_{null} phenotypes resulting from a lack of the entire Kell protein caused by splice site disruption or multiple SNPs leading to a premature stop codon. Individuals with this genetic configuration and phenotype are extremely rare.[19, 20]

1.5 Donor Genotyping

The ISBT WP maintain reference tables for each blood group system which link known combinations of SNPs, indels and SVs to blood group phenotypes. By using the information in these tables we can create algorithms that allow inference of antigen type given some DNA typing information for an individual (see Fig. 1.1).

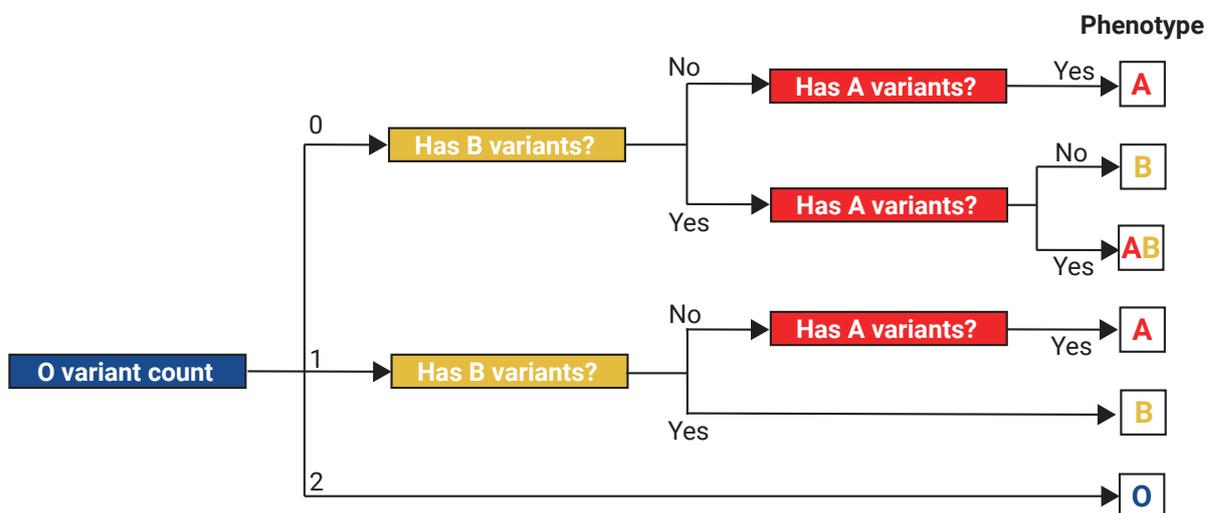


Fig. 1.5 Cartoon showing a simplified ABO genotyping algorithm. This could be applied to the variants shown in Fig. 1.4. Figure modified from [21]

A plethora of donor genotyping tests for typing RBC antigens and Human Platelet Antigens (HPA) antigens has been developed which use a wide range of technologies.[22] The earliest tests used Restriction Fragment Length Polymorphism (RFLP) in which restriction digest of a PCR product is changed by the variant of interest or PCR with sequence-specific primers (PCR-SSP) in which a PCR product is generated only when a specific sequencing is present.[23–25] Although analysis of these tests is cumbersome due to agarose gel electrophoresis being difficult to automate, many laboratory developed RFLP or PCR-SSP tests are still widely used by blood services today when more advanced assays are either not available or cost prohibitive.[26–28]

The next generation of antigen genotyping assays used a variety of techniques to automate analysis of PCR products. TaqMan real time based fluorescent detection of PCR products (RT-PCR) has been used for both RBC and HPA typing, with the latest market entry focusing on ABO antigen genotyping.[29, 30] The most important ongoing use of RT-PCR has been non-invasive foetal *RHD* genotyping.[31] LightCycler PCR product melt point analysis, in which the sequence specific melt-curve of a PCR product is used to detect presence/absence of SNPs, has been used to genotype donors for antigens in the Kell, Kidd, Duffy, MNS, Colton, Dombrock, and Lutheran systems but the technique is susceptible to errors due to unknown sequence variation.[32] Commercial and in-house developed MALDI-TOF assays have been used for typing HPA and Kell antigen typing and are cost effective to run, but they are limited in the number of DNA polymorphisms that can be detected in a single run.[33, 34]

The main commercial assays on the market today are the HEA BeadChip and the BloodChip assay. The HEA BeadChip is based on a short extension reaction of oligonucleotide probes bound to colour-encoded beads using a PCR product as template. The test can type antigens in the RHCE, KEL, FY, DO, LW, CO, SC, LU, DI, JK, MNS blood group systems, but lacks the ability to type the ABO and RHD system antigens.[35] The BloodChip assay is based on allele- and spatial- specific hybridisation of PCR products to complementary oligonucleotides affixed to glass arrays and can be used for clinical typing of 13 RBC antigens (JK, FY, KEL, RH, MNS and DO).[36, 37] Studies using these assays in combination with antibody-based typing have shown that the typing data produced can dramatically increase the availability of antigen negative blood, with one group reporting that they were able to provide blood for 99.8% (5661/5672) of complex blood requests using just 43,066 genotyped donors.[38] In cases where PCR based assays cannot explain a serology phenotype or alloantibody formation, then Sanger sequencing is performed, generally focusing on single exons.

Although many of the assays discussed have gone through regulatory approval for labelling of blood products they have not been widely adopted. This is because the current

commercially available RBC antigen genotyping assays can only detect a limited number of variants and for general typing of the most important systems such as ABO and RH are far more expensive than serological methods. Furthermore, the molecular genetics of some blood group antigens, such as N of the MNS system, is complex and poorly understood and it is thought that serological methods may be more accurate than their DNA based counterparts.

1.6 Population Scale Genotyping

The first draft of the human genome was released in the year 2000.[39] Since then, the reference sequence of the genome (reference genome) has undergone continued re-versioning with each iteration improving the accuracy of gene structures and overall sequence. By comparing the genome sequence of an individual to the reference genome, millions of genetic variants can be identified where the two sequences differ. Through international collaborations the minor allele frequencies (MAFs) of these variants have been defined for several of the ancestral populations at a genome-wide level.

Initially this was done at limited scale by the International HapMap project project in which the genomes of 1,397 individuals from several ethnic groups were genotyped using DNA microarrays.[40] The aim of the study was to identify loci in the genome for assessing risk of common diseases and variants which control quantitative traits of biomedical interest such as height, weight, and blood cell metrics.

In 2006 the Wellcome Trust Case Control Consortium (WTCCC) conducted the first genome-wide association study (GWAS) in which donor samples were used. In this study DNA samples from 2,000 NHS patients with seven common diseases and 3,000 common controls (1,500 were UK blood donors) were genotyped on the Affymetrix GeneChip Mapping Array. These first-generation genome-wide arrays contained probes for typing 500,000 DNA variants and captured about 60% of the known common sequence variation in the genome. The WTCCC study discovered 24 genetically independent loci associated with disease risk thus validating that GWAS could be used as a hypothesis-free method to identify risk loci for human diseases.[41]

The success of this pioneering study sparked a wave of successively larger GWAS meta-analyses in which the DNA of over 0.5M individuals has been typed using genome-wide array technology. Thousands of risk loci have been identified for the majority of common diseases and for biomedically relevant traits, such as the parameters which are measured by the routine full blood count analysis.[42–47] The most recent and more powerful GWAS use data produced by the UK Biobank (UKBB) study where genotype and phenotype was collected on a cohort of 0.5 million individuals in a consistent manner, thereby reducing

the 'noise' in association signals produced by the GWAS techniques applied between 2010 and 2018.[48, 49]. UKBB participants have been typed on a more advanced version of the genome-wide Affymetrix array, named the Axiom UK BioBank array, which contains probes for approximately 800,000 DNA variants. Using the genotyping results from this array and by applying imputation, the minor allele frequency (MAF) for approximately 80 million non-typed DNA variants can be accurately estimated through the principle of linkage disequilibrium (LD) down to a MAF of 0.001.

Several GWAS studies have specifically recruited and genotyped NHS Blood and Transplant (NHSBT) blood donors, using the genetic data produced to inform a future strategy, and to personalise the donation interval.[50, 51] 50,000 of the NHSBT donors enrolled in the INTERVAL randomised controlled trial were genotyped using the UKBB array and approximately 15,000 of these participants have also joined the national NIHR BioResource allowing their genotypes and associated metadata to be returned to NHSBT.

In 2008 the field of genomics was further transformed following a study that used short-read next-generation sequencing (NGS) to determine the genome sequence of a male Yoruba from Ibadan, Nigeria.[52] Illumina commercialised the NGS technology used and a year later the 1000 Genomes Project (1KGP) showed the feasibility of applying whole-genome sequencing (WGS) to a large number of samples.[53]

Today, sequencing data obtained by whole exome sequencing (WES) or WGS on almost 0.5 millions of individuals has been made publicly available by projects such as the African Genome Variation Project (UK - 1,000 genomes), the UK 10,000 project (6,000 by WES, 4,000 by WGS), gnomAD (global - v3 71,702 aggregated genomes), the NIHR BioResource Rare Disease project (UK - 13,072 by WGS), 100KGP (UK - 100,000 by WGS) and the TopMED (USA - 62,784 by WGS).[54–59, 58, 60] This data is easily accessed, and contains valuable information that relates to antigen expression (see Fig. 1.6).

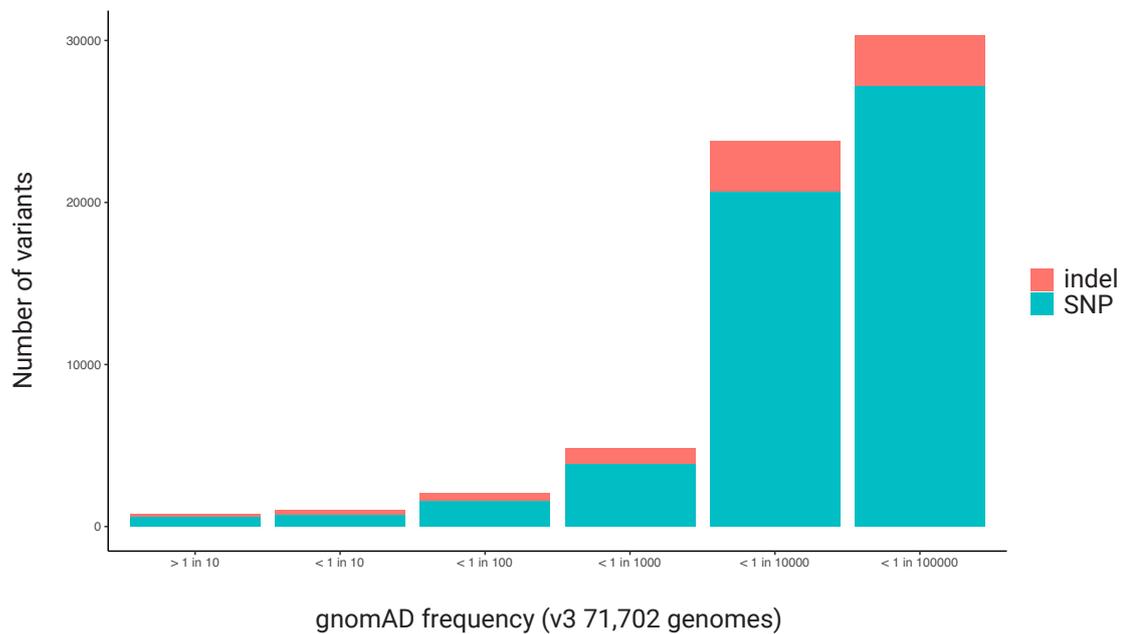


Fig. 1.6 Coding variants in the 44 RBC antigen encoding genes (2020 ISBT tables) observed in the 71,702 whole genomes of the gnomAD v3 dataset. Variants are binned by frequency with the proportion of SNPs and indels coloured in **blue** and **red**, respectively.

These initiatives, alongside execution of even larger scale future population genotyping projects like the Million Veterans Project (USA - 1M array genotypes), the Taiwan Precision Medicine Initiative (Taiwan - 20M array genotypes), GenomeAsia100K Consortium (Asia - 100,000 genomes), the FinnGen cohort (Finland - 500,000 array genotypes), All of Us Research Program (USA - 1M genomes) and the Early Disease Detection Research Project UK (UK - 5M array genotypes), means that by 2025 genomic data will be available for millions of individuals worldwide from a wide range of ethnicities. Furthermore, the DNA extraction, sequencing, and reporting of variants from projects such as the 100KGP have been performed to clinical standards, with sequencing done at an average of 30x read depth and >98.7% of the genome covered at greater than >15x depth. This level of coverage means the sequencing data can be used for diagnostic purposes as demonstrated by the 100KGP pilot study, the results of which led the NHS Executive to secure Illumina WGS capacity for 0.5 M DNA samples for the 2020-2023 period.[60] This means WGS will become the standard test in the UK for diagnosis of rare genetic diseases and identification of driver mutations for selection of precision treatment regimens in cancer patients. It is likely that health services in other countries will also begin implementing the technology over the coming decade.

Use of the technological advancements and vast data produced by the genomics community has been extremely limited within the transfusion medicine community. The first study to apply NGS for blood group antigen typing focused on the *RHD* locus and showed that the technology was superior for detecting variation in these complex loci.[61] Researchers then highlighted the utility of NGS genotyping for the management of transfusion-dependent patients. A study of 35 patients with sickle cell disease reported that 62% of unexplained alloimmunisation events could be resolved by NGS sequencing of *RHD* and *RHCE* genes, finding that current serological methods were not accurate in identifying variant D, C/c and E/e antigens.[62] WES has also been applied for the detection of variant RH antigens and inference of alloimmunisation risk in children with sickle cell anaemia.[63] The earliest attempts at automated interpretation of blood types from genomic data showed promising results, with the BOOGIE study reporting 94% antigen typing accuracy using WGS data. However, the software was never made publicly available for cross-validation limiting further development.[64] In 2016 Lane and colleagues explored the possibility of using NGS in a transfusion medicine setting by performing WGS on a single individual and using the data to infer accurate antigen types.[65] Another study focused on antigen gene sequence variation in the 1KGP dataset, and found that only 19% of the variation observed was present in ISBT reference tables.[66]

1.7 Aims

Even though many donor genotyping assays exist and there is evidence to suggest that their use can improve patient care and simplify the provision of blood, they have not been widely adopted by blood supply organisations. This is largely due to the limited number of antigens typed for by each assay, their cost and the lack of automated analysis software for the interpretation of results.

This means that despite the advancements made since the early days of antibody-based typing, 85% of donors in England have no typing data for RBC antigens encoded by common alleles (other than ABO and RH) and 94% of English donors have no typing data for rare antigens (information from look-up in NHSBT's PULSE database - 2019).

The results of recent studies are clear; the introduction of genomics technologies into transfusion medicine will enable us to better classify and type blood donors and patients thereby greatly simplifying the challenges of providing better-matched blood to patients requiring regular transfusion.

The purpose of this study is to assess the feasibility of determining human blood groups using the high throughput molecular techniques developed for population-scale genotyping initiatives, namely genotyping arrays and NGS. The ultimate aim is to use the vast amount of genomic data that is now available to guide the development of an affordable and automated donor genotyping test, validated for typing all transfusion and transplant relevant antigens which can be used by blood supply organisations to deliver genomics-based precision transfusion medicine to the patient bedside.

Chapter 2

Materials and Methods

2.1 Overview

This study involved several complex and interrelated components (Figure 1). We have written this thesis in such a way that each results chapter will have an introductory section to frame the work it describes.

Chapter 3 details the process of using RNA sequencing data from the BLUEPRINT study to select and curate fixed reference sequences for each of the RBC antigen encoding genes. Following this, we analyse whole genome sequencing data from the MedSeq, INTERVAL and, NIHR BioResource Rare Disease studies to show the benefits of computationally mapping the knowledge of genetic antigen expression to the human reference genome.

Chapter 4 details the development and validation of an array-based donor genotyping assay capable of typing all RBC, HPA, and HLA antigens. First, we used the Axiom UK Biobank version 1 array (UKBBv1 array) genotyping data from donors enrolled in the INTERVAL study to assess if the Axiom array technology is a suitable platform for antigen typing. Next, we selected known blood typing variants from various reference sources and designed the 384HT Axiom Blood Typing SNP Screen Array (384 BT array). We validated 384 BT array blood typing content using DNA samples from NHSBT blood donors enrolled in the NIHR Bioresource. Finally, combined the novel blood typing array content with the previous UK Biobank array content, creating the UK Biobank version 2 array (UKBBv2 array) which was then validated using DNA samples from blood donors from England and the Netherlands.

Chapter 5 details the development and validation of an affordable targeted NGS assay for in-depth investigation of antigen types when discordance is observed between antibody phenotyping results and genotyping results from donor screening assays such as the UKBBv2 array.

A high-level summary of this project is given in Fig. 2.1.

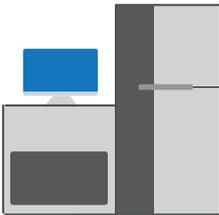
	Purpose	Technology	Data and Samples	
Chapter 3	1) Computationally map blood groups to human genome	 Illumina NGS	 BLUEPRINT RNA-seq n=90	 MedSeq 30x WGS n=110
	2) Demonstrate the power of WGS for antigen typing		 INTERVAL 15x WGS n=200	 NIHR Rare Disease 30x WGS n=13,037
Chapter 4	1) Develop array genotyping assay that can type all RBC, HPA and HLA antigens	 Thermo Fisher Axiom Array	 INTERVAL UKBBv1 array n=1,057	 NIHR BioResource 384 BT array n=507
	2) Validate by comparing array inferred antigen types to clinical blood types from electronic donor record		 COMPARE UKBBv2 array n=4,795	 DIS-III UKBBv2 array n=2,682
Chapter 5	1) Develop and validate targeted NGS assay for investigating genotype/phenotype discordance	 Targeted Illumina NGS	 NIHR Rare Disease 30x WGS + Targeted NGS n=48	 NIHR BioResource BGC capture n=48
	2) Use assay to resolve discordance observed in chapter 4	 Thermo Fisher Axiom Array	 COMPARE & DIS-III UKBBv2 array + Targeted NGS n=61 (discordant samples)	

Fig. 2.1 High level project overview showing the purpose of each chapter, the predominant technology used, and the sources of samples and data used.

2.2 An overview of Axiom array genotyping

The Thermo Fisher Scientific (TFS) Axiom genotyping platform is a ligation-based microarray utilising 30-mer oligonucleotide probes which are synthesised in situ on a glass slide. Two fluorescent colour channels and spatial position on the array slide are used to detect genotypes. There are multiple formats of the Axiom platform. The two used in this study are the 96-format - capable of genotyping 96 samples in parallel for 800,000 DNA variants, and the 384-format - capable of genotyping 384 samples in parallel for 50,000 DNA variants.

The surface of the glass array slide is divided into features which are 3 μm squares. Each feature contains millions of identical unique 30 base oligonucleotide probes that are complementary to the genomic sequence which flanks the variant of interest (either forward or reverse). Features are replicated multiple times on each array to improve the resolution of specific SNPs and provide redundancy in the case of failure. A \leftrightarrow T and C \leftrightarrow G variation must also be represented by two spatially separated features as only two dyes are used, one for A,T and another for C,G.

In a standard genotyping run sample DNA is amplified, fragmented, and then hybridised to the probe/array complex. A solution of detection probes, many DNA probes representing every possible combination of 9 DNA bases labelled with a base-specific 'hapten', is then washed over the array and covalently bonded to the array-probe/genomic-DNA complex. A stringent wash cycle is performed to remove unbound solution probes and genomic DNA. Staining is then performed with fluorescently labelled antibodies, laser excitation produces a signal, and fluorescent intensity measurements are used to infer genotype.

An overview of the Axiom genotyping process is given in Fig. 2.2.

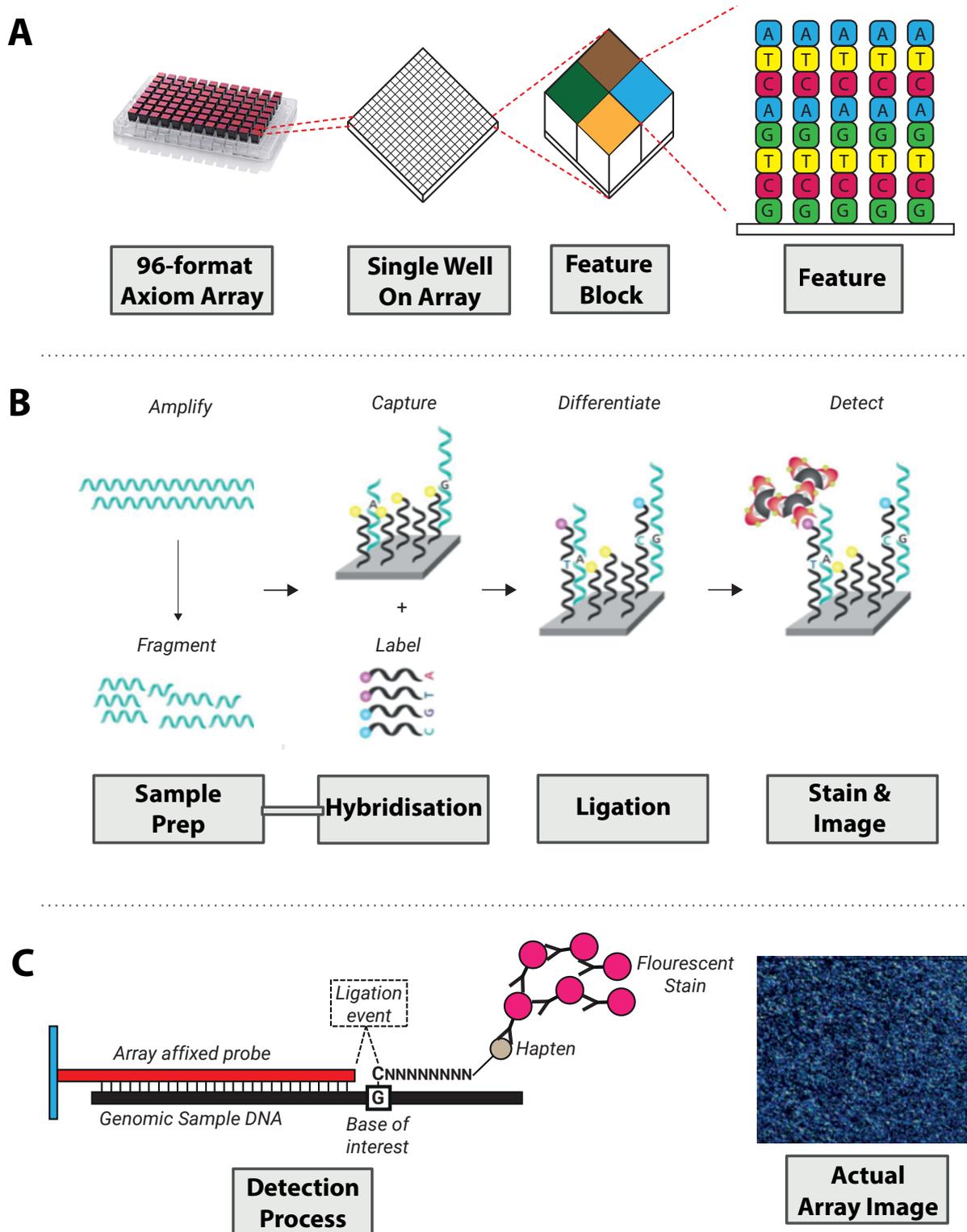


Fig. 2.2 Overview of Axiom array genotyping. (a) A zoom diagram showing the construction of an Axiom array. (b) Axiom genotyping workflow. From left to right; Genomic DNA is amplified then fragmented, fragments are captured and labelled via hybridisation to array probes and detection probes, ligation then covalently bonds the genomic DNA to the array/detection probe complex, finally fluorescent staining is performed followed by excitation and imaging. (c) Cartoon representation of the genomic DNA + array/detection probe complex. An actual image of an array during genotyping is included for reference.

Each DNA sample used for array validation was genotyped on the Axiom platform at the Applied Biosystems Microarray Research Services Laboratory, Santa Clara, CA. Genotyping was carried out in accordance with Axiom Best Practice Workflows.[67] 750ng total DNA at 30ng/μl from each sample was used.

Genotypes were called using the AxiomGT1 algorithm included with the Applied Biosystems Array Power Tools v2.10.2 software (APT).[67] Quality control metrics for each plate, sample and probeset were calculated automatically using APT. In brief, this involves ensuring that the fraction of 'AT' probesets for which fluorescence is 2 standard deviations outside GC probeset fluorescence is within range, and genotype call rates for each sample and variant are above 95%. Probes producing values below the recommended thresholds were not used for analysis. For variants with more than one probeset remaining, a single best probeset was selected using Fisher's linear discriminant as a measure of cluster resolution.

We also visually inspected genotype call plots for all newly added antigen typing probesets as per the original UKBB study methodology (see Fig. 2.3).[49] Data generated by any poorly performing probesets identified at this stage were not used in further analysis.

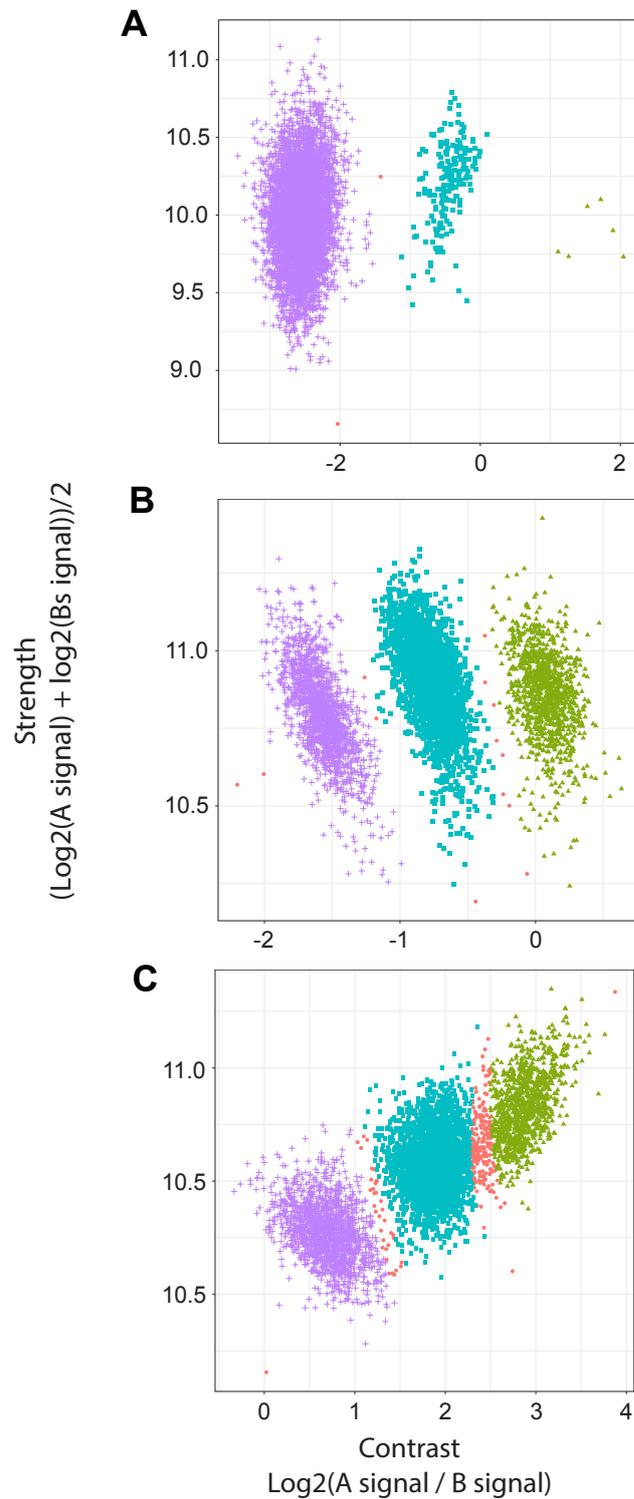


Fig. 2.3 Genotype call plots were visually inspected as part of DNA probe performance assessment. Examples of (a) a “good quality” call plot with clear separation between genotype clusters; (b) a “requires development” call plot, three unique genotype clusters can be observed, however, boundaries are too close; (c) a “poor quality” call plot with overlap between genotype clusters. Colour represents genotype call: homozygous reference (purple), heterozygous (blue), homozygous alternate (green); failed call (red).

2.3 Next Generation Sequencing

Illumina Next Generation Sequencing (NGS) technology uses the sequencing by DNA synthesis technique.[68] In brief, genomic DNA is fragmented and short oligonucleotide sequencing adaptors are ligated onto the ends of each fragment, the DNA fragments are then loaded into a flow cell and fixed via binding of adaptors to complementary oligonucleotides bound to the flow cell surface, bridge amplification via PCR is then conducted to form clusters of identical DNA strands in localised regions of the flow cell surface. A number of sequencing cycles follow this in which fluorescently-labelled reversible terminator bound dNTPs are added to the flow cell, DNA synthesis using the clustered DNA fragments as templates is performed and a single dNTP is incorporated each cycle, the reaction is stopped after a given time and unincorporated dNTP's are washed away, imaging is performed, and finally the terminator/label is chemically removed from the forming DNA strand. The actual DNA sequence for each read is then reconstructed using the images taken.

An overview of the Illumina short-read NGS process is given in Fig. 2.4.

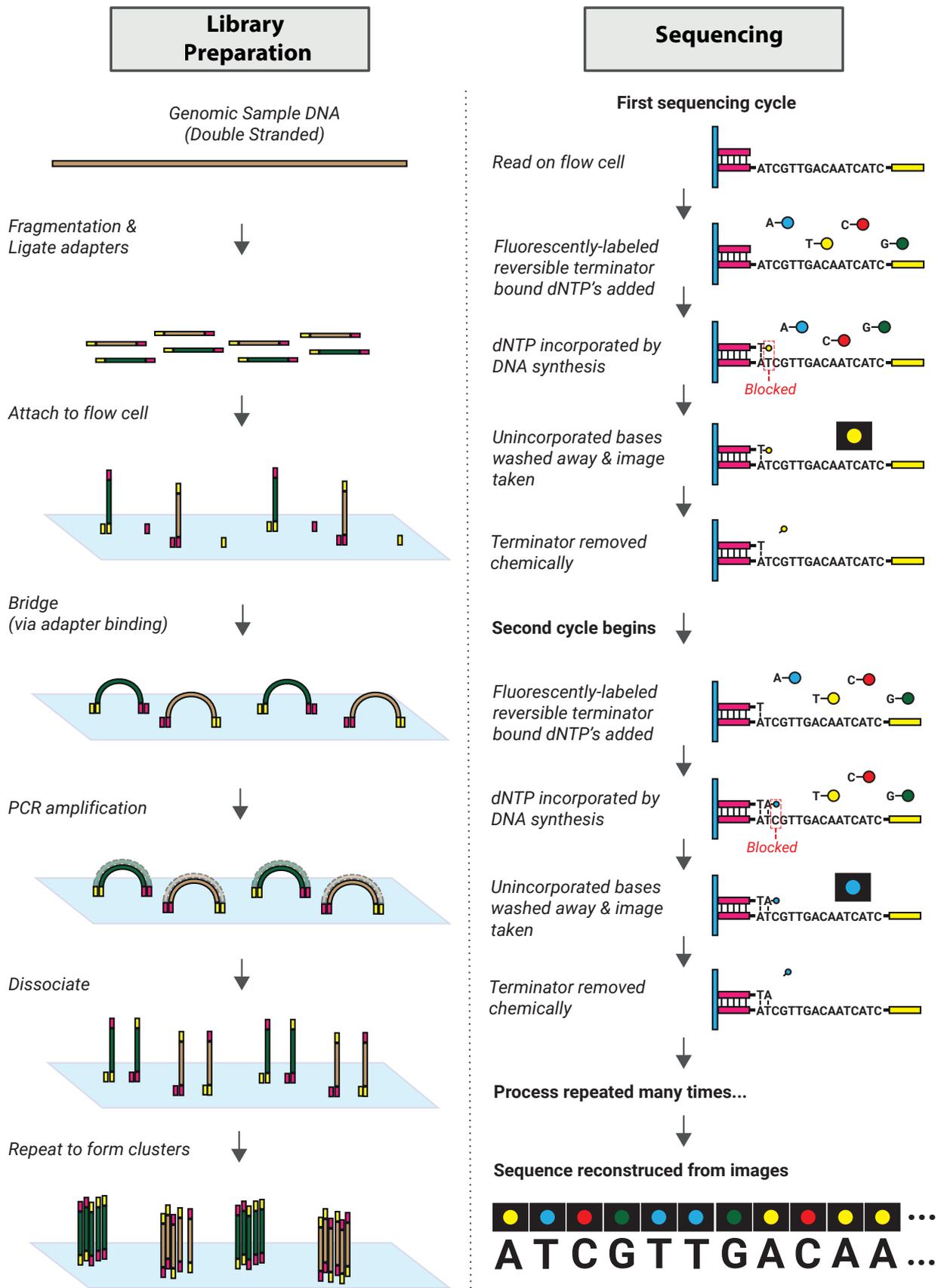


Fig. 2.4 Overview of Illumina NGS sequencing

Following sequencing, the millions of reads produced are computationally aligned to a reference genome (see Fig. 2.5).

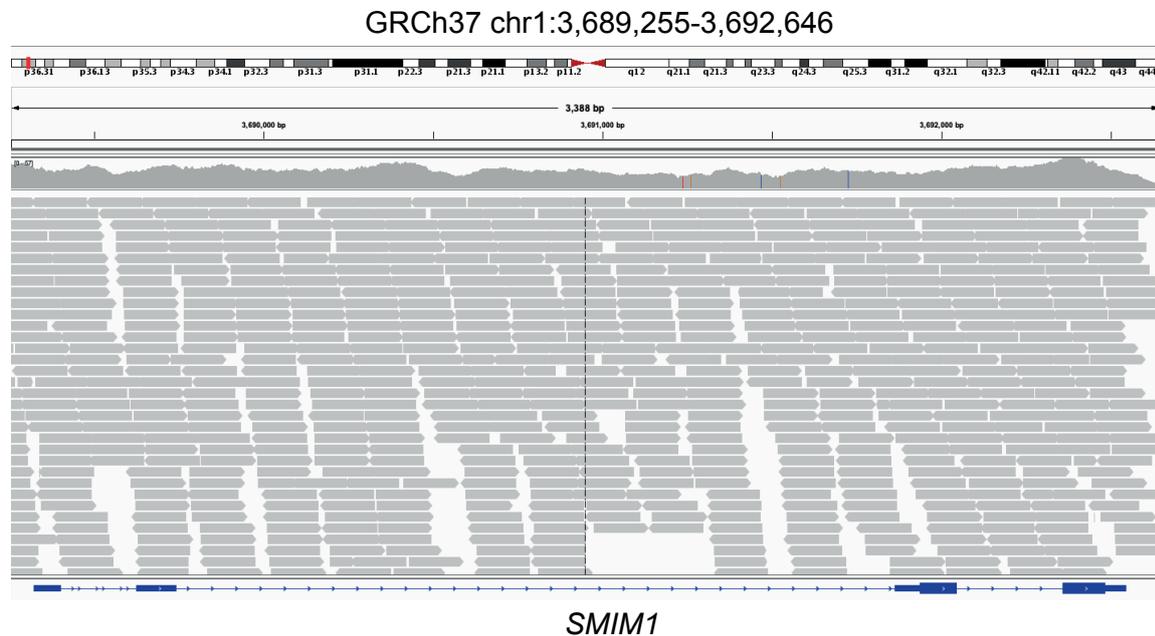


Fig. 2.5 NGS sequence alignment at the *SMIM1* loci which is responsible for Vel antigen expression. From top to bottom; A map view of the chromosome with red box indicating the location of the viewing window. Next, a histogram showing read depth (the number of reads aligned over a particular nucleotide position). Next, the alignment view mapped Reads are shown as grey boxes, a point at either end indicates the orientation of the read (forward > or reverse <). The structure of the *SMIM1* gene is shown in blue at the bottom of the figure, the thin line with arrows represents intronic sequence, thick and thin boxes represent transcribed and un-transcribed exonic regions, respectively.

Targeted sequencing

In chapter 5 a targeted or capture-based NGS sequencing assay was developed to resolve discordance between clinical and array antigen typing results. The only difference between standard NGS and capture sequencing is a hybridisation capture step between DNA fragmentation and adaptor ligation. In this step target DNA is hybridised to oligonucleotides primers or 'baits' affixed to synthetic beads, the bait oligos are complementary to the genomic region of interest. Following hybridisation a wash step is used to remove unbound 'off-target' DNA, thus enriching the sequencing library for the region of interest.

Targeted sequencing was performed as follows. Aliquots of original DNA samples were retrieved from research biobanks and 1 µg of each sample was fragmented using Covaris E220 (Covaris Inc., Woburn, MA) to obtain 150 bp average size fragments. Samples were processed using the TruSeq DNA LT Prep kit (Illumina Inc., San Diego, CA). Two DNA libraries were captured at the same time using a single reaction of the SeqCap BG capture array (ROCHE NimbleGen, Inc. Madison, WI). Enrichment was tested by qPCR measuring the abundance of four control target regions before and after DNA capture. The libraries were quantified using the Library quantification method (KAPA Biosystem, Ltd, Cape Town, South Africa) and sequenced by the CRUK Cambridge Institute Core Genomics facility in pools of 24 samples in one Illumina HiSeq 2000 lane using 150bp paired-end sequencing.

2.4 Bioinformatics

2.4.1 bloodTyper

All genetically inferred RBC and PLT antigen typing results presented in this work were produced using the bloodTyper algorithm. bloodTyper utilises a curated antigen allele database, containing all known antigen encoding variants, to infer antigen status from genomic data. An in-depth explanation of the bloodTyper algorithm has been previously published.[65]

An overview of the bloodTyper pipeline is given in Fig. 2.6.

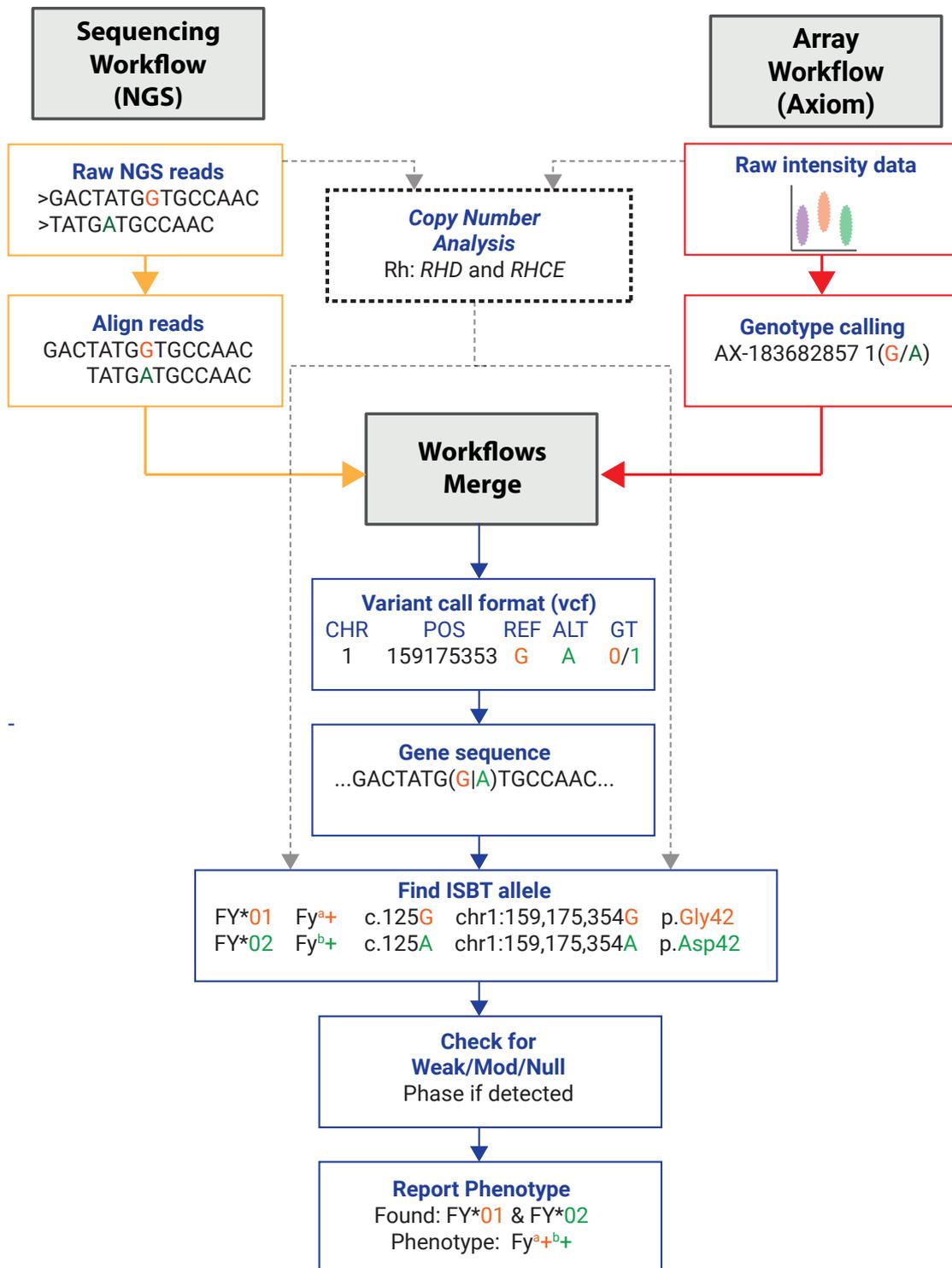


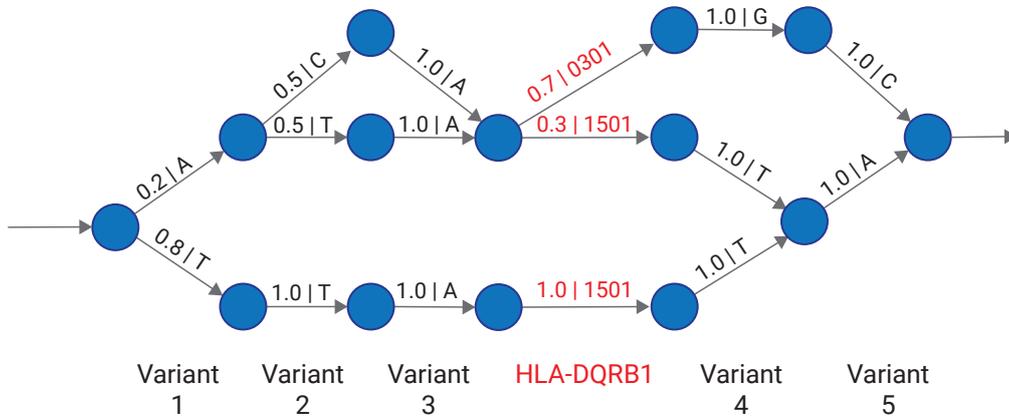
Fig. 2.6 Overview of the bloodTyper analysis pipeline.

2.4.2 HLA impute

All genetically inferred 4-digit HLA typing results presented in this work were produced using the HLA*IMP:02 imputation algorithm. HLA*IMP:02 uses a multi-population reference panel to build a graphical representation of haplotype structure at the HLA locus, accounting for genotype error and haplotypic heterogeneity. In layman's terms, the frequency information for all genetic variants across the HLA locus is extracted from a set of known HLA typed haplotype reference sequenced. Then the probability of each combination of the variants is mapped into a graph structure.

The model then computes a probabilistic HLA type for each sample, given their sequence. An in-depth explanation of the HLA*IMP:02 algorithm has been previously published.[69]

A simplified overview of the HLA*IMP:02 model is given in Fig. 2.7.

A**B**

Sample	Sequence	Path Through Graph	HLA-DQRB1 Estimate
Sample 1	V1 V2 V3 V4 V5 T T A T A	 T T A 1501 T A 0.8 1.0 1.0 1.0 1.0 1.0	1501 p= 0.8
Sample 2	V1 V2 V3 V4 V5 A T A T A	 A T A 1501 T A 0.2 0.5 1.0 0.3 1.0 1.0	1501 p= 0.03
Sample 3	V1 V2 V3 V4 V5 A C A G C	 A T A 1501 T C 0.2 0.5 1.0 0.7 1.0 1.0	0301 p=0.07

Fig. 2.7 Overview of the HLA*IMP:02 model. (a) A graphical representation of all haplotypes is created for the HLA locus. Blue dots represent nodes at each graph vertex (column). The paths between vertices represent different possible variants, the reference panel probability of a given nucleotide and the nucleotide itself is annotated on top of each path. We show the probability of different HLA-DQRB1 genotypes in red, these are linked to each possible path through the graph. (b) Use of the haplotype graph to impute HLA types. Sequence, haplotype graph path, and predicted HLA-DQRB1 type is shown for three theoretical samples.

2.4.3 BWA, SAMtools, and BCFtools

BWA v 0.7.12, SAMtools v1.9, and BCFtools v1.9 were used extensively during this project for the processing of genomic data. All three are available for download free of charge online at <https://github.com/lh3/bwa> and <https://github.com/samtools/>.

BWA is a software tool for alignment of sequencing reads to reference genomes. Three alignment algorithms are available. The latest, BWA-MEM, was used to produce all sequencing alignments in this study.[70]

SAMtools is a software library for processing genotyping and sequencing data. It is used to convert between alignment formats, sort and merge alignments, remove PCR duplicates, and generate per-position information in the pileup format. [70]

BCFtools is an extension to the original SAMtools software package and sits within the HTSLib environment. It is used for the manipulation of VCF files including; format conversion, querying, genotype annotation, and sequence reconstruction.[71]

2.4.4 Genome Analysis Toolkit, Manta and Canvas

Three variant callers were used in this study to detect genetic variants from NGS data. The Genome Analysis Toolkit (GATK) v3.1 was used for calling SNPs and indels.[72] Manta v1.6 and Canvas v1.3 were used to call structural variants (SVs).[73, 74] All three are available for download free of charge.

GATK is a standard software package for identifying genetic variants from NGS sequencing data. In addition to several variant callers, GATK also includes quality control utilities such as Picard for de-duplication of sequencing reads. GATK does not currently support calling of structural variants.

Manta is an SV caller that uses paired and split-read evidence (paired reads for which opposite ends align to different parts of the genome) to identify SV breakpoints. Canvas is also an SV caller, however uses differences in read depths for the detection of SV's. Calling of structural variants has a higher error rate than other types of variant, in this study we therefore cross-validate SV calls by analysing only those reported by both tools.

2.4.5 Axiom Analysis Suite

Axiom Analysis Suite, created initially by Affymetrix and now published by Thermo Fisher Scientific is a set of freely available tools for analysing Axiom genotyping data.[75] In this study it was used for genotyping calling and genotype quality control. Axiom Analysis Suite is available on request from Thermo Fisher Scientific.

2.5 Summary of external data sources

Samples and data from many different previously published studies have been used throughout this thesis. Rather than copy out the methods of these studies we have made reference to the publications where detailed study methodologies can be found. For ease of access, we have also summarised this information in the below table.

Table 2.1 Data sources

Study	Data Type	Participant Number	Participant summary	Reference
UK Biobank	Array	500,000	UK Population sample	[49]
NIHR BioResource	Array	150,000	Healthy Controls Common Disease Patients	[76]
gnomAD (v3)	WGS	71,702	Healthy Controls Rare Disease Patients	[59]
INTERVAL	Array and WGS	45,263	NHSBT blood donors	[50]
COMPARE	Array	29,874	NHSBT blood donors	[77]
NIHR BioResource Rare Disease	WGS	13,037	Rare Disease Patients Family Members	[60]
DIS-III	Array	3,046	Sanquin Blood Donors	[78]
MedSeq	WGS	200	Healthy Individuals Cardiac Patients	[79]
BLUEPRINT	RNAseq	90	NHSBT blood donors	[80]

2.6 Blood Donor DNA samples

We made use of 11,923 DNA samples from English and Dutch blood donors in this study. The samples and data from all participants were obtained after informed consent, see specific panel descriptions for more detail.

2.6.1 INTERVAL panel (n=1,257)

The INTERVAL randomised control trial used data from 45,263 English donors to find the optimum interval for which it is safe for different donors to give blood (Interval: study of optimum frequency of blood donations in England, Research Ethics Committee reference: 11/EE/0538).[81, 50]

WGS data for the 200 samples used in chapter 3 was generated using 15x Illumina short read sequencing. Array genotyping data for the 1,057 samples used in chapter 4 was generated using the UKBBv1 array and NHSBT clinical antigen typing data was available on request to the study organisers.

2.6.2 NIHR BioResource panel (n=507)

The NIHR BioResource is a panel of over 150,000 volunteers (NIHR BioResource – Research Tissue Bank, Research Ethics Committee reference: 17/EE/0025), with and without health problems. Data is made available to enable studies on the association between phenotype and genotype. Approximately 10,000 members of the NIHR BioResource were NHSBT blood donors and clinical antigen typing data was retrieved from NHSBT’s electronic donor record.

2.6.3 COMPARE and DIS3 panel (n=7,477)

Trial set: COMPARE study panel (n=4,795). The COMPARE study enrolled 29,066 English blood donors between February 2016 and March 2017 (Comparison of NHSBT’s current approach with three alternative strategies to assess haemoglobin levels in whole blood donors (Research Ethics Committee reference: 11/EE/0335).¹ The study aim is to find the optimum technology for haemoglobin screening. All participants were active blood donors and clinical antigen typing data was retrieved from NHSBT’s electronic donor record. The 4,795 participants used in this study were selected based on also being participants in the NIHR BioResource.

Trial set: Donor InSight III panel (n=2,682). The Donor InSight-III (DIS-III) enrolled 3,046 Dutch blood donors to form a research panel to allow scientific insight into donor characteristics, motivations and health (METC 2014/124, NL47865.018.14).[78] Additionally, 95 newly registered donors were enrolled between May 2017 and August 2017 using the DIS-III study protocol. All participants were active blood donors and clinical antigen typing data was retrieved from Sanquin databases. The 2,682 participants used in this study were selected based on the availability of extracted DNA samples.

The COMPARE and DIS-III panels, totalling 7,477 English and Dutch blood donors, were used for validation of the final donor typing array presented in this study, named the Applied Biosystems UK Biobank – version 2 Axiom Array (UKBBv2 array).

2.7 Clinical RBC and PLT antigen typing data

Antigen typing data for the 11,923 donors were generated using clinically accredited tests as part of routine donor typing by NHSBT and Sanquin. Only antigen types that had been verified by at least two independent measurements were used. We refer to these antigen types as “clinical types” as they are a combination of antibody- and DNA- determined types. Results were available for 48 RBC, 11 HPA and 6 HLA antigens (see Table. 2.2).

Clinical typing results for the ABO, D, C, c, E, e, K, k, Kp^a, Jk^a, Jk^b, Fy^a, Fy^b, Le^a, Le^b, Lu^a, M, N, S, and s antigens were produced primarily using commercial high throughput serological phenotyping systems, for example the Olympus PK7300 instrument. Clinical types for the other RBC antigens were produced using in-house manual antibody-based typing tests. In-house polymerase chain reaction (PCR) based techniques were used to type antigens such as Do^a, Do^b, and LAN where no reliable antibody testing reagents are available. HLA antigen typing results were produced using commercially available genotyping tests or in-house next generation sequencing tests. HPA genotyping was performed using PCR based tests developed in-house by NHSBT and Sanquin.

Table 2.2 Antigen typing data available in electronic donor record

Antigen	Results Available in Electronic Donor Record
ABO	7449
M	4012
N	1440
S	4069
s	3095
P1	0
D	7467
C	7441
c	7443
E	7453
e	7447
C(W)	3641
C(X)	2
V	1
VS	1
Vw	27
Lu(a)	882
Lu(b)	949
K	7424
k	876
Kp(a)	1112
Kp(b)	931
Js(a)	2
Js(b)	3
Le(b)	1069
Le(a)	1497
Fy(a)	3051
Fy(b)	2518
Jk(a)	4388
Jk(b)	4359
Di(a)	1
Di(b)	1
Yt(a)	97
Sc1	3
Sc2	1
Do(a)	7
Do(b)	5
Co(a)	130
Co(b)	212
Yk(a)	1
Kn(a)	5
In(b)	6
Wr(b)	8
I	1
JMHK	1
Wr(a)	611
Lan	149
Vel	698
HPA-1a	501
HPA-1b	338
HPA-2a	343
HPA-2b	343
HPA-4a	69
HPA-4b	69
HPA-5a	343
HPA-5b	342
HPA-6bw	1
HPA-15a	334
HPA-15b	334
HLA-A	2395
HLA-B	2413
HLA-C	2134
HLA-DPB1	346
HLA-DQB1	1426
HLA-DRB1	2225

Reformatting of clinical HLA typing data

Clinical HLA typing data for DIS-III participants were recorded using several different formats in the Sanquin database and required formatting before use. Where a multiple allele code was used (e.g. 07:GS), a lookup table developed and maintained by NHSBT was used to generate an allele string (e.g. *07:01*07:03*07:040). Where an allele name ended in G, the result was not changed (e.g. A*01:01:01G). A “G” code indicates groups of alleles that share the same sequence in the peptide-binding groove, many of the member alleles in each group only differ in the third field which was not typed in this study. Where an allele had the XX suffix (e.g. *02:XX) the result was converted into a single field result (e.g. *02). The XX suffix indicates that the second field can be any allele within the primary group making any second field comparison unreliable.

2.8 RBC and HPA Antigen typing concordance analysis

Antigens for which fewer than 10 comparisons between clinical and array antigen types were possible were excluded from concordance analysis. We also excluded the Le^a and Le^b antigens because anti-Le antibodies are not usually clinically significant and the ISBT table required for variant interpretation is lacking.[7] P1 antigen typing was disabled in bloodTyper as the molecular basis of this antigen was not defined at the time of platform design.

2.9 HLA antigen typing concordance analysis

Four-digit genotype inferred HLA types were compared to clinical HLA typing data extracted from NHSBT and Sanquin databases using the following match algorithm:

1. **Allele Match:** Both results match for the first two fields (e.g. *32:01 and *32:01:01)
2. **String Match:** Clinical typing result is an ambiguous string of ‘potential’ alleles which contains the genotype inferred result (e.g. *02:01 is within *02:01/*02:04/*02:07/*02:09)
3. **Group Match:** Both results match in the first field, but due to lack of clinical typing data a second field comparison cannot be made (e.g. *24:02 and *24) Mismatch: Both results are from different allele groups e.g. *25:01 and *26:01:01

2.10 Data availability

Genotype data produced for blood donor DNA validation samples can be retrieved from the European Genome-Phenome Archive (<https://www.ebi.ac.uk/ega/home>) at the EMBL European Bioinformatics Institute (Hinxton, Cambridge, UK; EBI) using the following study numbers: NIHR BioResource participants - EGAD00001005024; COMPARE study participants - EGAD00001005023; DIS-III study participants - EGAD00001005026.

The genotype data for UK Biobank samples will be made available through a data-release process that is being overseen by the UK Biobank (<https://www.ukbiobank.ac.uk/>).

Genotype and phenotype data from the National Institute for Health Research (NIHR) BioResource Rare Diseases Pilot can be accessed by application to Genomics England Ltd (<https://www.genomicsengland.co.uk/about-gecip/joining-researchcommunity/>).

2.11 My contribution to each chapter

The work presented in the thesis would not have been possible without the collaborative efforts of many experts from around the world. For examination purposes, I have detailed below my specific contributions to the work performed in each chapter.

Of the research in chapter 3, I was directly responsible for; 1) Bringing together the members of the Blood Transfusion Genomics Consortium by enlisting each member, organising £5000 funding to enable the first BGC meeting, and playing an essential role in designing the BGC research programme, 2) Analysis of Blueprint expression data for blood group genes and in-depth curation of reference transcripts through bioinformatic analysis and six months of weekly calls with members of the transcript curation the Locus Reference Genomics team at the EMBL European Bioinformatics Institute. I was also responsible for relaying suggested transcripts back to the relevant ISBT working party members for approval, 3) Quality control, alignment, variant calling and analysis of genotyping and blood typing results using the whole genome sequencing results for 200 INTERVAL participants, 4) Blood typing analysis of the entire NIHR BioResource Rare Disease cohort, including resolution of the complex *RHAG*_{null} structural variant (SV) in one of the cases and guiding re-validation of the SV variant calling pipeline used by the project as initially the large duplication and other SVs such as deletion of the *RHD* gene were not being reliably detected, 5) phasing, reconstruction and analysis of *KEL* gene haplotypes for the same cohort.

Of the research in chapter 4 I was directly responsible for; 1) Sample panel creation, 2) Quality control of genotype and clinical antigen typing data, 3) Concordance analysis and further development and validation of array specific RBC and PLT antigen calling algorithms,

4) HLA type imputation, 5) Genotype array design for all cohorts except INTERVAL, 6) All associated analysis of derived data, for example, the historical blood demand analysis. Of the research in chapter 5 I was directly responsible for; 1) Sample panel creation, 2) Development and quality control of the BGC capture platform, 3) Quality control and data processing of genotype and clinical antigen typing data including relatedness analysis to link BGC capture sequencing data to WGS data obtained from the same DNA samples, 3) Analysis of discordant genotype and antigen typing results, 4) Analysis of derived data, for example, investigation of samples with genetically complex antigen types and genotype concordance analysis.

Chapter 3

Mapping blood group antigens to the human genome

3.1 Abstract

In recent years genomic data for millions of individuals from around the world has been made publicly available. More recent population-scale studies such as the 100,000 Genomes Project (100KGP) in the UK and others have analysed DNA samples by whole genome sequencing (WGS) in accordance with clinical standards allowing the data to be used in a diagnostic setting. This investment in introducing WGS into clinical care provides many opportunities to deliver genomics-based precision medicine to the patient bedside. In order to achieve this within the field of transfusion medicine, the knowledge of the genetics which underpins blood group antigen expression must be computationally mapped to the human genome and international standards created for the analysis of genomic data with respect to blood group antigens.

In this chapter, we discuss the formation and initial work of the Blood transfusion Genomics Consortium (BGC). This collaboration has brought together experts in the fields of transfusion medicine, immunogenetics, genomics and data science from around the world with the common aim of addressing the challenges mentioned above so that genomics technologies can be safely integrated into clinical transfusion practice. Through collaboration with the EMBL European Bioinformatics Institute (EBI), the BGC members have conducted the first extensive data-driven review of the International Society of Blood Transfusion (ISBT) blood group gene reference transcripts and established fixed sequence records for each of them, taking the first step towards creating international analysis standards.

We go on to demonstrate the benefits of this work by using WGS data to; further develop and validate analysis tools for blood cell antigen genotyping, provide clinically relevant blood typing information for 13,037 patients and their close relatives who were recruited to the NIHR BioResource WGS project, and fully explore the sequence variation observed for these 13,037 individuals at the genetic locus which controls the Kell blood group system.

The description of part of the work in section 3.6.1 of this chapter is based on that given in: W.J. Lane, C.M. Westhoff, N.S. Gleadall, et al., (2018). “Automated typing of red blood cell and platelet antigens: a whole-genome sequencing study”. *Lancet Haematology*. 2018 Jun; 5(6): e241–e251.

The work underpinning resolution of the clinical case in section 3.6.2 of this chapter was used in: E. Turro, et al., (2020). “Whole-genome sequencing of patients with rare diseases in a national health system”. *Nature*. 2020 Jul;583(7814):96-102.

3.2 Introduction

While the initial focus of genomic medicine will be on rare disease and cancer patients, the sequence data being generated provides the transfusion medicine community with unrivalled opportunities to further characterise blood group relevant loci via; cataloguing DNA variant level frequency data in different ancestral populations, the definition of haplotype sequences for each antigen system, and identification of previously unobserved DNA variants which are clinically relevant because they affect antigen expression. However, in order to analyse this vast amount of data in a meaningful way, the decades of knowledge amassed on the genetics of antigen expression on blood cells must be computationally mapped to the human genome.

Several research groups have begun this task and started using publicly available genome sequencing and genome-wide array typing data. One such study has evaluated the sequence data from the 1KGP and compared it with known variation described in the ISBT antigen tables from the Red Cell Immunogenetics and Blood Group Terminology working party of the (ISBT WP), dbRBC, and The Blood Group Antigen FactsBook.[66] The conclusions drawn were that in many instances blood group gene reference sequences were inadequate and that considerable effort must be put towards phasing sequence data before attempting to infer the genotype status of variants and relate them to blood group alleles. Many errors of different types were discovered in the reference sources (e.g. error in the initial publication of an antigen or typographical errors in the WP tables). The results of this analysis reinforced the importance of mapping and curating the knowledge of molecular antigen expression to rigorous standards, which are internationally accepted.

Additionally, the analysis of the 1KGP data revealed that only 19% of the 1,241 non-synonymous (ns) single nucleotide variants (SNP) identified in loci encoding blood groups were found in the WP allele tables. This finding highlights that, whilst blood groups have been well characterised to date, there has been a significant underestimation of the underlying genetic diversity at all blood group loci. The 1KGP was the first to pilot WGS at scale and produced only a read-depth of 4x on average using first-generation Solexa (subsequently bought by Illumina) instruments, thus the confidence of calls of variants of low MAFs in the 1KGP dataset is limited. These issues have now been resolved with the high read depth (>30x) achieved for the 100KGP and in this chapter, we show that the latter dataset gives a highly reliable catalogue of sequence variation in antigen encoding loci.

In 2016 Lane and colleagues explored the possibility of using NGS in a transfusion medicine setting.[65] They performed WGS on a single individual for which the RBC groups had been previously established by clinically accredited methods. WGS data were compared and aligned with known reference sequences for blood group genes, following which the blood groups were deduced and compared with the clinically accredited results, with excellent concordance being observed. This study highlighted a number of further challenges which the transfusion community faces if it is to apply the new sequencing technologies effectively. These include; lack of consensus on reference gene sequences; the fact that many antigen reference sequences are 'virtual' composites that do not exist in humans, reference sequences continue to be improved by automated and manual curation, posing a risk that ISBT tables are outdated, ISBT allele nomenclature is restricted to variants that affect blood group phenotype, and blood group alleles with >2 variants are often not captured in their entirety.

In this chapter, we detail the creation of the BGC and the process by which some of the challenges mentioned above were met. In particular, we focus on the selection and curation of accurate fixed reference transcripts which allows for stable mapping of antigen encoding variants to the human genome. We go on to demonstrate the benefit of this mapping by using it to extract comprehensive antigen types and blood group allele variation information from WGS data of the NIHR BioResource WGS pilot project for the 100KGP.

3.3 Chapter Workflow

In order to extract transfusion relevant information from the vast genome-wide typing data that is currently being produced, we set out to further link the knowledge base of the transfusion medicine community with the modern human reference genome (see Fig. 3.1).

Realising this task would require input from experts in multiple fields we created the BGC with the aim to produce international standards for analysis of genotyping and sequencing data with respect to blood group antigens.

Our initial task was to perform an extensive data-driven review of the ISBT blood group gene reference sequences to address problems and inaccuracies identified in previous publications and to establish fixed reference sequences for the antigen encoding genes.

Using the newly curated sequences, we then created or updated Locus Reference Genomic (LRG) records for each gene and submitted this information to ISBT for review.[82] We also worked within ISBT to systematically curate the blood group reference tables - amending reference sequences where appropriate and linking all known antigen encoding variants to their unique dbSNP identifiers.

Finally, we show the benefits of linking antigen expression data to the human genome by analysing WGS data - First by extracting comprehensive antigen typing data from it, and then by using it to assess genetic variation in the Kell blood group system.

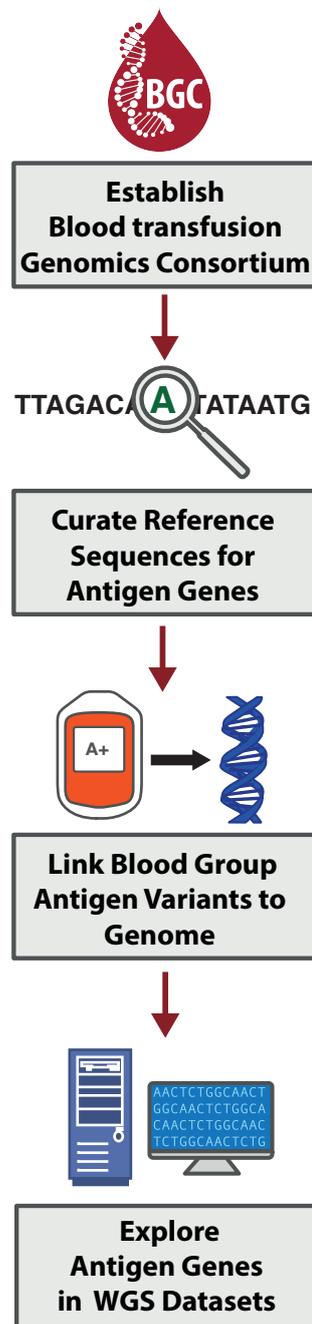


Fig. 3.1 Cartoon representing the overall workflow of this chapter. From top to bottom; first, the BGC was established to unite interested parties and define the standards which must be met for global adoption of donor genotyping. The BGC then set out to curate the reference sequences for each blood group encoding gene. Following reference curation, we worked alongside ISBT to link antigen encoding variants in their allele tables to the human reference genome. We then began to explore blood group-specific variation in WGS data.

3.4 The Blood transfusion Genomics Consortium

To ensure the high clinical standards required for safe transfusion practice are maintained as genotyping technologies are integrated into routine service, it is essential that international standards are set for management and interpretation of genotype data with respect to antigen expression. To this end, we established the BGC which has united experts from blood supply organisations, research institutions from around the globe and an industry partner with global reach (see Fig. 3.2).

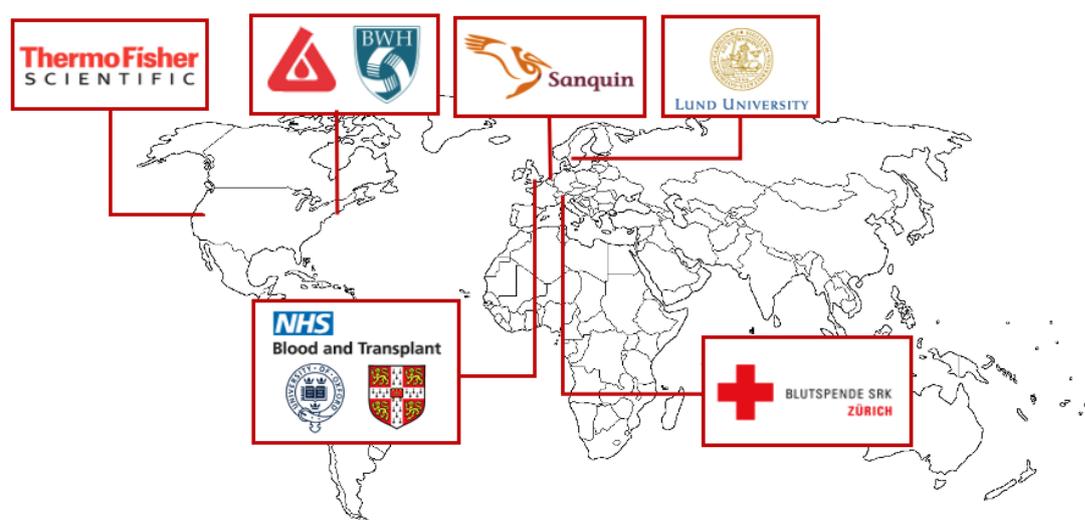


Fig. 3.2 Map of the world showing member organisations of the Blood transfusion Genomics Consortium. From left to right, Thermo Fisher Scientific (Santa Clara, USA), The New York Blood Centre (USA), The Brigham and Women's Hospital - Harvard University (Boston, USA), NHS Blood and Transplant, University of Oxford, University of Cambridge (UK), The Sanquin Blood Foundation (Amsterdam, the Netherlands), Lund University (Sweden), and Blutspende Zürich (Switzerland).

In 2016 the BGC held its first workshop meeting and several objectives were established based on a review of recently published studies. We now present the work that has been done toward the following selection of those objectives:

1. Review of the current reference sequences used by ISBT for antigen encoding genes.
2. Establish fixed and standardised reference sequences for each of the known red blood cell (RBC) antigen encoding genes.
3. Engage with ISBT to map variants contained in the blood group allele tables to genomic datasets.

3.5 Data driven review of ISBT reference transcripts

The ISBT WP has created a series of basic PDF tables which list the molecular changes in genes that give rise to blood group antigens and altered blood group phenotypes. Within each table a reference mRNA transcript is indicated against which the variation in each table is measured. Multiple studies have identified inaccuracies in the reference sequences used by ISBT, and no systematic evidence-based analysis has been performed to ensure that the current sequences are accurate.

In order to address this, we used RNA sequencing (RNA-seq) data generated by the BLUEPRINT Epigenomes of Blood Cells Consortium as a reference sequence curation tool.[83, 84] In this dataset, gene transcript and isoform expression <https://www.overleaf.com/project/5e16fb3> levels have been estimated by RNA-seq in multiple cell types from 90 cord-blood and adult-blood RNA samples.[80] We compared expression levels of each transcript in ENSEMBL90 for the 43 genes associated with blood group antigens by ISBT at the time of analysis in 2016 (an example of the expression data is given in Fig. 3.3).

We then compared the most expressed transcripts in BLUEPRINT to the reference transcripts listed in ISBT tables, suggesting amendment of the ISBT transcript where BLUEPRINT expression data provided no evidence for the presence of the original existing one (see Table. 3.1).

Table 3.1 ISBT transcripts compared with BLUEPRINT most expressed transcripts

Blood Group	Gene Name	ISBT reference transcript	Supported by RNAseq data	Suggested Transcript
ABO	ABO	NM_020469.2	Yes	-
MNS	GYPB	NM_002099.5	Yes	-
MNS	GYPB	NM_002100.5	Yes	-
MNS	GYPE	N/A	N/A	NM_198682.2
Rh	RHD	NM_016124.4	Yes	-
Rh	RHCE	NM_020485.4	Yes	-
Lutheran	BCAM	NM_005581.4	Yes	-
Lutheran	KLF1	NM_006563.3	Yes	-
KEL	KEL	NM_000420.2	Yes	-
Lewis	FUT3	NM_000149.3	Yes	-
Duffy	ACKR1	NM_002036.4	Yes	-
PIPK	A4GALT	NM_017436.4	Yes	-
Kidd	SLC14A1	NM_015865.6	Yes	-
Diego	SLC4A1	NM_000342.3	Yes	-
Cartwright	ACHE	NM_015831.2 and NM_001282449.1	Yes	-
Xg	CD99	NM_002414.3	Yes	-
Xg	XG	NM_175569.2	Yes	-
Scianna	ERMAP	NM_001017922.1	Yes	-
Dombrock	ART4	NM_021071.2	Yes	-
Colton	AQP1	NM_198098.2	Yes	-
Landsteiner Wiener	ICAM4	NM_001544.4	Yes	NM_001039132.2 (secondary)
Chido Rodgers	C4A	NM_007293.2	Yes	-
Chido Rodgers	C4B	NM_001002029.3	Yes	-
H	FUT1	NM_000148.3	No	NM_001329877.1
Kx	XK	NM_021083.2	Yes	-
Cromer	CD55	NM_000574.3	Yes	-
Knops	CR1	NM_000573.3 and NM_000651.4	Yes	-
Indian	CD44	NM_001001391.1	Yes	-
Ok	BSG	NM_198589.2	Yes	-
Raph	CD151	NM_004357.4	Yes	-
John Milton Hagen	SEMA7A	NM_003612.3	Yes	-
I	GCNT2	NM_145655.3	Yes	-
Globoside	B3GALNT1	NM_033169.2	Yes	-
Gerbich	GYPB	NM_002101.4	Yes	-
Gill	AQP3	NM_004925.4	Yes	-
Rh-associated glycoprotein	RHAG	NM_000324.2	Yes	-
Forsman	GBGT1	NM_021996.5	Yes	-
Junior	ABCG2	NM_004827.2	Yes	-
LAN	ABCB6	NM_005689.2	Yes	-
VEL	SMIM1	NM_001163724.2	Yes	-
CD59	CD59	NM_203330.2	No	NM_000611.5
AUG	SLC29A1	NM_001304463.1	Yes	-

Through this analysis we discovered inappropriate reference transcripts had been chosen for three blood group antigen encoding genes, *GYPE*, *FUT1*, and *CD59* of the MNS, H, and CD59 blood group systems, respectively.

For *ICAM4* (Landsteiner Wiener) evidence of a secondary transcript, NM_001039132.2, was observed - this transcript contains a frameshift mutation changing the last 140 amino acids (aa) of the protein and may be important to antigen expression.

Expression of *CD44* (Indian) transcripts was highly tissue specific. We confirmed that the reference transcript NM_001001391.1 was the highest expressed transcript in Erythroblasts.

For *ABCB6* (Lan) reads supporting a secondary transcript, ENST00000295750, were observed. This transcript was only present in the Ensembl database, with RefSeq only containing a single transcript for this gene, NM_005689.2.

GCNT2 (I) and *B3GALNT1* (Globoside) were not covered at sufficient read depth in the BLUEPRINT data, further experimental work is required to assess these genes.

Through this curation analysis, we also found out that the ISBT WP reference transcript for *SLC29A1* (Augustine) does not correspond to the reference protein. NM_001304463.1 is listed in the ISBT WP reference table with a protein length of 456 aa but the RefSeq entry linked to the transcript is 498 aa.

Results from this analysis were then used as a base to select fixed reference transcripts for each gene via creation of LRG records. An LRG record contains stable reference sequences that are used for reporting variants with clinical implications - in translation, this means that genotyping data from any technology can be mapped to a stable LRG record allowing genetically inferred blood types to be standardised globally.

Finally LRG records, detailing changes to the 'canonical' reference transcript, were then submitted to the individuals responsible for curating each ISBT WP allele table for their approval. As a result of this work 42 of the 43 blood group genes analysed now have approved LRG records, only *GYPE* is waiting for analysis by LRG curators.

In 2019, following engagement with the BGC and other genomics focused groups, the ISBT WP begun associating the antigen encoding variants to the genome by updating their reference sequences to the RefSeq equivalents of the LRG records and also by addition of dbSNP identifiers for each variant. With this development, the transfusion medicine community is now positioned to begin analysing the vast amount of genomic data available to it.

3.6 Exploring whole-genome sequencing data

We next sought to demonstrate the value of having standardised this knowledge by analysing WGS data.

3.6.1 Further validation of the bloodTyper WGS algorithm

As previously mentioned, in 2016 Lane and colleagues first demonstrated that accurate and comprehensive antigen types could be inferred from WGS data. This was done by using the purpose built bloodTyper software to infer blood types for a single genome. In 2018 a larger scale validation study was carried out in which bloodTyper was trialled using 30x WGS data from 110 American blood donors enrolled in the MedSeq project and 15x WGS data from 200 English donors enrolled in the INTERVAL trial.[50, 21] An initial subset of 20 MedSeq samples were analysed using bloodTyper and antigen typing results compared to those on the donor records. Concordance was 99.5% with 1,194 results correct in 1,200 comparisons across 38 RBC and 22 HPA antigens. Discordance was observed for ABO in two donors and for the RH system in four donors, three for C and one for D.

Following analysis of each discordant result, improvements were made to the bloodTyper analysis software and the remaining 90 MedSeq samples were analysed. Concordance between WGS inferred results and donor record typing was 99.8% with 5,390 results correct in 5,400 comparisons across 38 RBC and 22 HPA antigens.

Finally, bloodTyper was trialled in a blinded fashion using 15x WGS data from 200 INTERVAL participants. Concordance between WGS inferred and donor record antigen types was 99.2%, with 3486 correct results in 3515 comparisons across 21 RBC antigens. Analysis of discordant results revealed that the majority of errors were due to the low 15x average coverage of the genome data for the INTERVAL participants. In most cases the correct antigen defining nucleotides were detected but at variant positions with a coverage < 4x, bloodTyper does not infer blood types. In particular, the low sequencing coverage of the genomes of INTERVAL participants caused difficulties when typing for the M antigen, with some genomes having the antigen encoding region of the *GYP A* gene covered with an average coverage of 1x and even 0x read depth.

Based on these encouraging results we now consider bloodTyper validated for use on 30x WGS data, and advise caution when genomes have been sequenced to a lesser depth.

A full description of this validation experiment, including an in-depth discussion of discordant results, is given in: W.J. Lane, C.M. Westhoff, N.S. Gleadall, et al., (2018). “Automated typing of red blood cell and platelet antigens: a whole-genome sequencing study”. *Lancet Haematology*. 2018 Jun; 5(6): e241–e251.

3.6.2 Antigen typing 13,037 whole genomes

Following validation of the bloodTyper WGS algorithm, we used it to infer antigen types for the 13,037 participants enrolled in the NIHR BioResource WGS study for rare diseases patients and their close relatives.[60] This study was one of the pilot studies for the 100KGP and WGS was performed to clinical standard. The vast majority of samples were analysed by WGS at 150 bp read length, but samples sequenced early in the project have 100 bp and 125 bp read length.[60] Individuals in this study, which covered 15 distinct rare diseases domains, were recruited from 57 National Health Service (NHS) hospitals in the United Kingdom and 26 hospitals in other countries.

Demonstrating the value of typing blood groups from WGS data generated on DNA samples obtained from patients, we identified 144 study participants with rare antigen-negative phenotypes. If such patients would have formed alloantibodies against the cognate antigen then it will be challenging to identify suitable donors, should they require transfusion support in future (see Table. 3.2); 22 of these individuals have been diagnosed with a bleeding, thrombotic or platelet disorder and the likelihood of this category of patients requiring transfusion is increased.

Table 3.2 Patients in the NIHR BioResource WGS project with rare antigen negative phenotypes

System	Phenotype	Patients Identified
CO	Co4-	4
MNS	U-	1
MNS	S- s- U _{weak} He-	1
RH	Rh26LOCR-	1
RH	CEAG-	5
RH	Rh _{null}	1
LU	Lu ^b -	12
LU	Lu8-	2
LU	Lu13-	10
KEL	k-	20
KEL	Kp ^b -	1
KEL	Js ^b -	2
YT	Yt ^a -	24
SC	Sc1-	1
DO	Hy-	1
DO	Jo ^a -	5
CO	Co ^a -	19
KN	Kn ^a -	7
KN	McC ^a -	17
KN	Sl3-	8
VEL	Vel-	2

One patient in this cohort, a young female from Belgium with severe haemolytic anaemia, provided a particularly interesting example of the added value WGS brings. She had been serologically typed as Rh_{null} (D-, C-, c-, E-, e-) and the Rh proteins were lacking from her RBCs. A complete absence of the protein of the RH complex is known to be causally associated with haemolytic anaemia and this inherited condition is incredibly rare, with only 43 Rh_{null} individuals being reported to the ISBT WP since the discovery of the phenotype in 1961. Investigative genotyping focussed on the coding fraction of the *RHD*, *RHCE* and *RHAG* genes performed at the Institut National de Transfusion Sanguine (INTS) in Paris failed to identify the molecular basis of her disorder. Freely translated and in short, the INTS report states: 'Using targeted sequencing of the coding fraction and flanking sequences of the *RHD* and *RHCE* genes no variants with possible causality were identified. This prompted the sequencing of the coding fraction and flanking sequences of the *RHAG* gene which resulted in the identification of the splice site variant LRG_822:c.157+1g>a on a single allele. Further

sequencing is required to identify a putative other causal variant on the alternate allele of the *RHAG* gene. Results of this test will follow in due course"

Mutations of the *RHAG* gene are known to underpin the Rh_{null} phenotype if present in homozygosity or in compound heterozygosity with another nonsense *RHAG* variant (see Fig. 3.4). The INTS only identified a single causal variant in *RHAG*.

15. JUN. 2017 16:53 CNRGS N° 325 P. 2



INSTITUT NATIONAL DE LA TRANSFUSION SANGUINE
Laboratoire de Biologie Médicale
Dr Syria Laperche – Biologiste responsable
Unité d'Immuno-Hématologie Spécialisée
Centre National de Référence pour les Groupes Sanguins
Dr Thierry Peyrard Dr Jérôme Babinet Dr Joëlle Nataf Dr Vincent Thonier

Suite des résultats d [REDACTED]
Prélèvement numéro : [REDACTED]
Référence courrier : CR 28504/TP

AVIS - INTERPRETATION

Nous avons reçu des échantillons sanguins de ce sujet pour confirmation d'un phénotype exceptionnel de type Rh_{null}.

L'étude phénotypique réalisée au CNRGS confirme le phénotype RH:-1,-2,-3,-4,-5.

Le séquençage des gènes *RHD* et *RHCE* (10 exons et régions flanquantes) n'a révélé aucune anomalie (présence d'un gène *RHD* sans mutation et génotype *RHCE*^{*ce}/*RHCE*^{*cE}).

Compte tenu de ces résultats, l'implication d'une altération du gène *RHAG* conduisant à un phénotype Rh_{null} de type régulateur apparaît hautement probable. Les résultats de séquençage NGS que vous nous avez transmis confirment la présence d'une mutation du gène *RHAG* mais à l'état hétérozygote (c.157+1g>a, *RHAG**01N.03). Il est donc probable que la mutation sur le deuxième allèle *RHAG* n'ait pas été détectée lors du séquençage à haut débit (problème d'"allele dropout" ou de délétion impliquant un nombre significatif de nucléotides) ; il importerait donc de séquencer intégralement le gène *RHAG* avec une méthode alternative. Cette étude sera prochainement réalisée au CNRGS mais les résultats ne seront probablement pas disponibles avant le courant du mois de septembre 2017.

Fig. 3.4 INTS genotyping report of Rh_{null} patient

bloodTyper was used to analyse the WGS data for this patient and we confirmed the INTS finding of the presence of the LRG_822:c.157+1g>a Rh_{null} variant on a single allele of *RHAG*. However, inspection of the raw sequence alignment at the *RHAG* loci by BGC members revealed the presence of a large SV in this individual (see Fig. 3.5). SV callers Manta v1.6.0 and Canvas v1.40.0 were then used to characterise the variant. Both callers identified a tandem duplication at GRCh37 chr6:49575934-49588875.

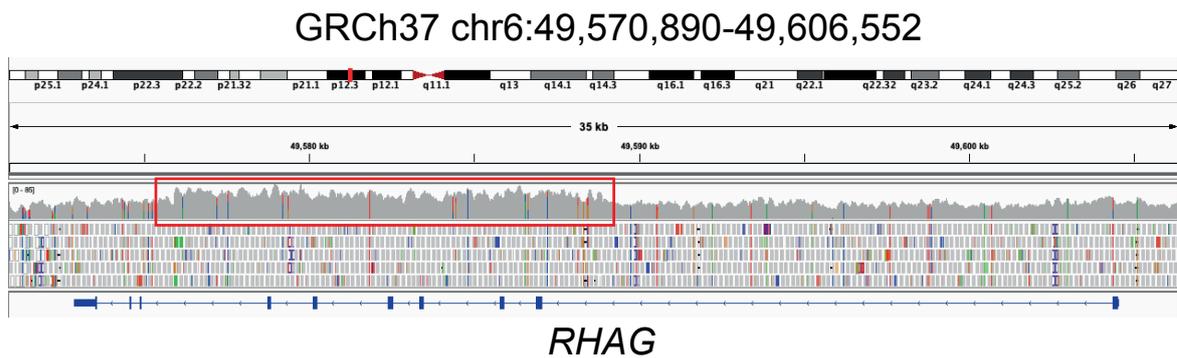


Fig. 3.5 WGS sequence alignment at *RHAG* loci of the Rh_{null} patient. The gene structure of *RHAG* is shown in blue at the bottom of the figure, thick and thin sections represent exons and introns, respectively. A red box is drawn around an area of increased read depth and area of improperly mapped reads indicating presence of a structural variant.

Confirmatory Sanger sequencing was performed for both variants on parental DNA samples in order to assess the mode of inheritance. It was revealed that the LRG_822:c.157+1g>a splice site variant was of maternal origin and the inline duplication had been inherited from the father (see Fig. 3.6). The identification of the two variants is highly likely causally associated with the non-functional *RHAG* gene in this patient. A non-functional *RHAG* gene results in the RHD and RHCE proteins being unable to incorporate into the RBC membrane causing a Rh_{null} phenotype. The results obtained in this patient show the power of WGS and the genetic results can now be used for family planning purposes.

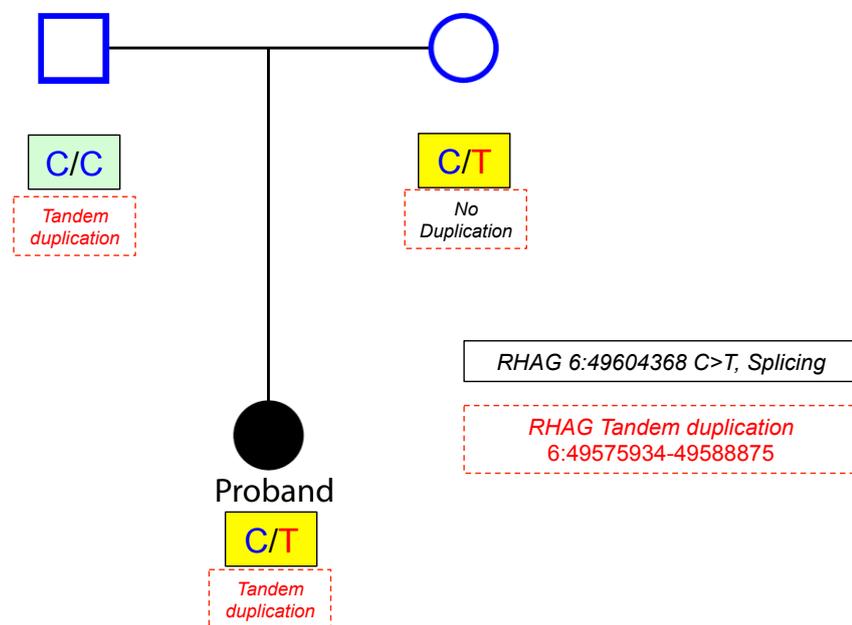


Fig. 3.6 Pedigree of Rh_{null} patient with Sanger sequencing results depicted. (**black box**) shows the genotype of the *RHAG* chr6:49604368c>t known Rh_{null} causing variant. (**red-dashed box**) shows heterozygous presence of the *RHAG* tandem duplication chr6:49575934-49588875.

3.6.3 Variation in the Kell blood group system

We then used the high coverage WGS data from the 13,037 NIHR BioResource individuals to assess the extent to which ISBT WP has underestimated variation in antigen encoding genes via the adoption of a retrospective and incomplete approach to recording haplotype data. For the purposes of this thesis we have used the Kell blood group system as an example locus.

First, 10,1049 variants with an overall pass rate (OPR) of >0.98 were extracted from merged variant call files (VCF) for all samples in a 2 Mb window centred around the *KEL* gene (GRCh37 chr7:141638200-143659802) using bcftools v1.9. Genotypes for all variants in each sample were then phased using Eagle v2.1.9. This was performed without a phasing reference panel as Eagle2 will drop variants not present in the reference panel resulting in loss of sequence information. Phasing of genotypes is an essential step when trying to recreate haplotypes for antigen encoding genes, as the presence of nonsense (or null in blood group terminology) variants has a pronounced effect on overall blood group phenotype (see Fig. 3.7). Hence it is important to determine in which haplotype the nonsense variant is localised.

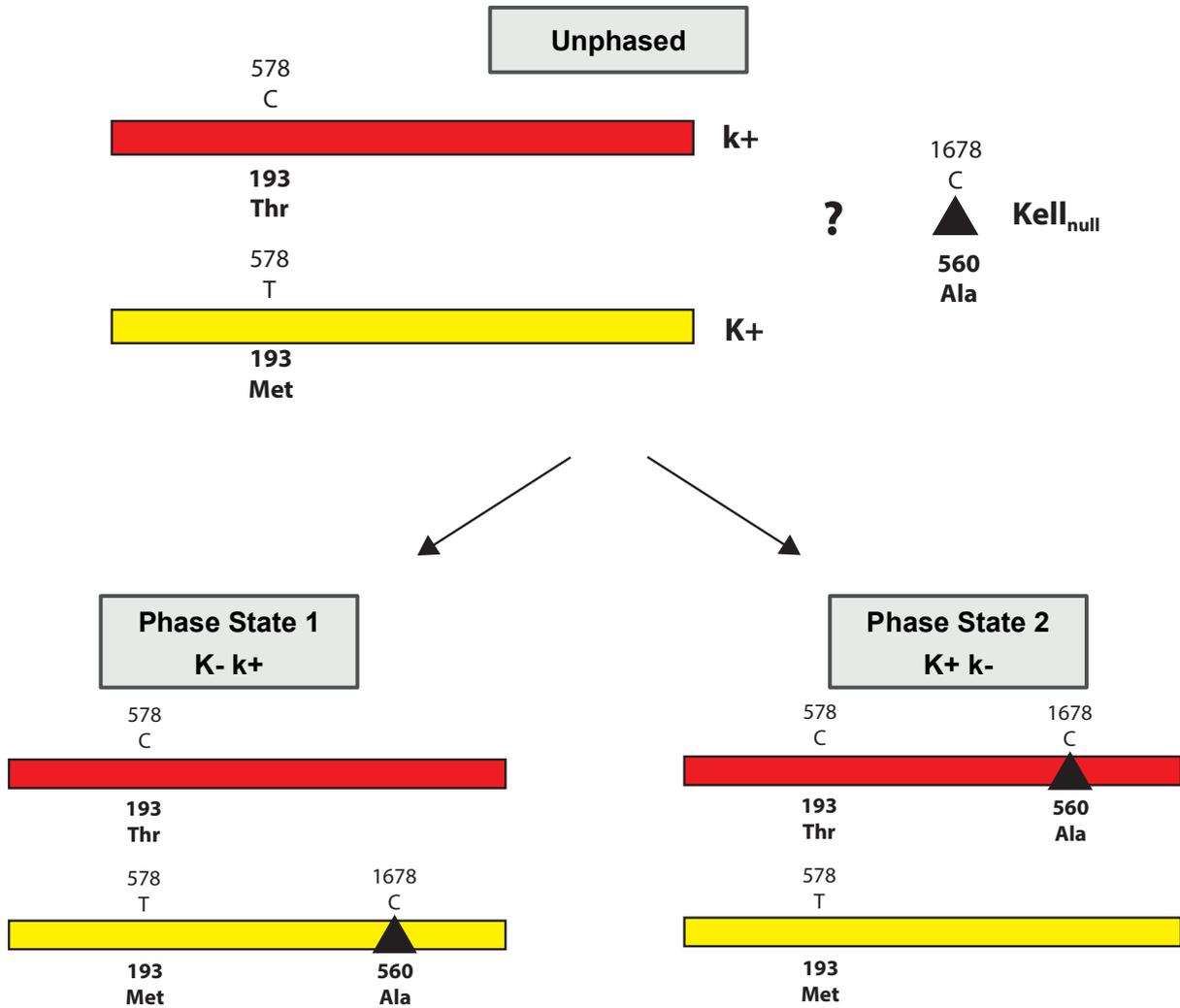


Fig. 3.7 Cartoon showing the importance of phasing when reconstructing antigen encoding haplotypes. (**Unphased**) Variation detected at the *KEL* locus of a theoretical individual. Coloured blocks represent 'background' haplotype sequences with phenotype defining cDNA changes above each block and resulting amino acid changes and its position below. Here the individual has both a K-/k+ phenotype allele shown in **Red**, and a K+k- allele shown in **Yellow**. A **Kell_{null}** phenotype variant has also been detected and is represented by a **Black Triangle**. (**Phase State 1**) Here the **Kell_{null}** variant is in phase with the K+ allele, nullifying it, and resulting in an overall K-k+ phenotype. (**Phase State 2**) Here the **Kell_{null}** variant is in phase with the k+ allele, nullifying it and resulting in an overall K+k- phenotype. As can be seen, the phase of the **Kell_{null}** allele has a pronounced effect on overall phenotype

We then created two sets of *KEL* gene consensus sequences for each individual, with two haplotype sequences in each set, by using phased variant genotype data to modify the *KEL* reference sequence (LRG_799).

The first set of consensus sequences was constructed only using genotypes for known antigen encoding variants contained in the ISBT WP Kell system table. These sequences can be thought of as 'ISBT reference haplotypes' and serve to capture the known antigen encoding variation of each individual. We identified 35 unique ISBT haplotypes in the 13,037 whole genomes and bloodTyper was used to compute phenotype and ISBT allele codes for each of them (see Fig. 3.8). Interestingly, 7 of the 35 haplotypes were novel combinations of ISBT Kell variants that have not yet been given ISBT allele codes. One individual possessed a potentially new K_{null} allele composed of a *KEL**02.01 (k+) and a premature stop variant (LRG_799:p.Arg700Ter).

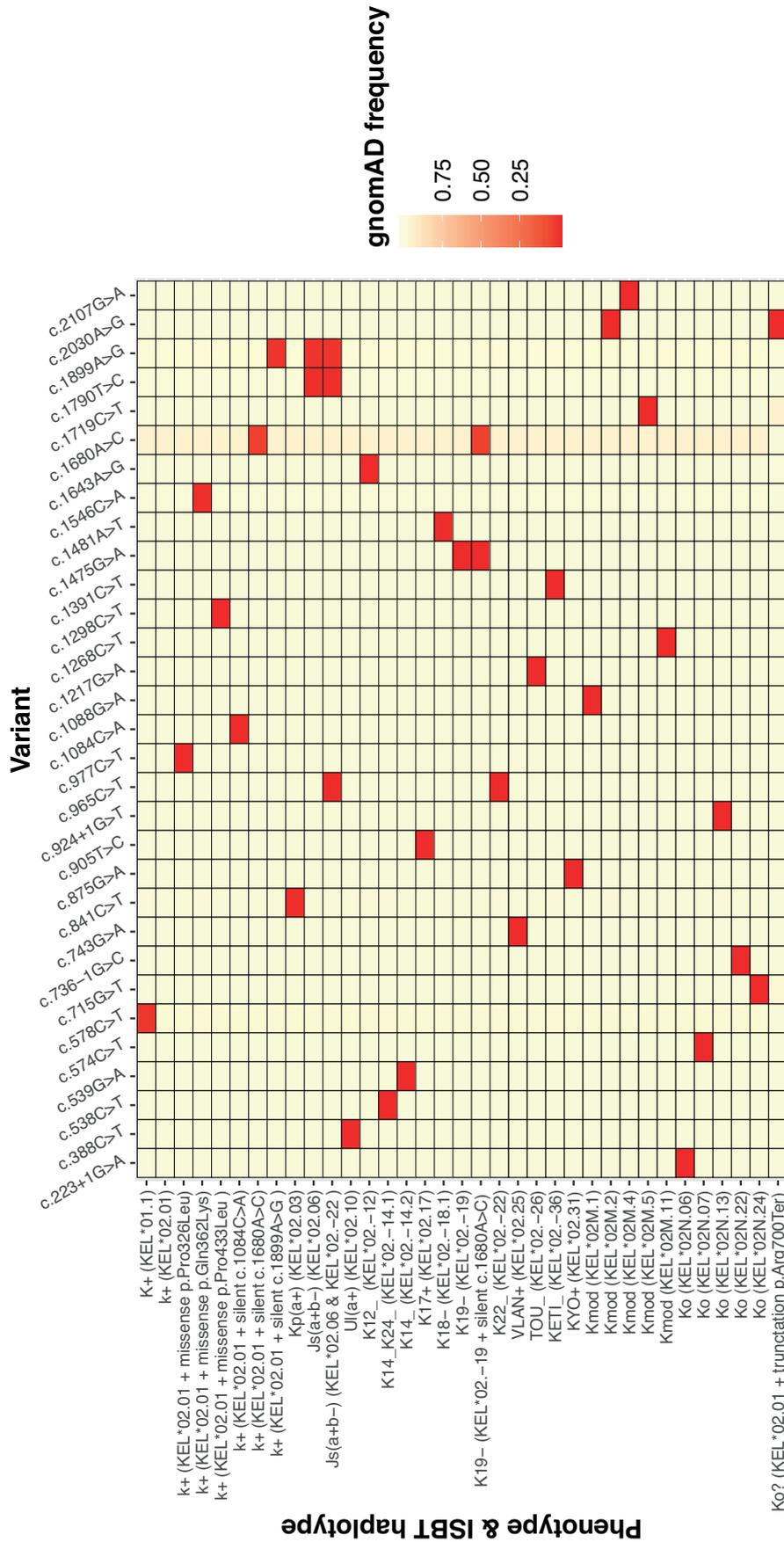


Fig. 3.8 Heatmap showing the ISBT *KEL* haplotypes identified in the WGS data of 13,037 individuals. Each row represents one of 35 unique haplotypes identified, overall phenotype and ISBT allele nomenclature for each is annotated on the Y-axis. Columns represent the ISBT listed variants which were used to reconstruct each haplotype. Colour represents gnomAD minor allele frequency of the genotype at each position in each sequence, with **Red** and **Yellow** representing the rare and common end of the frequency spectrum, respectively. Rare variant positions can be thought of as haplotype defining.

The second set of consensus sequences was constructed using genotypes from all variants identified within the *KEL* gene, these are 'complete haplotypes' reflecting all observed variation in the gene for each individual. We identified 1,732 unique complete haplotypes in the same 13,037 whole genomes, a nearly 50 fold increase from the number identified when using ISBT variants alone (see Fig. 3.9). Again, bloodTyper was used to compute antigen phenotype and ISBT allele codes for each of them; 12 of the haplotypes identified contained a nonsense variant leading to a premature stop before the K/k antigen defining polymorphic position (LRG_799:c.578). The Kell protein is a type II transmembrane protein and therefore some of the shorter isoforms encoded by the 12 haplotypes containing a nonsense variant may be expressed on the RBC membrane, however, due to the presence of these truncating variants the K/k antigenic site will not be present. The results of this analysis show that antigen encoding variation within the *KEL* system has been grossly underestimated by the ISBT WP. Inclusion of these Kell_{null} alleles in the ISBT reference tables is critically important for patient blood group genotyping.

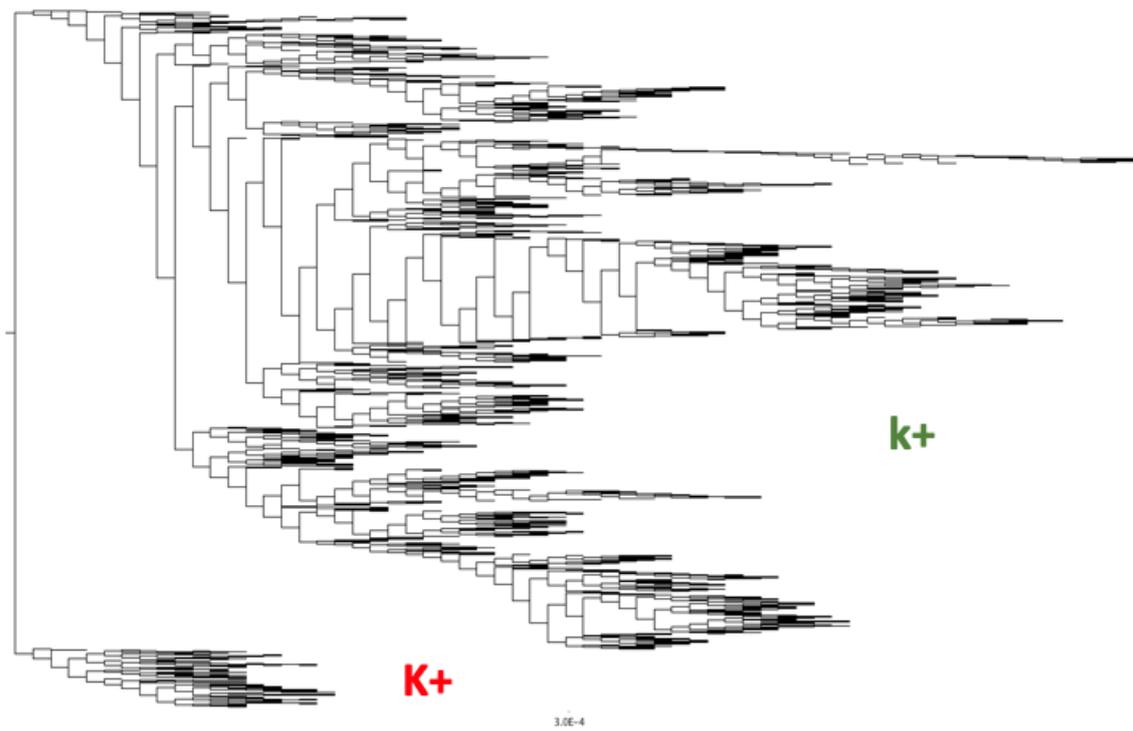


Fig. 3.9 Phylogenetic tree showing the complete variation in the *KEL* gene sequences of 13,037 individuals. Each branch represents one of 1,732 unique *KEL* sequences identified. The tree is rooted on the LRG_799:c.578C>T variant which defines K+k- or K-k+ phenotype status. Clusters are annotated with K/k phenotype.

3.7 Discussion

Over the past decade, WES and WGS sequencing data from nearly 0.5 M research participants has been made publicly available. Based on the results of the pilot study of the 100KGP, the NHS has taken the decision that WGS will become the standard of care for rare disease and cancer patients. Implementation of this new approach is now rolled out via the reconfigured 7 NHS Clinical Genomics laboratories. In this chapter, we have shown that the WGS data can be used to infer the blood cell antigen status of both donors (the INTERVAL WGS study) and patients (the NIHR BioResource WGS study). The transfusion medicine community must act on this information and leverage this comprehensive data to further its knowledge of blood group genetics and to improve health outcomes for patients. In order to achieve this two things must first happen; 1) The knowledge of blood group antigen genetics must be carefully curated and computationally mapped to the human reference sequence, and 2) Global standards for the interpretation of genotype data with respect to antigen typing must be established in order to ensure the safe integration of genomics technologies into the clinical typing laboratory.

Using the RNA-seq expression data produced by the BLUEPRINT study we have delivered the first data-driven review of the ISBT WP blood group gene reference sequences. In line with the results of previous studies, we found that some of the historical references selections were inadequate. For three loci we have recommended to the ISBT WP that new reference transcripts are adopted as no supporting data for original transcript expression could be found in the BLUEPRINT data. Additionally, at six loci we found evidence of moderately expressed secondary transcripts or highly tissue-specific expression of transcripts; in these cases, our data provide the first confirmation that the ISBT transcripts are the most appropriate and most abundantly expressed. Most importantly, this analysis allowed us to work with the EBI to extensively curate and establish fixed LRG records for each of the antigen encoding genes. Fixed reference sequences that do not change with genome build updates provide a standard to which genotype information produced by any technology can be mapped.

We communicated the results of our analysis to the ISBT WP and in turn, they have now updated the reference tables to use the RefSeq equivalents of the LRG records defined here. Through continued engagement with the ISBT WP we were able to use these references to map each known antigen encoding variant to dbSNP identifiers, cementing their position in the human reference genome. This development, in combination with the earlier work done by Lane et al., in 2015 and Moller et al., in 2016, means the transfusion medicine community is now in a position to explore and use the vast amount of WES, WGS and

array genotyping data, which will become available as clinical genomics is introduced across healthcare systems.

In the second part of this chapter, we demonstrated the value of having a carefully curated and standardised link between antigen phenotype and genotype by using it to analyse WGS data.

First, we used WGS data from 310 American and English blood donors to further develop and validate the bloodTyper algorithm. Overall excellent concordance between genotype and phenotype was observed with 10,070 concordant results in 10,115 comparisons (99.5% concordant). The bulk of discordant results (64%, 29 of 45) in this analysis were caused by poor coverage of antigen encoding genes in the lower 15x depth sequencing for donors who participated in the INTERVAL trial highlighting the importance of using clinical quality genomes (30x) for inference of antigen status from WGS data.

We then used bloodTyper to produce comprehensive antigen typing data for 13,037 rare disease patients and their close relatives enrolled in the NIHR BioResource project for whom 30x clinical-grade WGS data was available. We identified 144 individuals with rare antigen-negative blood group phenotypes who may be difficult to provide blood for should they need transfusion support and shown to have formed alloantibodies. This information is particularly relevant for 22 of these individuals who have each been diagnosed with a bleeding, thrombotic or platelet disorder as this category of rare disease patients has a higher likelihood of requiring transfusion. For one patient, analysis of her WGS data elucidated the genetic cause of her haemolytic anaemia - a fine example of how WGS data provide a higher sensitivity than targeted sequencing to detect the genetic basis of rare inherited disorders.

Finally, we used the 30x WGS data of the 13,037 patients and their close relatives to determine the extent which variation in the blood group encoding genes has been underestimated by the transfusion medicine community. Previous work has highlighted the problem, reporting that only 19% of the variation observed within blood group genes in the HapMap and 1KGP datasets are reflected in the ISBT blood group allele tables. By phasing variant genotypes and reconstructing sequences that captured variation across the entire *KEL* gene locus we identified 1,697 (50x) more haplotypes than when performing the same analysis using only variants in the ISBT WP reference tables. Importantly, in the 35 'ISBT haplotypes' we identified, several were novel combinations of ISBT WP recognised ones, which have not yet been given ISBT WP allele codes. This means that although the inclusion of fixed sequences and dbSNP identifiers in the reference tables was an important step in the right direction, the transfusion medicine community must use high coverage WGS datasets such as those made available by the NIHR BioResource and the 100KGP to produce fully phased haplotype references for each of the blood group systems. Such an improved catalogue of

reference haplotypes would substantially increase the accuracy of genotype phasing, which is essential when considering the effect of nonsense (null) variants on overall antigen phenotype and that antigens within the same system are not inherited separately but travel together on haplotypes. In the next step, the catalogue must be enriched for haplotypes present in individuals of non-European ancestry, this will become feasible in the near future as a large number of WGS projects have been initiated across the world, as reviewed in the introduction to this chapter and of this thesis.

This chapter provides an example of how integrating the transfusion community knowledge with a detailed understanding of genomic data can result in the development and validation of tools such as bloodTyper. We have used the genotype-phenotype link and WGS data to not only further our understanding of the molecular basis of antigen expression, but also to increase the accuracy of inferring antigen types of patients and donors from genotype data. The BGC and ISBT WP have now taken the first steps towards producing the international standards required for analysis of genomic data with respect to blood group antigens, setting the scene for an era of precision transfusion medicine and genomics guided development of clinically accredited blood group genotyping tests which will be discussed in the next chapter.

Chapter 4

A universal donor genotyping platform

4.1 Abstract

In this chapter, we present a genotyping platform that is capable of typing almost all clinically relevant Red Cell, Platelet and Leukocyte antigens of an individual. The platform is composed of two pieces of technology; 1) an array-based genotyping assay that has been designed to interrogate all known genetic variants which underpin blood type antigen expression and 2) the bloodTyper automated analysis software for the interpretation of blood groups from genotyping data. These two components were developed in tandem using an iterative design process. A typical development cycle involved; genotyping DNA samples from real blood donors, interpretation of genotype data using bloodTyper, comparison of genotype inferred and blood service determined antigen types, and finally improvements to array content or the bloodTyper algorithm guided by concordance analysis. Following two initial rounds of development with a small number of samples, a validation study was performed using a larger number of samples from 7,927 European, 27 South Asian, 21 East Asian and 9 African ancestry blood donors enrolled in two national biobank studies in England and the Netherlands.

In addition to detailing the development and validation process, we show how the extremely dense antigen typing data produced can be used to simplify the challenges of supplying blood at national scale for patients with antibodies against several red blood cell antigens. This was done by extracting antibody data from NHS Blood and Transplant patient records over a five year period and comparing the number of potentially compatible individuals that could be found using genotyping or current donor record typing data available for the 7,984 validation set donors. We further demonstrate the utility of the platform by typing Human Platelet Antigens (HPA) and Human Leukocyte Antigens (HLA) and by

querying the genotyping data produced for the presence of variants linked to donor health - all using the same affordable test.

The description of work in this Chapter is based on that given in N. Gleadall, B. Veldhuisen, J.Gollub, et al., (2020). “Development and validation of a universal blood donor genotyping platform: a multinational, prospective study”. *Blood Advances*, 2020

4.2 Introduction

The Alliance of Blood Operators collects 31 million units of blood each year which is used to provide life-saving support to an estimated 15 million individuals with a wide range of medical conditions.[85] Currently, it is common practice to match red blood cells (RBC) for only the ABO and Rh blood groups.[86] While this policy ensures transfusion safety and prevents the majority of fatal haemolytic transfusion reactions (HTRs), sensitisation to non-self RBC antigens remains an unavoidable consequence of this matching strategy.

Annually, an estimated 3% (0.5 million) of patients become sensitised to RBC antigens after a single transfusion episode, with 60% of patients receiving regular transfusions becoming immunised.[87–91, 63, 92] Multiple sensitisation events, or sensitisation due to absence of high frequency antigens can render haemoglobinopathy patients un-transfusable due to lack of available compatible donors. Sensitisation can also cause haemolytic disease in pregnancy, which is potentially life-threatening to the foetus and newborn. Notwithstanding these serious side effects, the introduction of a more-precise matching policy is resisted because of perceived logistical challenges and the cost of typing thousands of blood donors for all antigens.[93]

Antibody-based typing tests are currently the gold standard for RBC antigen typing however, reliable reagents and high-throughput techniques are not available for all clinically relevant antigens. It is for these reasons that donor typing data is so incomplete, even for commonly typed, clinically relevant antigens such as those of the Duffy, Kidd, MN/Ss, and Lutheran systems. DNA-based tests have been used by blood services to overcome these limitations and a plethora of “in-house” and some commercial assays have been developed for specifically donor genotyping.[94, 95, 36] These range from single variant PCR based assays to typing arrays such as the BioArray™, HEA BeadChip™ from Immucore which covers a wider range of different antigens. Studies using these assays in combination with antibody-based typing have shown that antigen-negative blood could be supplied for 99.8% (5661/5672) complex blood requests to a US blood bank using 43,066 donors genotyped for only 16 common antigens and antibody types for rarer phenotypes such as Vel-negative.[38] Multiple small-scale clinical studies have shown that

the higher resolution typing data provided by genotyping assays can improve clinical care for transfusion-dependent haemoglobinopathy patients, particularly when used to facilitate prophylactic matching for variant RH antigens.[96, 97]

Despite the growing evidence which supports their use, antigen genotyping assays have not been introduced into routine service by the global blood banks. The main reasons for this lack of uptake are; the cost of current genotyping assays, the fact that no existing test can type all clinically relevant RBC antigens, and the lack of an algorithm for automated interpretation of results. Furthermore, existing tests do not include typing for other transfusion relevant antigens such as HLA and HPA, which are required for supporting cancer patients refractory for random donor platelets because of anti-HLA and/or anti-HPA antibodies.[98, 99]

In previous studies, we have demonstrated that whole-genome sequencing (WGS) (see Chapter 3) and whole exome sequencing (WES) methodologies can be used to produce comprehensive and accurate RBC and HPA antigen typing data.[65, 21, 100] Although sequencing-based tests are becoming the standard of care for transfusion-dependent patients, these assays remain too expensive for typing vast numbers of blood donors. The unbalanced use of genotyping technologies has created a situation in which some patients have extensive high-resolution antigen typing data that cannot be used to improve their transfusion care as only a relatively small number of extensively typed donors are available, which can be used for extended matching. This issue of limited availability of typing data is even more pressing for haemoglobinopathy patients receiving regular exchange transfusion as their requirement can be as high as 10 units of blood per procedure, on a fortnightly basis. To remedy this situation blood services endeavour to increase enrolment of donors of African ancestry and ideally all these donors are to be typed with a comprehensive and affordable genotyping test. To remedy this situation and improve clinical care, an affordable and comprehensive genotyping test is required.

A universal donor typing platform must be able to identify all clinically relevant RBC antigens for blood transfusions and HLA and HPA for platelet transfusions. The physical test must be combined with software for automated data interpretation and formatting so that it is immediately usable by blood supply organisations. Importantly, the platform must be cost-effective and scalable to millions of donors and patients. In 2016 the Blood transfusion Genomics Consortium (BGC) was established to capitalise on array technology recently applied in studies to genotype thousands of individuals worldwide.[47, 50, 49] In this chapter, we describe the validation of a high-throughput genome-wide test re-purposed for extensive blood donor antigen typing that is available at a cost of approximately £30 per sample, inclusive of equipment, labour cost and analysis.

4.3 Chapter workflow

Our donor genotyping platform consists of two interrelated components: a high-throughput genotyping assay that generates DNA typing data, and automated analysis software for antigen typing. These two components were developed in parallel, using a strategy of iterative improvement and evaluation using DNA samples and data from currently active blood donors. The project can be thought of as having three main stages (see Fig. 4.1).

Stage 1 - involved assessing the suitability of the Axiom platform for donor genotyping. Here we sought to conduct a 'proof of principle' experiment to show that an array designed for population screening could also be used for donor typing. To do this we inferred the blood antigen types of 1,057 donors using pre-existing genotype data and compared our results to NHSBT donor records. The samples and data used for the analysis were from NHSBT donors enrolled in the INTERVAL randomised controlled trial. The DNA samples from the donors were typed with the original UK Biobank Axiom Array (UKBBv1 array), capable of typing approximately 800,000 DNA variants for 96 individuals in a single test cycle. It is important to note that while this version of the array contained content for typing some RBC antigen controlling variants, it was not designed with comprehensive donor RBC antigen typing in mind.

Stage 2 - involved re-designing the array for donor typing. Following the encouraging results of our initial analysis (stage 1), a donor typing only array was designed based on a smaller testing format capable of typing 50,000 DNA variants for 384 individuals in a single test cycle. This platform was called the 384HT Axiom Blood Typing SNP Screen Array (384 Blood Typing Array). This included novel probes for typing 1,602 antigen typing variants which were identified via consultation with experts in immunogenetics. In order to assess typing accuracy of the novel 384 Blood Typing Array design, we genotyped 507 donors enrolled in the NIHR BioResource, inferred blood types from the data generated, and compared these to existing blood types in NHSBT's electronic donor records. The main reasons for performing this smaller-scale trial were; to identify any manufacturing errors that might have been made in the array design and fabrication process, examine raw data from novel probesets to assess the quality of their design, and provide validation data for checking software accuracy and further development of the bloodTyper software. **Stage 3** - involved performing a large scale array validation experiment. Guided by the results from previous stages, the next round of array optimisation was performed and new antigen typing content was integrated into the original UKBBv1 array design. This new array was called the UK Biobank version 2 Axiom Array (UKBBv2 array). We then trialled the UKBBv2 array using DNA samples from 7,477 English and Dutch blood donors enrolled in the COMPARE and

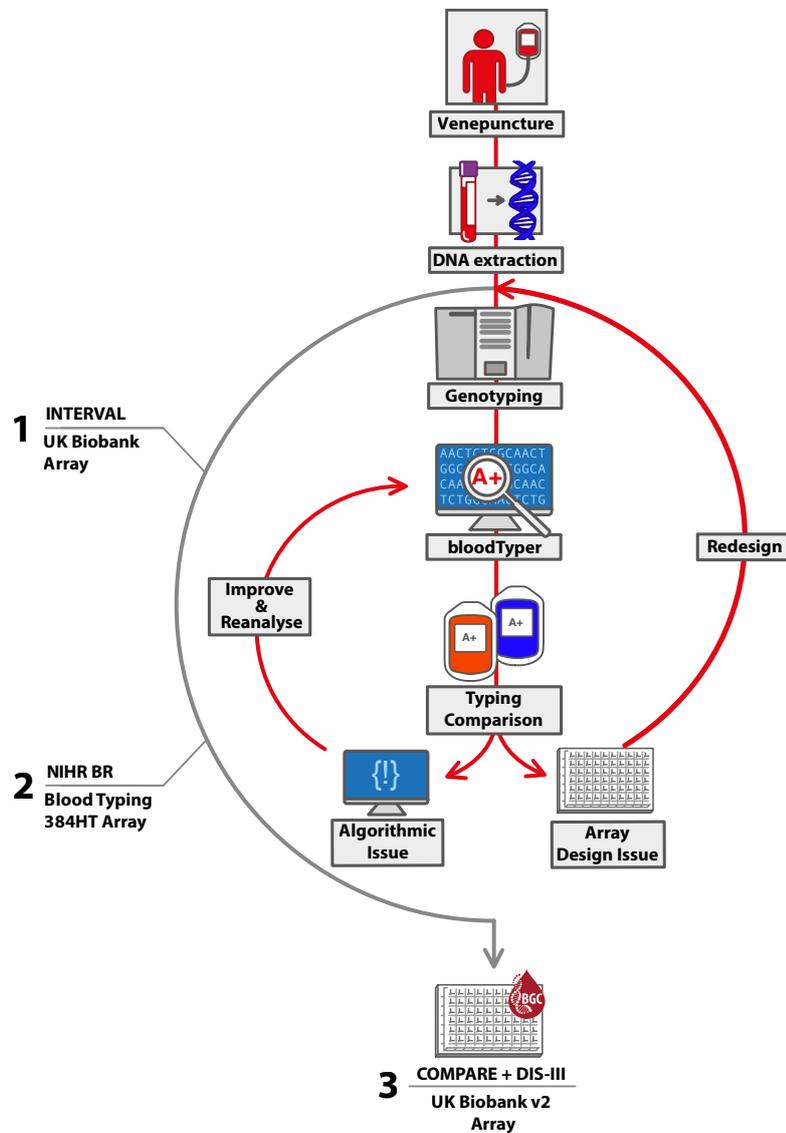


Fig. 4.1 Cartoon representing iterative cycles of array development and overall chapter workflow. (**Red Track**) Typical array development cycle. (**Grey track**) Overall project workflow, with significant array iterations indicated. **UK Biobank array**, Applied Biosystems UK Biobank Axiom Genotyping Array; **NIHR BR**, National Institute for Health Research BioResource; **UK Biobank v2 array**; Applied Biosystems UK Biobank version 2 Axiom Genotyping Array; **BGC**, Blood transfusion Genomics Consortium, **DIS-III**, Donor InSight III study.

Donor Information Study III (DIS-III) studies. Array determined blood antigen types were then compared to those available NHSBT and Sanquin donor records.

4.4 Stage 1 - Proof of principle

Using data available from donor records, 1,057 participants from the INTERVAL study cohort were selected on the completeness of their RBC antigen typing data. All selected individuals were of European ancestry by genotype. We then extracted genotype data for these individuals from the main INTERVAL UKBBv1 Axiom variant calls file using a combination of in-house developed scripts and bcftools v1.9. As both the serology and genotype data contained in the INTERVAL dataset has already undergone extensive quality control prior to release, no additional control steps were performed for this analysis.

We did however randomly select 100 individuals (including the 2 HapMap control samples: NA19315 and NA19318) from the remaining participants in the INTERVAL cohort and analysed their Axiom genotyping data through the pipeline to evaluate file formats and data integrity. The custom interpretive blood typing software bloodTyper utilises a curated antigen allele database, to infer antigen status from genomic data. Originally developed to infer RBC antigen typing from WGS data, adaptations were required to enable bloodTyper to process array data. Several issues with file formats were identified, and the decision was taken to reformat array files to comply with the Variant Call Format (VCF) 4.2 specification - a standard file format for storing genotype data. Furthermore, bloodTyper is designed to analyse genetic variation with respect to the human reference genome GRCh37/hg19, however, while array reported variants are positionally linked to this reference genome the reference (ref) and alternate (alt) alleles for each array variant are not. In-house software we developed to correct this by; 1) looking up each variant position in the reference genome, 2) comparing the ref and alt values to array ref and alt values, 3) correction of genotypes and ref/alt values in the event a swap has occurred. Following the implementation of these changes, a final round of pipeline compatibility testing was performed by repeat analysis of the 100 samples to verify that all previously identified compatibility issues had been fixed and no new ones were identified.

Inspection of these initial results revealed that some probes were clearly not working as intended. For example, all 100 individuals typed Kell_{null} with the specific allele detected being KEL*01N.01 (LRG_799:c.1678C>G). This highlighted the need to introduce a list of variants 'to be ignored' because of the lack of a good probeset on the array for reliable genotype detection. The Kell_{null} phenotype is extremely rare; the causal variant was unobserved in the exome aggregation consortium (ExAC) database which at the time of analysis contained the

sequence data of 125,748 individuals. The analysis was repeated after the $Kell_{null}$ variant was added to the ‘ignore list’.

Next, UKBBv1 array genotype data for each of the 1,057 selected INTERVAL samples were analysed using the bloodTyper algorithm and a comparison between array inferred and donor record antigen types were performed. Concordance between array inferred and donor record antigen types was 90.0% (14,276 concordant results in 15,862 comparisons) for 21 RBC antigens (see Fig. 4.2).

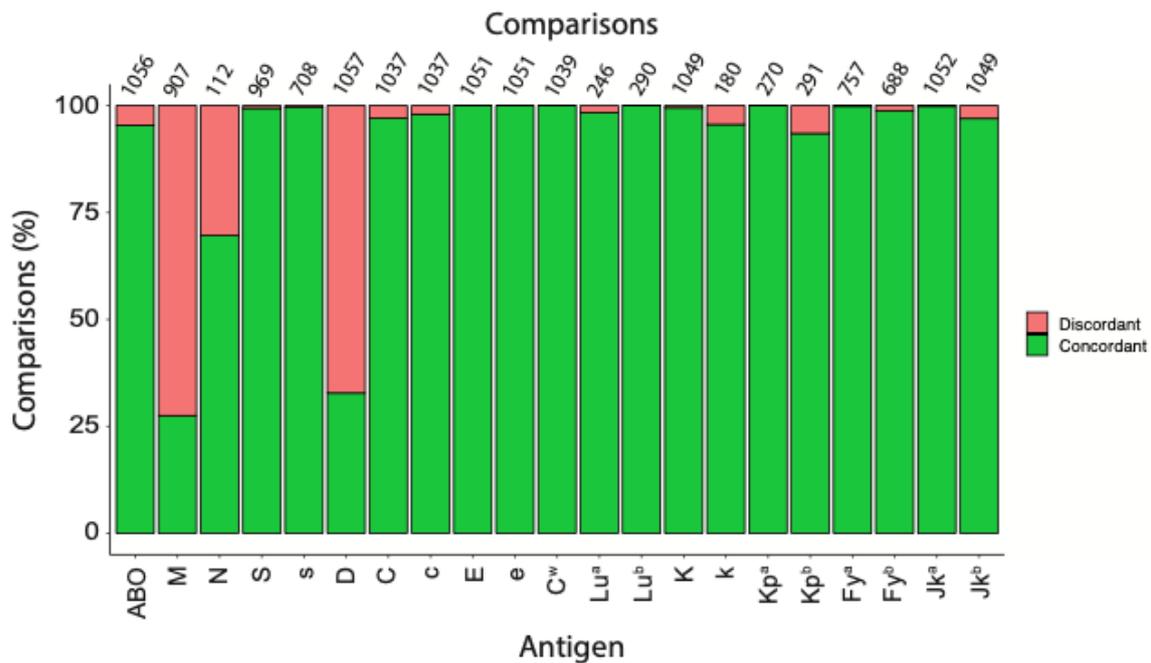


Fig. 4.2 Results from initial blood typing trial of the original UK BioBank (UKBBv1) array. Level of concordance per antigen is shown as a percentage of the total number of comparisons which is given at the top of each bar. Concordant and discordant results are in green and red, respectively.

For the ABO, M, N, D, C, c, Lu^a, k, Kp^b, Fy^b and Jk^b antigens concordance was unacceptably poor. Detailed investigation revealed that either a lack of appropriate variant typing or poorly performing probesets were the main reason for discordances. Some exemplars follow:

47 discordant results were observed between ABO typing results. Further analysis revealed that 39 (82%) of these cases were serologically group B individuals being genotyped as group AB or serologically group O individuals being genotyped as group A. These were explained by the lack of a probeset on the UKBBv1 array which interrogates the

LRG_792:c.802G>A variant responsible for the second most common allele (ABO*O.01.02) underlying the O phenotype.

Concordance for RhD antigen typing was only 32.8%. On inspection of typing results, it was revealed that all individuals had been reported as D-negative and that any concordance between typing results had happened by chance as approximately 30% of the NHSBT donors participating in INTERVAL are D-negative. The most common genetic cause of the D-phenotype in European ancestry individuals is the deletion of the entire *RHD* gene, and thus, the primary method of molecular D antigen typing is the detection of this deletion. While Axiom array technology can accurately detect deletions of this size, not enough *RHD* probesets were included on the UKBBv1 array and therefore accurate copy number typing of the *RHD* gene cannot be performed.

Similarly, RhC/c typing accuracy was only 28%. Further investigation revealed that the probeset interrogating the causal variant underpinning the C and c phenotype, LRG_797:c.307C>T, was not functioning as expected. In addition to discordance in antigen typing results, the MAF observed in this dataset for this variant did not correspond with MAFs observed in ExAC or the 100,000 genomes project. A working probeset was available for an alternative variant, LRG_797: c.48G>C, that is in high linkage disequilibrium (LD) with LRG_797:c.307C>T and can therefore be used as tag-SNP for C/c typing. A ‘force-call’ list was implemented to the analysis pipeline, allowing blood groups to be predicted from user defined variants such as the one in high LD. Although force-calling using this variant improved results, this approach is not ideal as LRG_797:c.48G>C does not directly encode for the amino acid polymorphism which underlies the C/c antigen difference. Furthermore, the LD is not equal to 1 and the G allele is not present in all individuals with a C phenotype.

The poor typing performance of other antigens was similarly explained by the lack of properly performing probesets or probes for a critical variant causal of the antigenic difference was entirely lacking. All together we can conclude that because of the high error rate between genotype inferred and clinically determined antigen types, UKBBv1 array data cannot be used for inference of RBC antigens in the clinical laboratory.

However, for some antigens such as Kp^a and Lu^b, 100% concordance between DNA- and antibody-based antigen typing results were observed. This excellent concordance indicated that, with further development, the Axiom technology could become suitable for reliable comprehensive donor typing. Most importantly, this analysis revealed that with inspection of results the root cause of each discordance could be understood and a strategy formed for resolution - namely introduction of blood typing content to the array.

4.5 Stage 2 - Designing an array for donor genotyping

Guided by the results of our proof of principle experiment, we set out to improve the design of the UKBBv1 array with respect to donor typing by including probesets for all known antigen encoding variants. To this end, we identified 1,602 variants for inclusion from three sources: 1) all antigen encoding variants known to ISBT, 2) variants suggested by the BGC panel of experts which were known to underpin blood group antigen expression but not yet published, and 3) variants designed to type unique regions of the *RHD* and *RHCE* gene to be used for gene copy number identification.

Probes for identified variants were then integrated into the 384 Blood Typing Array design. We additionally included: 4,347 variants located within RBC and HPA antigen encoding genes, 6,572 variants located in *HLA* encoding genes and 31,946 variants for sex and ancestry inference from the UKBBv1 array.

After the fabrication of the new array, Thermo Fisher genotyped 314 DNA samples from the HapMap study and ran automated probeset quality analysis as part of their routine internal QC process for new array designs. Additionally, genotype call-plots were drawn from this data for the 1,602 antigen typing variants and were independently inspected by three BGC members and ranked according to performance (See Fig. 2.3 in Methods chapter). 38 variants important to antigen typing were identified in this analysis which did not have working probesets and subsequently excluded from further analysis (see Table. 4.1). These 38 variants were also given a priority ranking according to the clinical importance of phenotype, and those identified as priority 1 (highest) variants were flagged for probeset re-design in future validation studies.

Table 4.1 Antigen typing variants for which no working probeset could be identified

Variant Priority	Gene	rsID	Phenotype
1	<i>RHD</i>	rs371803235	Partial D (Type 11) or Del
1	<i>RHD</i>	rs142037235	D _{weak} (Type 18)
1	<i>RHCE</i>	rs144348222	Rh26+ LOCR- / Rh26- LOCR+
1	<i>RHAG</i>	rs16879498	Rh null
1	<i>RHD</i>	rs17418085	Partial D typing, multiple
1	<i>RHD</i>	rs1053355	Partial D typing, multiple
1	<i>RHD</i>	rs1053356	Partial D typing, multiple
1	<i>RHD</i>	rs1053359	Partial D typing
1	<i>RHD</i>	(null)	Partial D (DWI)
1	<i>RHD</i>	rs1132760	Partial D (DVII type 2) Tar+
1	<i>A4GALT</i>	(null)	p (P1- pk - NOR-)
1	<i>A4GALT</i>	(null)	p (P1- pk- NOR-)
1	<i>ABCB6</i>	rs200125320	Lan-
1	<i>CR1</i>	rs41274768	Kn ^{a+} / Kn ^{b+}
1	<i>SLC14A1</i>	rs759505281	JK _{null}
1	<i>SLC14A1</i>	rs113578396	Jk ^a _{weak}
1	<i>ITGB3</i>	rs13306487	HPA-6a+ / HPA-6b+
1	<i>ITGA2B</i>	rs5911	HPA-3a+ / HPA-3b+
1	<i>ART4</i>	rs185001341	D _{onull}
1	<i>RHD</i>	rs150606530	DAR3.1/4/5 (DEL)
1	<i>RHD</i>	(null)	D-
1	<i>RHD</i>	(null)	D-
1	<i>RHCE</i>	(null)	C-/c+
1	<i>ABO</i>	rs8176743	B
2	<i>RHCE</i>	rs144163296	STEM-
2	<i>BCAM</i>	rs28399659	LU:-13
3	<i>RHD</i>	rs590813	D _{weak} (Type 66)
3	<i>RHD</i>	(null)	D _{weak} (Type 58)
3	<i>RHD</i>	rs770829982	D _{weak} (Type 24)
3	<i>RHD</i>	rs17421158	D _{weak} (Type 14)
3	<i>ABO</i>	(null)	O
3	<i>ITGAM</i>	rs1143679	HNA-1b Neutrophil adherence receptor
3	<i>XK</i>	(null)	Kx-
3	<i>GP1BB</i>	rs375285857	HPA-12bw- / HPA-12bw+
3	<i>C4A</i>	rs28357076	CH+RG- / CH-RG+
3	<i>FUT2</i>	rs1047781	FUT2:06
3	<i>B3GALNT1</i>	rs2231257	GLOB+
3	<i>SEMA7A</i>	rs56367230	JMHK+ / JMhk-

Then 507 NIHR BioResource samples, selected by completeness of RBC antigen typing data on NHSBT's electronic donor records, were genotyped using the 384 Blood Typing Array. No samples were excluded during quality control of the genotyping data. Genotypes were analysed using the bloodTyper pipeline, then array and donor record antigen types compared. A significant improvement in the accuracy of genotype inferred antigen types was observed (see Fig. 4.3).

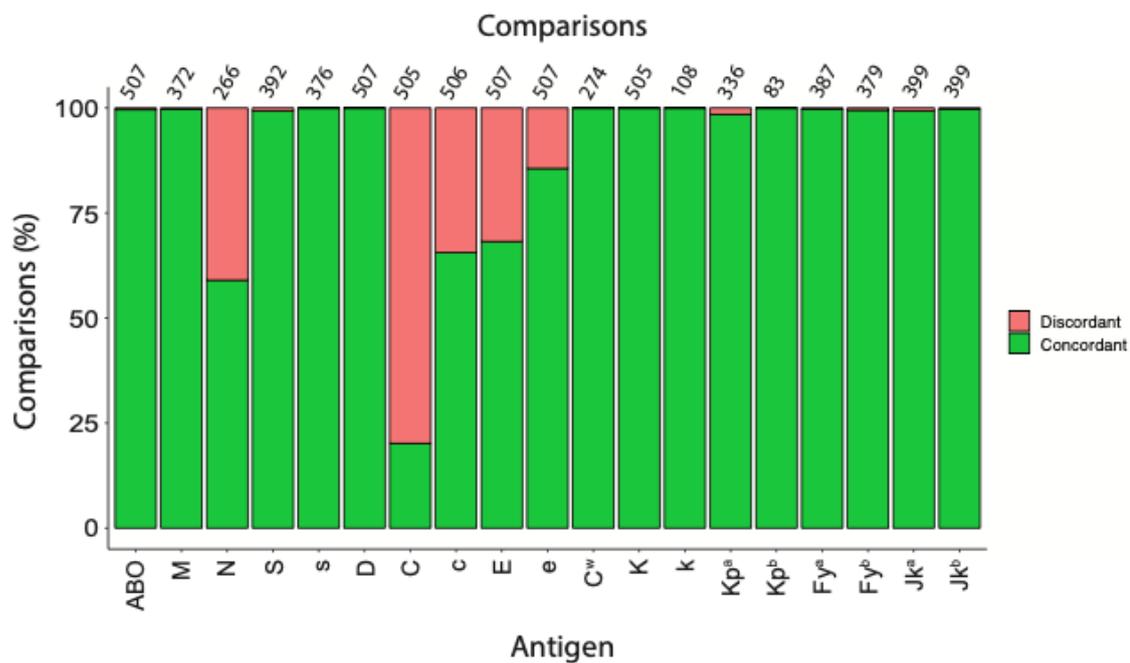


Fig. 4.3 Results from antigen typing trial of the novel 384 Blood Typing array. Level of concordance per antigen is shown as a percentage of the total number of comparisons which is given at the top of each bar. Concordant and discordant results are in green and red, respectively.

ABO genotyping accuracy increased from 95.4% in the previous experiment to 99.5%. This was due to the inclusion of important ABO typing variants such as the already mentioned LRG_792:c.802G>A (ABO*O.01.02) variant which encodes a group O phenotype. Two discordant results were observed; one serology group O to genotype group B and the other serology group O to genotype group A. Inspection of genotype and call-plots for these cases did not reveal any reason for discordance. From this, we concluded that it is likely that genetic variants not previously linked to the group O phenotype are responsible for the discordances. Sequencing the *ABO* locus of these individuals would be the logical next step of investigation for these two cases, however, no further DNA or blood samples were available.

D genotyping accuracy increased dramatically from 32.8% to 99.8% due to accurate detection of *RHD* gene copy number by the new array. Just a single discrepancy was observed, serology D- to genotype D+. Genotyping results indicated that this individual had 1 copy of *RHD* and this call was produced by 114 independent probesets. Further analysis of genotyping results revealed the presence of two variants in the *RHD* gene of this individual, LRG_796:c.697G>C and LRG_796:c.712G>A. In combination, these variants encode the RHD*05.09 allele which is associated with variant expression of the D antigen (DV type 9). This could suggest that result in the electronic donor record is incorrect. A monoclonal anti-D is used by NHSBT for routine D typing of donors and experts indicated mistyping would be unlikely but not impossible for individuals carrying a variant *RHD* gene. A similar D antigen mistype was recorded in the last 10 years by the NHSBT International Blood Group Reference Laboratory (IBGRL), where tests using the standard monoclonal anti-D failed to detect a similar phenotype (DV type 4). The RHD*05.04 (DV type 4) allele shares the variant LRG_796:c.697G>C with the RHD*05.09 (DV type 9) allele suggesting that the epitope for the monoclonal antibody may not be present on the RBCs of individuals with this genotype.

Genotyping accuracy for some antigens, however, decreased when using the new assay. RhC/c and RhE/e antigen genotyping accuracy decreased significantly to 42.93 and 76.93% concordance for C/c and E/e, respectively. The original UKBBv1 array did not directly interrogate the variants that encode C/c and E/e antigen expression but relied on “tagging” variants that are in high LD with the phenotype encoding variant. While these ‘tag-SNPs’ performed well, the decision to stop using them to infer antigen types was taken due to the fact that the use of tag-SNPs is not an acceptable solution as LD was not 100%. In translation, this means there would have been a hard-coded higher error rate in the bloodTyper analysis algorithm for these antigens. Following this design choice, probes for typing the variants that directly encode expression of C/c and E/e antigens, NM_020485.5:c.307C>T and NM_020485.5:c.676C>G, were included on the 384 Blood Typing array. However, genotypes for these variants were initially difficult to interpret.

Genotyping of Rh antigens is challenging due to high sequence homology between the *RHD* and *RHCE* genes. This often causes miss-binding of array probes or erroneous alignment of NGS sequencing reads. We developed novel algorithms to overcome these difficulties (see Fig. 4.4).

Firstly, we used data from 114 copy-number probesets tiled across the *RHD* gene at loci with the lowest sequence homology to the *RHCE* gene to ascertain *RHD* copy number. These probes allowed us to accurately identify *RHD* copy number using the standard CNVmix algorithm (Fig. 4.4a,b).

Secondly, genotype calling algorithms expect samples to gather in three clusters in correlation to genotype. However, due to cross binding of *RHCE* probes to the *RHD* gene, multiple overlapping clusters were observed for *RHCE* variants. Grouping samples by array determined *RHD* copy number allowed us to identify that the degree of interference caused by cross-binding is directly linked to *RHD* copy number. Three groups of clusters were observed (Fig. 4.4c).

Thirdly, samples were split according to *RHD* copy number and genotypes for *RHD* and *RHCE* variants were re-called for each copy-number sample group (Fig. 4.4d).

Application of this algorithm significantly increased concordance in the test set to 98.71% and 99.61%, for the C/c and E/e Rh group antigens, respectively (Fig. 4.4e).

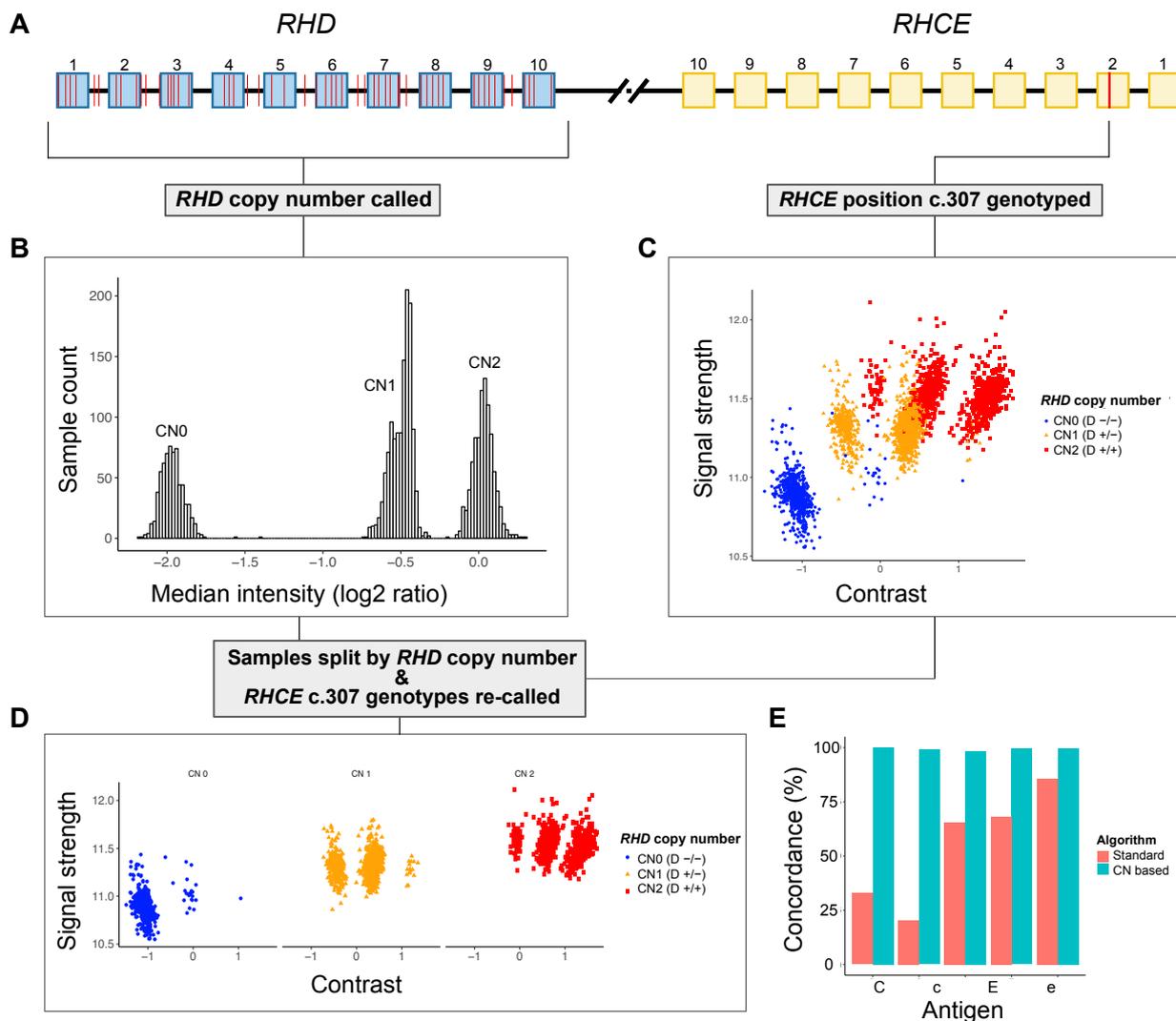


Fig. 4.4 (a) Diagrammatic representations of the *RHD* and *RHCE* genes, the binding position of array probes are indicated in red. The probes spanning *RHD* are used for *RHD* copy number assessment and for *RHCE*, as an example, the probe for typing variant LRG_797:c.307C>T, which is used for C/c typing. (b) All samples in the DIS-III cohort binned by *RHD* median copy number probeset intensity ratio (log₂), copy number clusters are annotated. (c) Genotype call plot for the C/c antigen variant, LRG_797:c.307C>T, in the *RHCE* gene. Samples are coloured by array determined *RHD* gene copy number. Three unique sets of genotype clusters can be observed, with a 'right shift' in contrast linked to *RHD* copy number. (d) Genotype call plots for the same samples and variant, LRG_797:c.307C>T, post copy number stratified genotyping. (e) Improvement in Rh(C/c) and Rh(E/e) antigen typing concordance resulting from application of the copy number genotyping algorithm.

The N antigen also proved extremely challenging to type. All 109 discordances concerning this antigen were serology N- to genotype N+. Inspection of bloodTyper calls excluded an analysis pipeline issue, therefore discrepancy is caused by either the genotype calling algorithm or due to a malfunctioning probeset. Molecular typing of N can be difficult because of sequence homology between the *GYP A* and *GYP B* genes. There is a region of 100% sequence identity between the two genes where the variant controlling N expression in *GYP A* is located. This biological artefact is responsible for the antibody determined phenotype 'pseudo N in people who are N-. Following the addition of copy number probes located in *GYP A* and *GYP B*, the copy number aware algorithm used to improve *RHD* and *RHCE* variant calling could be applied to N antigen typing. The probes used for N antigen typing were also flagged for review and redesign in future array versions.

4.6 Stage 3: Large scale blinded trial of the UKBBv2 array

Guided by the results from the previous stages the donor typing content from the 384 Blood Typing array was integrated into the original UKBBv1 array design.

We also added 9,180 variants in 49 genes relevant to antigen expression which had minor allele frequencies (MAFs) >0.02% in large-scale sequencing datasets (see Table. 4.2). These were included to 'future proof' the array by capturing non-synonymous variants with possible clinical consequences for antigen expression discovered in population scale sequencing studies. This approach also allows us to link these variants to antigen expression in future studies. The resultant Applied Biosystems UK Biobank – version 2 Axiom Array (UKBBv2 array) has significantly increased content in the RBC and HPA antigen-encoding genes (see Fig. 4.5).

In addition to physical array changes, algorithms were introduced to bloodTyper for handling specific technical events. For example, in the event a probeset fails and therefore no genotype is called for an important variant in the *ABO* locus underlying a group O phenotype, such as LRG_792:c.261delG; no ABO antigen inference is made by bloodTyper for safety reasons.

Table 4.2 Genes relevant to antigen expression selected for coding variant enrichment on UKBBv2 array

HGNC gene ID	Inclusion Reason
<i>A4GALT</i>	Blood Group Antigen
<i>ABCG2</i>	Blood Group Antigen
<i>ABO</i>	Blood Group Antigen
<i>ACHE</i>	Blood Group Antigen
<i>ACKR1</i>	Blood Group Antigen
<i>AQP1</i>	Blood Group Antigen
<i>AQP3</i>	Blood Group Antigen
<i>ART4</i>	Blood Group Antigen
<i>B3GALNT1</i>	Blood Group Antigen
<i>BCAM</i>	Blood Group Antigen
<i>BSG</i>	Blood Group Antigen
<i>C4A</i>	Blood Group Antigen
<i>C4B</i>	Blood Group Antigen
<i>CD109</i>	Platelet Antigen
<i>CD151</i>	Blood Group Antigen
<i>CD31</i>	Blood Group Antigen
<i>CD44</i>	Blood Group Antigen
<i>CD55</i>	Blood Group Antigen
<i>CD59</i>	Blood Group Antigen
<i>CD99</i>	Blood Group Antigen
<i>CR1</i>	Blood Group Antigen
<i>ERMAP</i>	Blood Group Antigen
<i>FUT1</i>	Blood Group Antigen
<i>FUT2</i>	Blood Group Antigen
<i>FUT3</i>	Blood Group Antigen
<i>GBGT1</i>	Blood Group Antigen
<i>GCNT2</i>	Blood Group Antigen
<i>GP1BA</i>	Platelet Antigen
<i>GP1BB</i>	Platelet Antigen
<i>GYP A</i>	Blood Group Antigen
<i>GYP B</i>	Blood Group Antigen
<i>GYP C</i>	Blood Group Antigen
<i>ICAM4</i>	Blood Group Antigen
<i>ITGA2</i>	Platelet Antigen
<i>ITGA2B</i>	Platelet Antigen
<i>ITGB3</i>	Platelet Antigen
<i>KEL</i>	Blood Group Antigen
<i>KLF1</i>	Blood Group Antigen Related
<i>RHAG</i>	Blood Group Antigen
<i>RHCE</i>	Blood Group Antigen
<i>RHD</i>	Blood Group Antigen
<i>SEMA7A</i>	Blood Group Antigen
<i>SLC14A1</i>	Blood Group Antigen
<i>SLC29A1</i>	Blood Group Antigen
<i>SLC4A1</i>	Blood Group Antigen
<i>SMIM1</i>	Blood Group Antigen
<i>XG</i>	Blood Group Antigen
<i>XK</i>	Blood Group Antigen

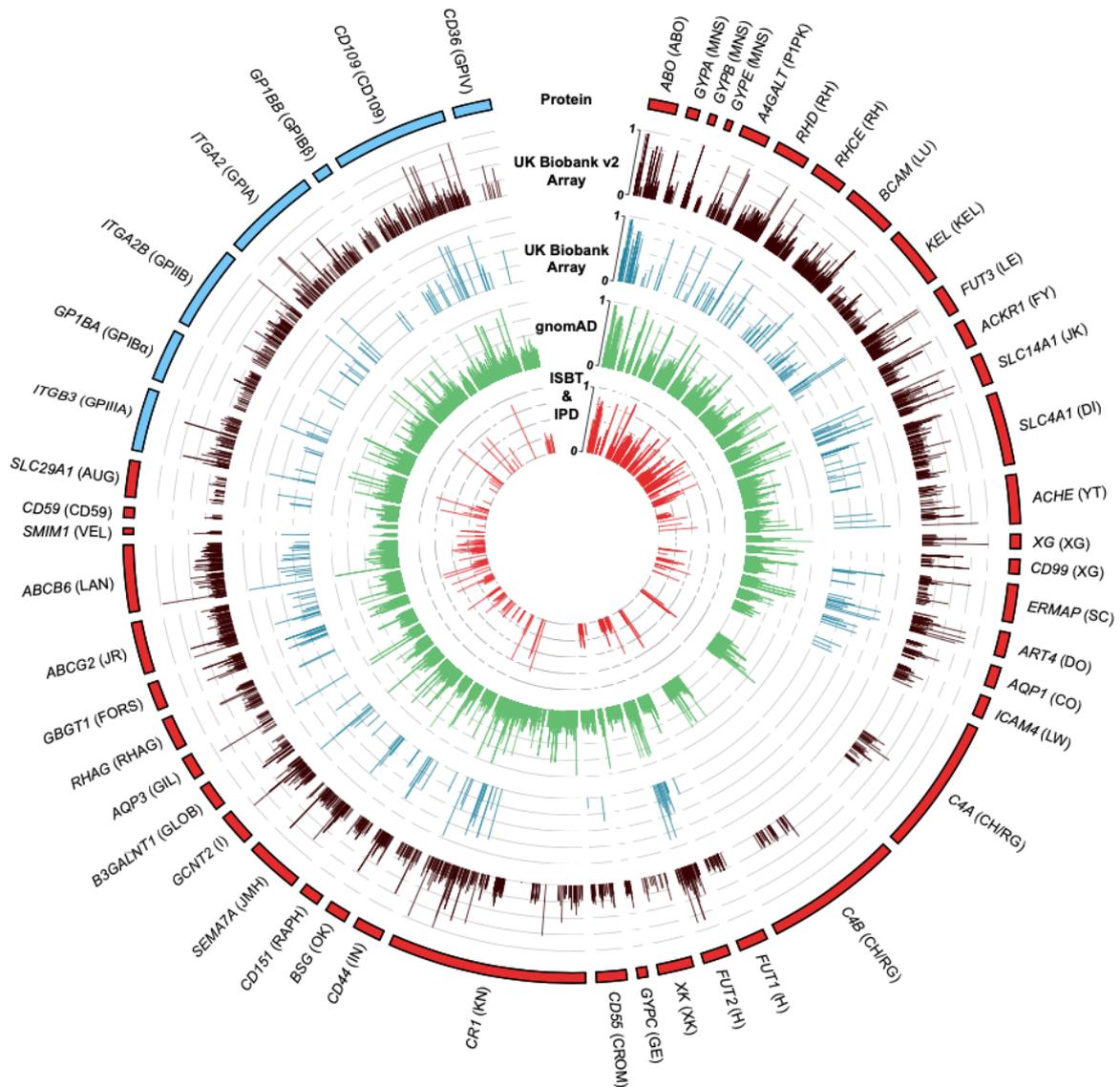


Fig. 4.5 Variation in antigen encoding genes compared to array content. Track (1-5) from outside to inside, (1) representations of Red Blood Cell (RBC, red) and Human Platelet Antigen (HPA, blue) antigen encoding genes, box length equates to the number of amino acids in the protein. RBC and HPA system names are annotated next to gene labels; (2-3) coding variants included on the UK Biobank (UKBB) v2 (crimson) and UKBB (teal) array, respectively; (4) all coding variants with consequences in the gnomAD database (green); (5) antigen encoding variants in the ISBT and IPD-HPA databases (red). Bar length for tracks 2-5 represents gnomAD alternate allele frequency.

4.6.1 Selection of trial set samples and initial genotyping quality control

Following UKBBv2 array fabrication, we selected 7,477 blood donors from England (n=4,795) and the Netherlands (n=2,682) who had consented to participate in the COMPARE and DIS-III studies. DNA samples from these individuals were used as a 'trial set' to validate the UKBBv2 array.

In addition to blood donor samples, each array plate of 94 donor samples had two HapMap DNA samples as controls included in it. These control samples have been extensively genotyped and analysed by WGS. Samples NA19315 and NA19318 were included on 52 plates with the samples from the English donors, and samples HG00097 and HG00264 were included on 32 plates with the samples from the Dutch donors. This repeat genotyping allowed us to assess array-wide repeat genotype concordance which was >99%, and concordance with WGS genotype data which was >99% across all 168 repeated samples tests (see Fig. 4.6).

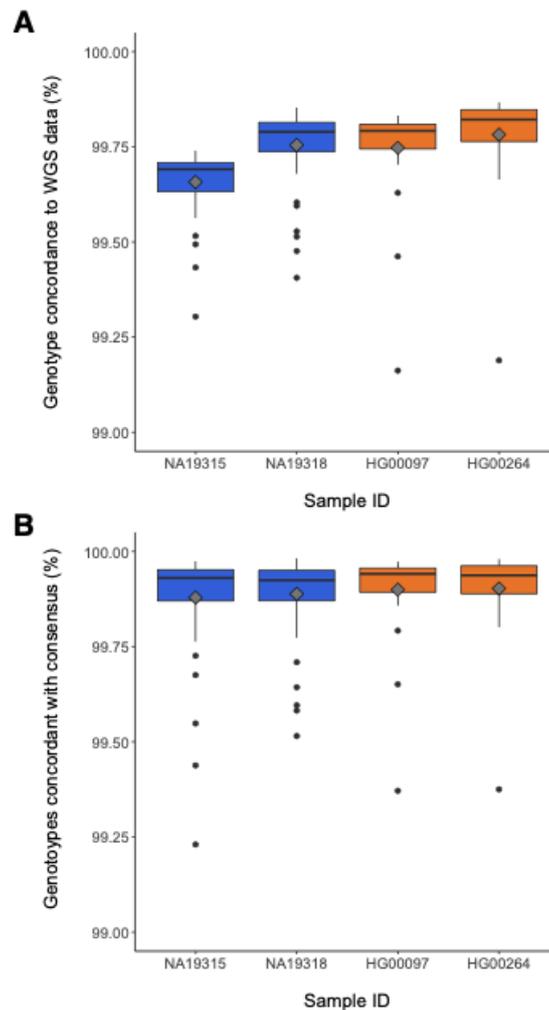


Fig. 4.6 (a) Array genotype concordance to next generation sequencing data across multiple repeats of each sample for all typed variants. A drop in UKBBv2 array and WGS genotype concordance can be observed for NA19315. This sample is from an individual of African ancestry and it is therefore likely there is bias in the read alignment and genotype calling for the WGS data. (b) Genotype repeatability across all repeats of each sample for all typed variants. Samples NA19315 and NA19318 were used for the English donor samples (blue) and repeat-tested 52 times each, and HG00097 and HG00264 used for the Dutch donor samples (orange) and repeat-tested 32 times each, on separate genotyping runs. Grey diamonds display means, middle box lines, box bounds and whiskers represent median, upper and lower quartiles, and values spread beyond middle 50% of overall distribution are outliers and represented by circles.

As genotype concordance data from repeat testing of HapMap samples was excellent, we proceeded with QC of the 7,477 trial set blood donor samples. 15 samples with imputed versus declared sex mismatches and more than five antigen typing discordances were identified. Investigation revealed that these discrepancies were either caused by erroneous handling of samples (n=4) or by incorrect data entry into the study database (n=11). The data from the former category were excluded from further analysis (see Table. 4.3). Samples in the latter category were retained following correction of sample linkage information. This left 7,473 samples for analysis in the final trial.

On further investigation, it was revealed that self-declared sex vs. genotype inferred sex discordance arose for one sample as the contributing participant no longer identified with their birth sex. In the case of this validation study, the sample was removed from further analysis in accordance with study procedures as sex discordance analysis together with discordance analysis for the ABO groups provides an added layer of quality assurance to identify possible procedural errors, including sample handling errors. In practice, the BGC recommends that sex mismatches should be treated as incidental ‘secondary findings’ and not be reported back to the individual, except if explicit consent has been obtained.

Table 4.3 Trial set samples removed from further analysis

Sample	Cohort	Submitted sex	Array inferred sex	Discordant antigens count	Antigens
1021232667	COMPARE	Male	Female	7	ABO;Fy(a);Fy(b);Jk(a);K;S;E
1032864627	COMPARE	Female	Male	6	ABO;D;Fy(b);M;C;c
1032864628	COMPARE	Male	Female	6	ABO;D;Fy(b);M;C;c
1032864332	COMPARE	Female	Male	4	Fy(b);M;S;c

4.6.2 Allele frequency validation of the UKBBv2 array

The 7,473 trial set blood donors were each genotyped for 789,550 DNA variants using the UKBBv2 array. 10,923 of these variants are used for inference of RBC, HPA and HLA antigens. We compared the GRCh37 alternative allele frequencies (AAF) of the genotyped variants for the 7,416 unrelated European ancestry donors in the trial set with the corresponding AAFs measured by WGS in 8,510 unrelated European ancestry participants in the NIHR BioResource pilot phase of the 100,000 Genomes Project. No significant difference in AAF was detected for 99.02% of the 716,102 variants with $MAF > 0.01\%$ in both WGS and array data at a Bonferroni adjusted critical threshold ($\alpha = 0.05/716,102$, Fig. 4.7a,b). Frequency comparison data for HLA, and RBC/HPA antigen typing variants is shown in Fig. 4.7c-d. For 84 of RBC antigen typing variants measured AAFs differed significantly. These variants were also given a priority ranking according to importance of phenotype, and those identified as priority 1 (highest) variants were flagged for re-design in future validation studies.

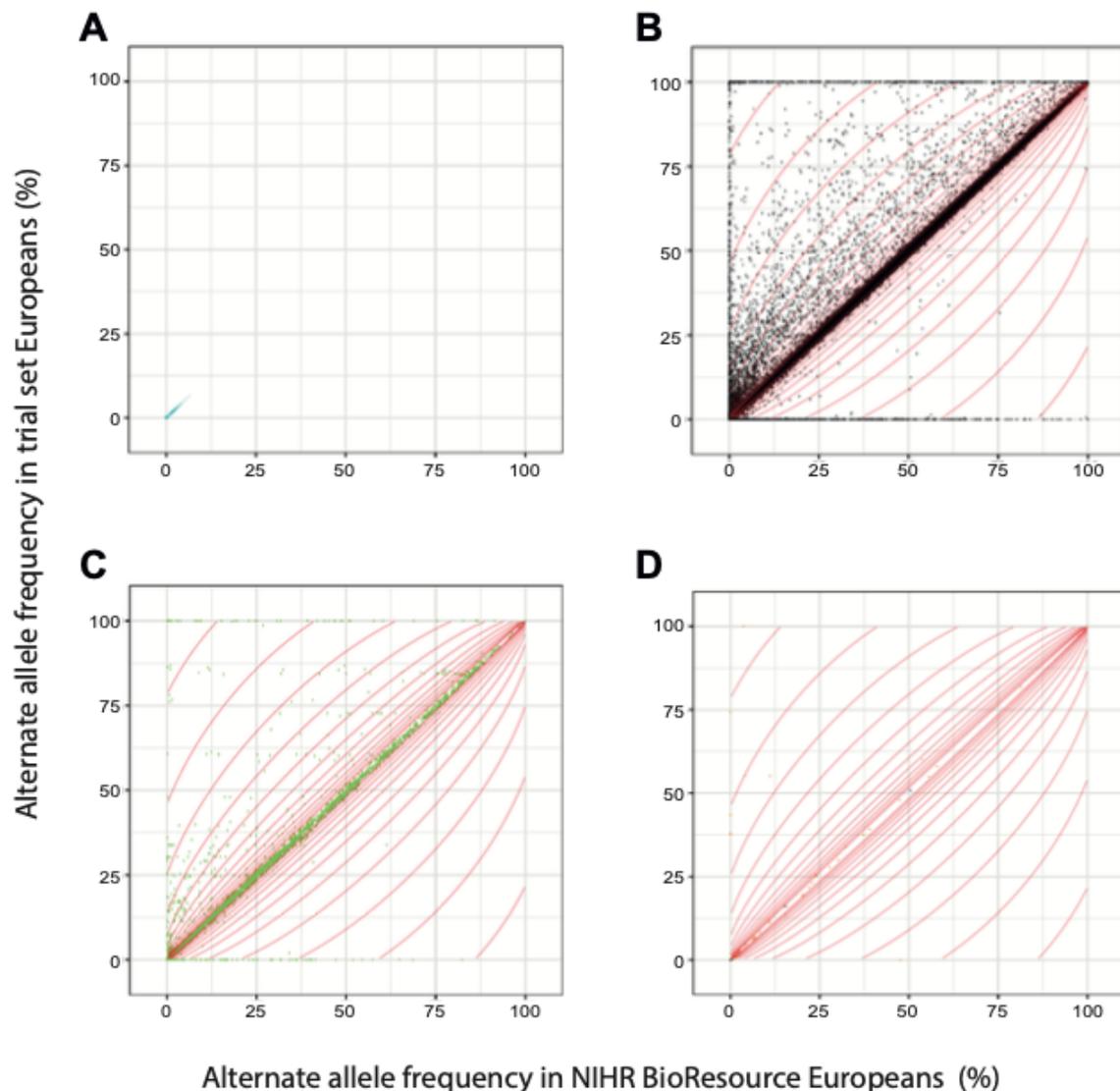


Fig. 4.7 (a) A kernel density plot showing excellent average concordance between alternative allele frequency in 7,416 UKBBv2 array typed unrelated European ancestry English donors and 8,510 WGS typed unrelated European ancestry NIHR BioResource volunteers, across the 709,713 variants with alternative allele frequency $>0.1\%$ in both datasets. (b) Bi-variate scatter of the data used to generate the kernel density plot, emphasising outliers. Each red arc corresponds to the critical threshold for a Pearson test of a given size comparing allele frequencies in the two datasets, under the assumptions that genotypes were observed for all participants and that Hardy-Weinberg equilibrium holds in both datasets. The arcs are uniformly separated on a log-scale. (c) Variants in the HLA locus. (d) Variants encoding RBC or HPA antigens.

4.6.3 Overall antigen typing performance

Overall concordance between clinical and UKBBv2 antigen typing was 99.82% in 103,326 comparisons across 28 RBC antigens, 10 HPA antigens and 6 HLA loci for which clinical typing data were available (see Fig. 4.8). All but 57 of the 7,473 donors enrolled in this study were of European ancestry. Of the non-European ancestry individuals; 27, 21 and 9 were of South Asian (SAS), East Asian (EAS4) and African ancestries, respectively. Antigen typing concordance for these 57 individuals was 100% in 835 comparisons across 28 RBC and 10 HPA antigens compared.

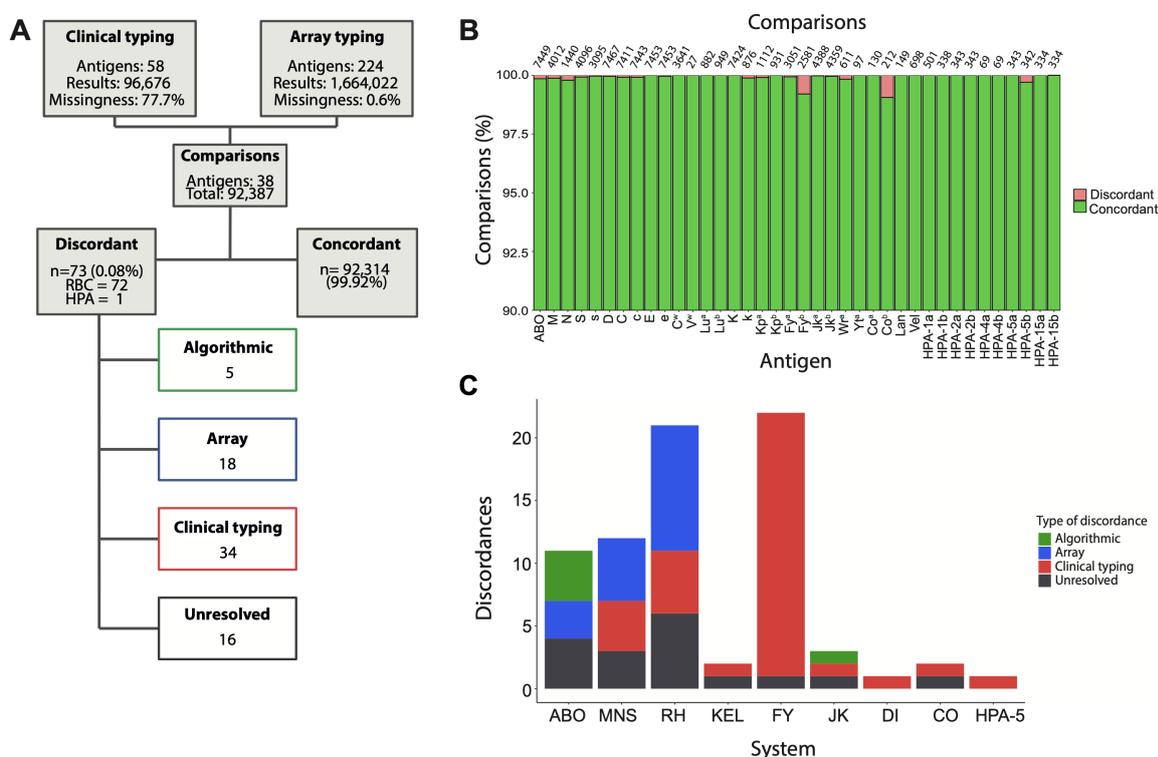


Fig. 4.8 (a) Overview of concordance analysis and high-level categorisation of results. (b) Concordance per antigen is shown as a percentage of the total number of comparisons (given at the top of each bar) with concordant and discordant results in green and red, respectively. (c) Categorisation by error type of discordances for the different RBC antigen systems and for the HPA-5 system.

4.6.4 RBC antigen typing performance

Concordance between RBC antigen typing results was 99.91% in 89,371 comparisons across 28 antigens (see Fig. 4.8a-b). We categorised the 72 (0.08%) remaining results according to the reason for discordance (see Fig. 4.8a-c). 33 (45.8%) discordances were explained by erroneous clinical typing results. In 19 of these cases, variants encoding variant antigen expression were detected by the array, examples include Del (RHD*11, NM_016124:c.885T), K_{mod} (KEL*02M.01, NM_000420.2:c.1088G>A) and Fy^x (FY*02W.01, NM_001122951.2:c.265C>T and NM_001122951.2:c.298G>A). For carriers of these alleles, the chances of false-negative antibody-based typing results are greatly increased. A unit of blood being erroneously typed as negative for a given antigen may boost antibody levels in a previously sensitised patient. The remaining 14 discordances in this category involved antigens where current typing reagents have been known to give incorrect results (n=4) or where sequencing analysis confirmed array genotyping results (n=10).

Sixteen of the discordances remain unresolved. However, in six of these cases previously unobserved DNA variants, which were likely to underpin an antigen-negative phenotype were discovered using the targeted sequencing platform designed for discordance resolution (see chapter 5). The absence of these newly identified variants from the ISBT reference tables prevents their use in antigen phenotype inference. To determine the effect of these six unique variants functional antigen expression studies are required, therefore we regard these cases as unresolved. For the remaining 10, we were unable to resolve the cause of the discordance due to a lack of a DNA sample for further analysis. The cause of these discordances and the genotyping method developed to investigate them will be detailed in chapter 5.

4.6.5 HPA antigen typing

A proportion of patients who are refractory to platelet transfusions due to HLA class I antibodies also form HPA antibodies. For these patients, platelet concentrates lacking the relevant HLA class I and HPA antigens are required. We report 99.96% concordance in 3,017 comparisons across 10 HPA antigens, with only a single discordant result observed for the HPA-5a antigen (Fig. 4.8a-c).

4.6.6 HLA antigen typing

Most blood supply organisations maintain panels of HLA class I typed apheresis platelet donors to support patients who are refractory to ABO-compatible platelet concentrates due to HLA class I antibodies. Clinical HLA typing data were available for 1,221 of the 7,473

trial set donors. Concordance between typing results across six HLA class I and II loci, at two-field resolution, was 99.03% in 9,289 comparisons (see Fig. 4.9).

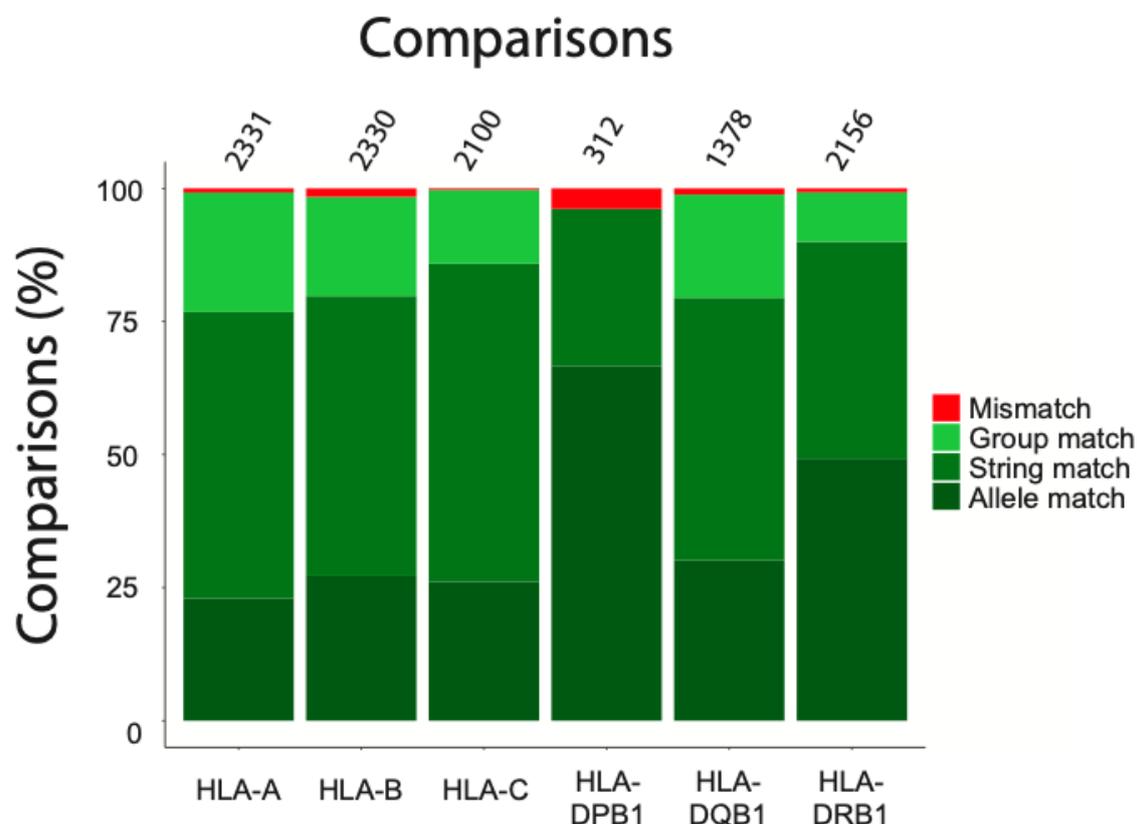


Fig. 4.9 HLA typing result comparison. Concordance per antigen is shown as a percentage of the total number of comparisons (given at the top of each bar) with concordant and discordant results in green and red, respectively. Concordance was assessed at group, string and allele levels.

4.7 Impact of an extensively genotyped donor panel

For 7,473 trial set donors 37 RBC and 11 HPA antigens had at least one clinical typing result available. These 48 antigens form a dataset that is representative of extended donor typing programmes conducted on English and Dutch donors (Fig. 4.10a-b). In contrast to clinical typing data, genotype inferred antigen typing is almost complete with only 458 (0.13%) of 354,624 possible results missing (Fig. 4.10c-d). Genotyping yielded 3.8 times more types (47.9 vs 12.6 results per donor) compared to current practice. Genotyping also allowed us to type the donors for 185 additional antigens for which no clinical typing results were available. In total 1.2 million antigen types, on average 214.6/donor, were produced by genotyping.

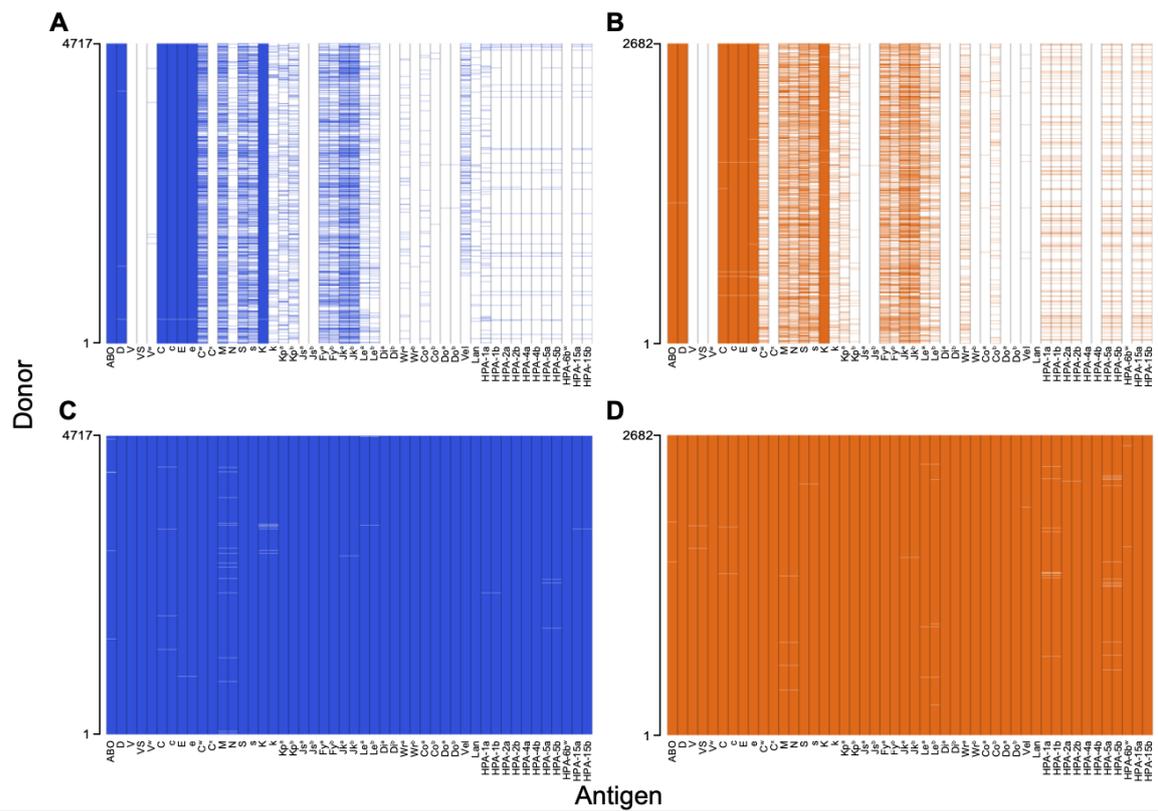


Fig. 4.10 Comparison of clinical and UKBBv2 array antigen typing availability in the trial set of samples. Presence of colour represents presence of a typing result (positive or negative) and absence of colour indicates lack of a typing result. (A-B) and (C-D) represent clinical and UKBBv2 array typing data for English (blue) and Dutch donors (orange), respectively.

To investigate the potential benefits of densely genotyping donors we used data from patient referrals to NHSBT over a five-year period. Filtering this dataset for 'complex cases', defined as patients with alloantibodies against at least three different RBC antigens, produced a set of 3,146 referrals with 1,253 unique alloantibody combinations (Fig. 4.11a). As expected, we observed an inverse relationship between the probability of finding a compatible donor and the number of alloantibodies identified in a patient (Fig. 4.11b). Using the trial set of 7,473 donors as a 'virtual stock', we report a 2.6-fold greater probability of identifying ABO and RhD compatible donors who are negative for the relevant combination of antigens needed to support patients with one of the 1,253 alloantibody combinations, when using the genotype data. This translates to an additional 176 patients for whom a matched donor could be identified. When finding donors for the most common alloantibody profiles, genotyping data was equivalent to clinical typing data, but for the rare combinations genotyping increased the average number of identified donors by 72 (Fig. 4.11c). We were even able to identify a compatible donor for a patient with alloantibodies against nine different RBC antigens.

4.7.1 Identification of individuals with rare phenotypes

The ISBT defines a rare donor as one with a phenotype less frequent than 1 in 1000 in a given population. This includes individuals negative for high frequency antigens (HFAs). On average, NHSBT and Sanquin each type approximately 52,000 donors per year for HFAs to identify those with HFA negative phenotypes. Fy^{a-b-} , $k-$, Lu^{b-} , and $Vel-$ and Yt^a- are examples of rare phenotypes that are in high demand. The genetic variants determining these particular examples are typed by both the UKBBv1 and UKBBv2 arrays. Querying the genotype data for 7,473 trial set donors identified 38 donors who lacked at least one of these five HFAs. Expanding the query to the UKBB array typing data from 533,308 UKBB and INTERVAL participants identified a further 8,951 individuals with potential HFA negative blood types. This demonstrates that blood supply organisations could improve the supply of rare blood types, for which level of demand is often a challenge, by typing donors with a comprehensive DNA based test or recruiting from national scale genotyping studies where data is readily available.

4.7.2 Immediate clinical benefit of extended typing data

Illustrating the clinical need for denser typing of donors, we present a female patient with myelodysplastic syndrome and anti-E, anti- Co^b and anti- Wr^a . Because of her bone marrow failure she required regular transfusions, however, no suitable blood donors could be identified in the 341,509 Dutch donors registered with Sanquin. Querying genotyping results from the

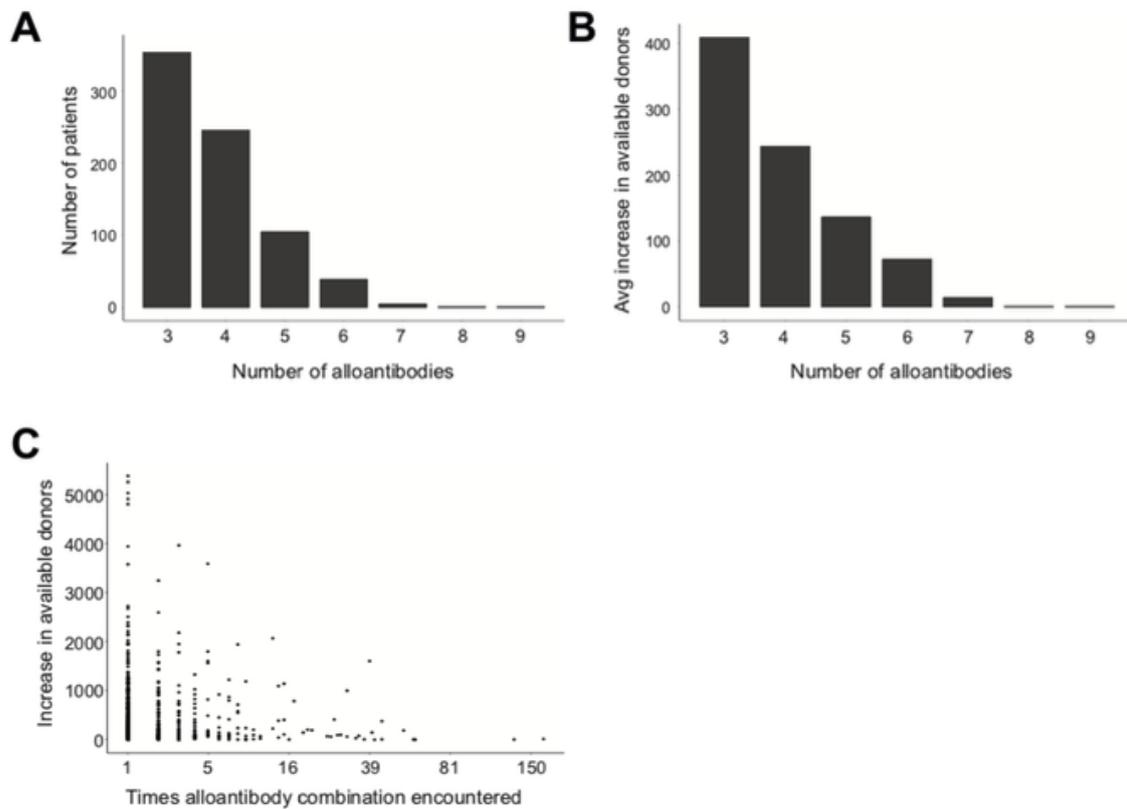


Fig. 4.11 (a) Number of alloantibodies identified in patient samples, with number of patients and antibodies on the vertical and horizontal axis, respectively. (b) Average increase in the number of compatible donors when using UKBBv2 array instead of clinical typing data to identify compatible donors, categorised by number of RBC alloantibodies identified in patient samples. (c) Increase in the number of compatible donors when using UKBBv2 array instead of clinical typing data for each complex alloantibody profile identified in the data from 3,146 patients with RBC alloantibodies.

Table 4.4 Examples of rare blood group phenotypes and number of homozygous individuals identified

Phenotype	rsID	ISBT allele	Homozygous Negative Individuals			
			COMPARE	DIS-III	INTERVAL	UK Biobank
Fy(a-b-)	rs2814778	FY*01N.01	5	1	351	7427
k-	rs8176058	KEL*01.01	7	1	73	variant uncalled
Lu(b-)	rs28399653	LU*01	9	2	2	226
Vel-	rs566629828	VEL*-01	1	0	3	104
Yta-	rs1799805	YT*02	6	6	55	710

2,682 Dutch donors in this study identified five compatible active Dutch donors and their donations are now being used to provide life-saving transfusion support.

4.7.3 Donor and patient health information

Due to the large number of genetic variants typed by the UKBBv2 array, the utility of this platform extends far beyond antigen typing for extended matching. For example, homozygosity for the variant NM_000410.3:c.845G>A in the HFE gene is the most common cause of hereditary haemochromatosis. A recent study showed that 21.7% of male and 9.9% of female UKBB participants with this genotype presented with iron-overload pathologies. Using the UKBBv2 array data we identified 78 (1%) donors in the trial set testing homozygous for this variant. This information provides blood supply organisations with an opportunity to reduce the risk of such pathologies developing in these individuals by recommending more frequent donations.

Currently there is no approved process for returning genotype information such as this back to donors and therefore the homozygous individuals identified in this study have not been made aware of these results. We argue that due to the clinical significance of these variants, the ease of typing for them using donor typing arrays, and the health benefits that can be conferred to the individual that blood services begin to put in place the ethics, procedures, and studies which will allow information on clinically actionable variants to be reported to donors through appropriate channels.

4.8 Discussion

To ensure transfusion safety, blood is routinely matched for ABO and RhD compatibility. However, matching is not generally applied to other RBC antigens and each year approximately 0.5 million patients become sensitised as a consequence.[88] These patients subsequently require extended matching of blood to prevent HTRs, which are occasionally lethal. This situation can become particularly precarious in transfusion-dependent patients and it has been shown that extended matching from the beginning of transfusion support brings immediate benefits by reducing the incidence of alloantibody formation.[96] To reduce the frequency of this serious hazard of transfusion, some blood supply organisations have introduced an extended matching policy for these patient groups. However, the cost of implementing generalised extended matching policies will not be justified without conclusive evidence from randomised trials of consequent reductions in morbidity or mortality. For

such trials to be possible, a comprehensive and affordable extended antigen typing assay is required to test large numbers of donors and patients.

Here we report a technology, already used to genotype millions of individuals, now optimised for donor typing. We show by genotyping 7,473 donors, 99.92% concordance between clinical and array antigen typing results in 103,326 comparisons across 44 clinically relevant RBC and HPA antigens. We observed one HPA and 72 RBC discordances between clinical and genetically inferred antigen type (0.08% of comparisons).

Currently there are no internationally recognised guidelines set for the validation of array based high-throughput antigen genotyping tests. Therefore, the BGC has developed an initial policy on how to deal with genotype-phenotype discordances in the event that genotyping becomes standard of care in the future.

In short, the principles underlying this policy are:

First, international guidelines stipulate that ABO grouping has to be performed on every donation. Considering the seriousness of ABO mismatched transfusions ABO typing will remain antibody-based. The concordance testing between phenotype- and genotype-based ABO grouping will be used as a quality assurance measure in addition to concordance testing for sex. However the BGC will continue to resolve discordant results for ABO using approaches outlined in my thesis.

Second, the international recommendation for typing blood cell antigens for systems other than ABO is to perform these tests on two independently obtained samples of blood before the results can be used in prescribing. The BGC entirely underwrites this recommendation and foresees that with a comprehensive genotyping array these results will be generated far more cost-efficiently for every donor. The results in Chapters 4 and 5 illustrate the methods by which discordances can be investigated in case large number of donors are genotyped and also shows a streamlined approach to resolving the discordances, respectively.

Third, in planning the next phase of the BGC four blood supply organisations (Australian Red Cross Blood Service, NHS Blood and Transplant, New York Blood Centre and Sanquin) are placing Thermo Fisher GENETITAN instruments in their clinical laboratories. In preparation of the study for obtaining FDA approval and CE-marking of the Axiom array for blood cell antigen typing approximately 80,000 blood donor samples will be genotyped. Based on the results presented in this thesis I expect that this will reveal approximately 600 discordances. By including samples from the South African National Blood Service and from New York in this validation study larger numbers of samples of non-European ancestry will be genotyped. It has been agreed that all the discordances will be resolved by the approaches outlined in my thesis. As illustrated in this thesis the results of this resolution

will, in a step-wise manner, improve the bloodTyper algorithm and the knowledge about the genotype-phenotype link.

Fourth, if the Axiom array becomes an accredited and licensed diagnostic product then the results will be used in a stratified manner based on the acceptability of error in antigen typing, in the following categories:

- A quality assurance result only, e.g. the genotype determined ABO grouping results.
- A result which can be used for the labelling of units of blood, e.g. this would apply to all common red blood cell antigens and selected set of the rare antigen types.
- A screening results which requires confirmation by an orthogonal antigen typing test, which can be performed on the same DNA sample.

Fifth, the BGC foresees that if the Axiom test obtains regulatory approval then there will be a transition phase in which there will be an immense opportunity to identify further discordances. This is possible because of the wealth of extended typing data already available to blood supply organisations. As an example, the four organisations which will introduce genotyping next year jointly provide services for a population of approximately 120 million. In aggregate, these services have a donor base of almost 2.5 million active donors. Genotyping of 20% of these already registered donors would result in 0.5 M donor records with antigen genotyping results. The results from Chapter 4 show that an estimated 75,000 of these donors will have extended donor typing results. In addition in the UK the Early Disease Detection Research Project UK (also named ADD) will genotype 5 M UK citizens. It is expected that at least 0.5M of these study participants will be NHSBT donors.

Finally, the BGC has agreed to continue the analysis of discordant results after regulatory approval has been obtained. If approval has been obtained Royalty funding will flow from Thermo Fisher to the BGC providing funds to maintain the BGC collections of samples and data and for resolving discordances. The results of the discordance resolution will be shared in the public domain and made available to the International Society for Blood Transfusion and the Histocompatibility and Immunogenetics Societies to support future improvements of international standards for DNA-based blood cell antigen typing.

In conclusion, I foresee that by 2025 a tipping point will have been reached where erroneous types generated by the current clinical typing tests will far exceed those made by genotyping. I do however realise the important challenge of maintaining global collaboration on data sharing, which is particularly important if donors of non-European ancestry are genotyped at scale. As outlined above the BGC may provide the framework for continued international collaboration.

The question of whether extended typing of blood donors improves the efficiency of matching blood units to patients with multiple RBC alloantibodies has been much debated. We observed a 2.6-fold increase in the number of donors identified to support 3,146 complex patients with multiple alloantibodies when extended genotyping data were made available. The immediate clinical benefit of having densely typed donors is demonstrated by a female patient with severe anaemia due to bone marrow failure, who required blood with an antigen typing profile possessed by 1 in 400 donors. Notwithstanding this relatively high frequency, no suitable donors were present in the Netherlands because donors are not routinely typed for Co^a and Wr^a because reliable or affordable typing reagents are lacking. Five compatible Dutch donors were identified in the trial set of this work.

We did not alter the UKBB array content at the HLA locus because previous studies suggest that HLA class I and II antigens can be imputed with good accuracy from the original UKBB array data. However, the level of accuracy has never been empirically verified by direct comparison of array versus clinical-grade HLA typing data.[69] We provide this comparison and show excellent concordance at a 99.03% accuracy at two-field resolution. Typing at this resolution has immediate value for the provision of HLA class I matched platelet concentrates for refractory patients and shows that the platform if widely applied could be used to improve the composition of platelet apheresis panels via identification of donors homozygous for rare HLA class I haplotypes. Furthermore, the results can be used to select donors with HLA haplotypes underrepresented in international bone marrow registries for confirmation of their HLA type by next-generation sequencing.

We used the *HFE* locus to illustrate that the array data can also be used to inform policies about donor health and confirm that approximately one percent of donors are homozygous for the haemochromatosis causing variant in the *HFE* gene. Based on this observation an estimated 12,000 individual donors in England and the Netherlands could donate more frequently and reduce their risk of developing ill-health, whilst adding to blood stocks.

Reflecting the donor registries of The Netherlands and Britain, only 0.7% of donors available for this study were of non-European ancestry. Due to the low frequency of RH variant antigens in European ancestry individuals and their high frequency in haemoglobinopathy patients of African ancestry further validation must be performed before this test can be used to clinically type these patients, and required donors, for variant antigens. However, the complete concordance between 835 antigen typing results for the 57 non-European ancestry individuals and identification of two erroneous serological typing results due to variant RhD expression reported here provides justification for further studies using samples from donors and patients recruited by ancestry and specific blood antigen types.

All together we present an affordable genotyping assay capable of simultaneously typing all clinically relevant human red blood cell group antigens that has been trialled using thousands of blood donors. As the study was embedded in the national blood services of England and The Netherlands, we were able to take a first look at how high-volume comprehensive donor typing can be used to simplify the provision of compatible blood for sensitised patients with alloantibodies against multiple red blood cell antigens.

With many health care systems making preparations for the era of precision medicine, an increasing number of countries are genotyping a proportion of their population. The assay presented here is nested within the platform already used by the FinnGen Biobank, Million Veteran Program, Taiwan Precision Medicine Initiative, and the impending UK 5 Million studies. This means that full blood cell antigen types will become available for millions of individuals, providing blood supply organisations with the opportunity to implement a policy of genomics-based transfusion medicine.

Chapter 5

Typing genetically complex antigens

5.1 Abstract

In this chapter, we present the Blood transfusion Genomics Capture (BGC) capture assay that is designed for the targeted sequencing of the genes underlying the blood cell antigens in humans. The BGC capture assay design includes all consensus coding sequence and 5' and 3' UTRs present in Ensembl and RefSeq for the 54 relevant Red Blood Cell (RBC), Platelet (PLT) and Neutrophil (NEUT) antigen encoding genes. This next-generation sequencing (NGS) test has been used by the BGC to aid the resolution of the discordant results between the blood cell antigens inferred from UK Biobank version 2 (UKBBv2) genotyping results and the clinical antigen types that we reported in Chapter 4. The test also targets the regulatory elements (REs) that control these genes as determined by the BLUEPRINT of Haematopoietic Epigenomes project and by BLUEPRINT-related studies which aimed to identify the long-range interactions between gene promoters and these REs.[83, 101, 84] In total 1,094,363 bp are targeted for sequencing by the BGC capture NGS assay and this discontinuous genomic space is referred to as regions of interest (ROI) hereafter.

Two validation experiments were carried out to determine the performance of the assay; In one, 48 samples for which whole-genome sequencing data at 30x coverage (WGS) were available and the samples were also sequenced using the BGC capture assay, followed by a direct genotype to genotype comparison. In the other, DNA samples from 24 NHSBT blood donors were tested on the BGC assay and the genotype-inferred antigen types were compared with the antigen types retrieved from electronic donor records. Excellent concordance was observed in the validation experiments between results for both genotype (99.65%) and phenotype (99.05%).

We then used the BGC capture assay to sequence DNA samples for which we reported discordance between UKBBv2 array and clinical antigen typing results in Chapter 4. The

73 discordances reported in the previous chapter were observed in the results for 69 donors, with DNA samples for 61 of them being available for recall and BGC capture sequencing. The bloodTyper algorithm was used to infer antigen types from the BGC capture data and a three-way comparison between UKBBv2 array, BGC capture, and donor record antigen types were performed. In addition, the genetic antigen typing reports and the donor record antigen typing reports were reviewed by experts from the BGC and the donor typing laboratories of NHSBT and Sanquin.

With this approach, the cause of 78% of discordances could be identified. Hence the BGC capture assay is a viable second-line test for the resolution of discordances between UKBBv2 array inferred and clinical antigen types. We also show that the capture test has some advantages for resolving discordances over alternative capture tests like WES.

5.2 Introduction

To continue improving the design of array-based donor genotyping assays and accompanying blood group interpretation software, it is essential to further investigate samples for which discordances between genotype and phenotype are observed. There are several possible reasons for discordances between array genotype inferred antigen typing results and antigen types obtained by currently used typing tests, which are mainly antibody-based tests for red blood cell (RBC) antigens;

The first concerns individuals with *de novo* variants that control antigen expression or novel combinations of known antigen encoding variants which have not been documented by the ISBT. For the former, previous studies suggest we can expect approximately 70 *de novo* variants per generation with 90% being single nucleotide polymorphisms (SNPs), 9% being short insertions or deletions (indels), and 1% being structural variants (SVs) - there is no way to preemptively include these variants on array designs and they will therefore go untyped.[53] For the latter, algorithmic inference of blood type is extremely difficult as the link between phenotype and genotype is incomplete and limited haplotype frequency data is available to allow accurate phasing of the variants.

The second concerns antigens for which it has been widely reported that types determined by haemagglutination using commercially available or in-house typing reagents may be erroneous. Weak expression of an antigen, expression of a rare variant antigen and contamination of polyclonal typing reagents with other specificities are well-documented causes of erroneous types.

The third involves the incorrect calling of genotypes. In the case of arrays we have already reported in chapter 4 that in some cases array probes simply do not work and have

to be excluded from analysis, and that on some array designs probes for genotyping critical antigen defining variants are entirely lacking.

Finally, the calling of structural variants (SVs) from array genotyping data remains challenging and from most types of sequencing data. To reliably call SVs breakpoints must be characterised at nucleotide resolution. It is particularly difficult to accurately type SVs using array data as when DNA is hybridised to probes on the array it has been fragmented and only a single nucleotide position is interrogated. This means that although genotyping of specific variants is possible post DNA fragmentation, information about genome structure is lost (see Fig. 5.1).

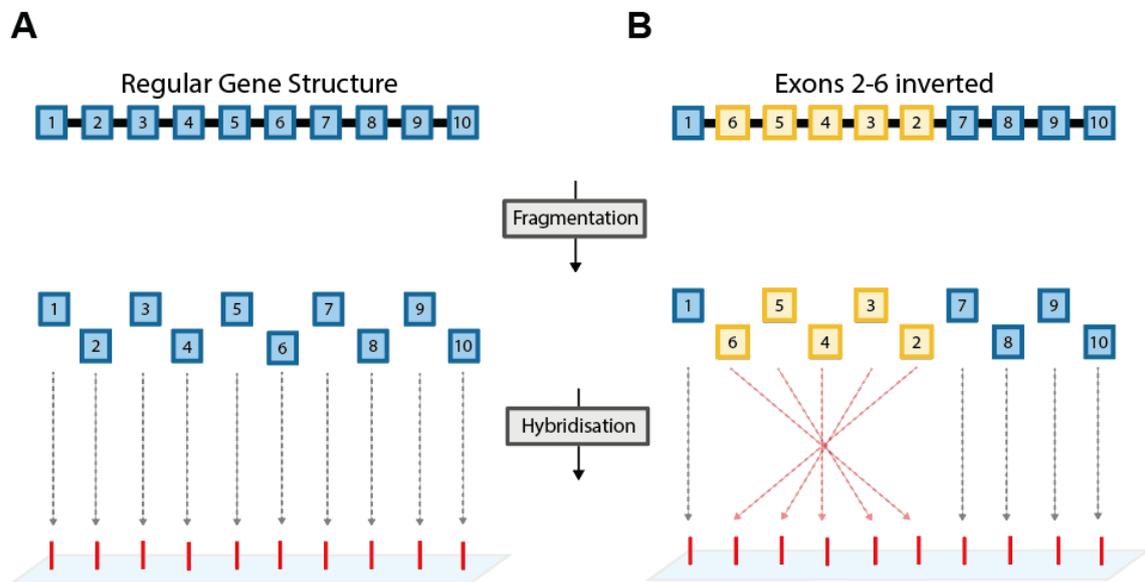


Fig. 5.1 Cartoon showing how genome structure information is lost in the array genotyping process. (a) Example genotyping of a gene with regular structure. (b) Example genotyping of a gene with exons 2-6 inverted. Here it can be seen that fragmented sections of the gene hybridise to the array in their correct positions regardless of original gene structure. Although individual sites in exons 2-6 of both genes would have genotyped correctly, structural information is lost.

The limitations of the array with respect to the identification of SVs require careful consideration when inferring antigen status. In chapter 4 we showed that reliable calling of the deletion type SV which underpins the D-negative phenotype required specific alterations to the array design and bloodTyper algorithm. Specific array designs and algorithms for the 38 other phenotypes that we know are encoded by complex SVs do not exist, and therefore they cannot currently be typed array-based technology (see Table. 5.1). The same applies to novel SVs which have not yet been linked to antigen expression.

Table 5.1 ISBT alleles which cannot be typed using current array technologies

System	ISBT Allele	Molecular Configuration
H	FUT2*0N.03	Fusion gene between <i>FUT2</i> and <i>SEC1</i>
	FUT2*0N.04	Fusion gene between <i>FUT2</i> and <i>SEC1</i>
MNS	GYP*101.01	<i>GYP A</i> :c.233_271del with <i>GYP B</i> translocation
	GYP*101.02	<i>GYP A</i> :c.232G>A with <i>GYP B</i> translocation
	GYP*101.03	<i>GYP A</i> :c.233_271del with <i>GYP B</i> translocation
	GYP*101.04	<i>GYP A</i> :c.58G>T; <i>GYP A</i> :c.67A>T; <i>GYP B</i> ;c.233_271del
	GYP*201.01	GYP(A1-232–B233-312)
	GYP*202.01	GYP(A1-232–B233-312);c.239C>T
	GYP*203.01	GYP(A1-271–B272-369);c.59C>T;c.71G>A;c.72T>G
	GYP*301.01	GYP(A1-159–Bψ160-177–A178-450);c.191T>A
	GYP*301.02	GYP(A1-159–Bψ160-177–A178-450)
	GYP*302.01	GYP(A1-202–Bψ203–A204-450)
	GYP*302.02	GYP(A1-202–Bψ203-212–A213-450);c.203G>C;c.212A>C
	GYP*303	GYP(A1-238–B239-242–A243-450);c.239G>C;c.242T>G
	GYP*401	GYP(B1-136-A137-354)
	GYP*402	GYP(B1-175-A176-354)
	GYP*501	GYP(B1-136-Bψ137-204-A205-229-B230-366)
	GYP*502	GYP(B1-136-Bψ137-204-A205-229-B230-366);c.236C>G
	GYP*503	GYP(B1-136-Bψ137-210-A211-229-B230-366)
	GYP*504	GYP(B1-136-Bψ137-159-A160-232-B233-369)
	GYP*505	GYP(B1-12-A13-78-B79-168)
GYP*506	<i>GYP B</i> (1-26)- <i>GP ψ B</i> (27-54)- <i>GPYA</i> (55-57)- <i>B</i> (58-103)	
<i>GYP A</i> *01N	<i>GYP A</i> :c.38_453del and <i>GYP B</i> :c.1_37del	
<i>GYP B</i> *01N	<i>GYP B</i> :c.38_276 and <i>GYP E</i> :c.1_37del	
GYP*01N	<i>GYP A</i> :c.38_453del; <i>GYP B</i> :c.1_276del	
RH	RHCE*01.29	<i>RHD</i> exon 4-9
	RHCE*01.34	<i>RHD</i> exons 4-7
	RHCE*03.02	<i>RHD</i> exons 2-3
	RHD*01N.07	<i>RHCE</i> exons 4-7
	RHD*01N.42	<i>RHCE</i> exons 1 and <i>RHCE</i> exons 7-10
	RHD*01N.43	<i>RHCE</i> exons 1-3
	RHD*01EL.23	<i>RHCE</i> exons 5-7
	RHD*01EL.44	<i>RHCE</i> exons 4-9
	RHD*01N.02	<i>RHCE</i> exons 1-9
	RHD*01N.03	<i>RHCE</i> exons 2-9
	RHD*01N.04	<i>RHCE</i> exons 3-9
	RHD*01N.05	<i>RHCE</i> exons 2-7
	RHD*01N.06	<i>RHCE</i> exons 4-7;c.733C>G;c.1006G>T

In previous work we have shown that sequence data generated by WGS and WES can be used for the accurate and comprehensive typing of blood cell antigens and studies have documented the use of WGS to discover and confirm the link between genotype and antigen expression.[102] Due to the ability of short-read Illumina NGS to genotype almost any position in the genome, thus being able to genotype novel variants, the technology could be applied as a second-line test to investigate discordances reported by first-line screening assays such as the UKBBv2 array.

Currently, the cost per sample of WGS is still too high for its widespread use by blood supply organisations for resolution testing. WES is a cheaper option than WGS, however, these assays are not designed with antigen typing in mind and often do not adequately sequence regions of the genome important for blood cell antigen typing (see Fig. 5.2). Furthermore, a recent comparison of the quality of WES data sets against the ‘gold’ standard coverage of WGS showed a significant difference in sensitivity for detecting SNVs and indels between different laboratories providing WES services.[60]

Targeted sequencing tests present a third short-read NGS option. The underlying principle of the test is similar to WES, differing in that the baits used to capture the region of interest (ROI) are customised and application-specific. The cost and turn-around-time of capture sequencing tests are similar to WES, and because the ROI is much more specific and on average 50 times smaller than for WES, very high sequencing coverage can be achieved for the ROI.

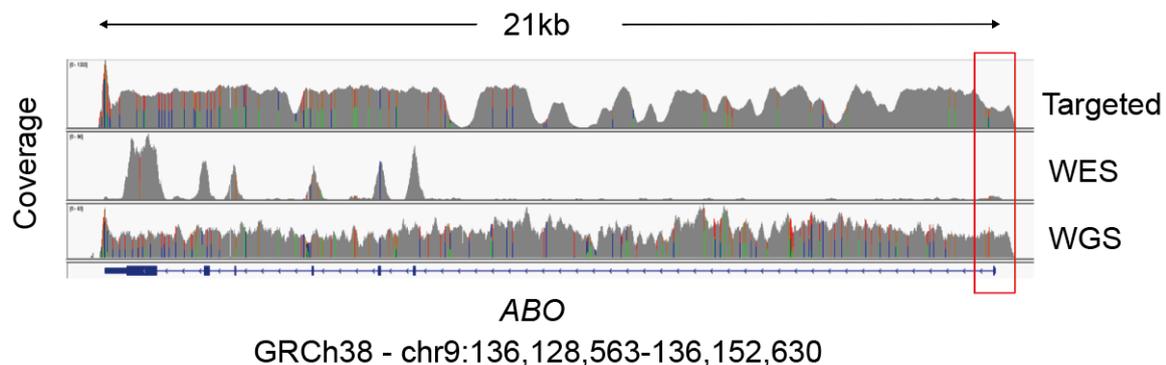


Fig. 5.2 Read depth profiles of Targeted, Whole Exome and Whole Genome sequencing across the *ABO* gene of the same DNA sample. The gene structure of *ABO* is shown in blue at the bottom of the figure, thick and thin sections represent exons and introns, respectively. A red box is drawn around exon 1 to highlight the poor coverage of WES in this region.

In this chapter, we detail the development, validation and use of the BGC capture NGS assay. We show its utility in identifying the cause of most of the discordances observed in

Chapter 4 between UKBBv2 inferred antigen types and clinical antigen types retrieved from the donor record.

5.3 Chapter workflow

In order to investigate samples producing discordant results by array technology, we developed the BGC capture NGS assay. This assay is a targeted sequencing test designed to sequence genes underpinning RBC, PLT and NEUT antigen expression. The work reported in this chapter has been split into three stages (see Fig. 5.3).

Firstly, the assay was designed by overlapping knowledge from the Transfusion Medicine field with the data produced by population genomics and epigenomics studies.

Secondly, we performed two validation experiments - one in which samples were sequenced using the BGC capture NGS assay and resulting genotypes were compared with clinical-grade WGS data, and another in which DNA samples from NHSBT blood donors were sequenced using the BGC capture assay and their genotype-inferred antigen types compared to the clinical ones held in the electronic donor record.

Finally, we sequenced DNA samples for which discordances were observed in our previous UKBBv2 array experiment described in chapter 4. The cause of the discordant results was then investigated on a case by case basis.

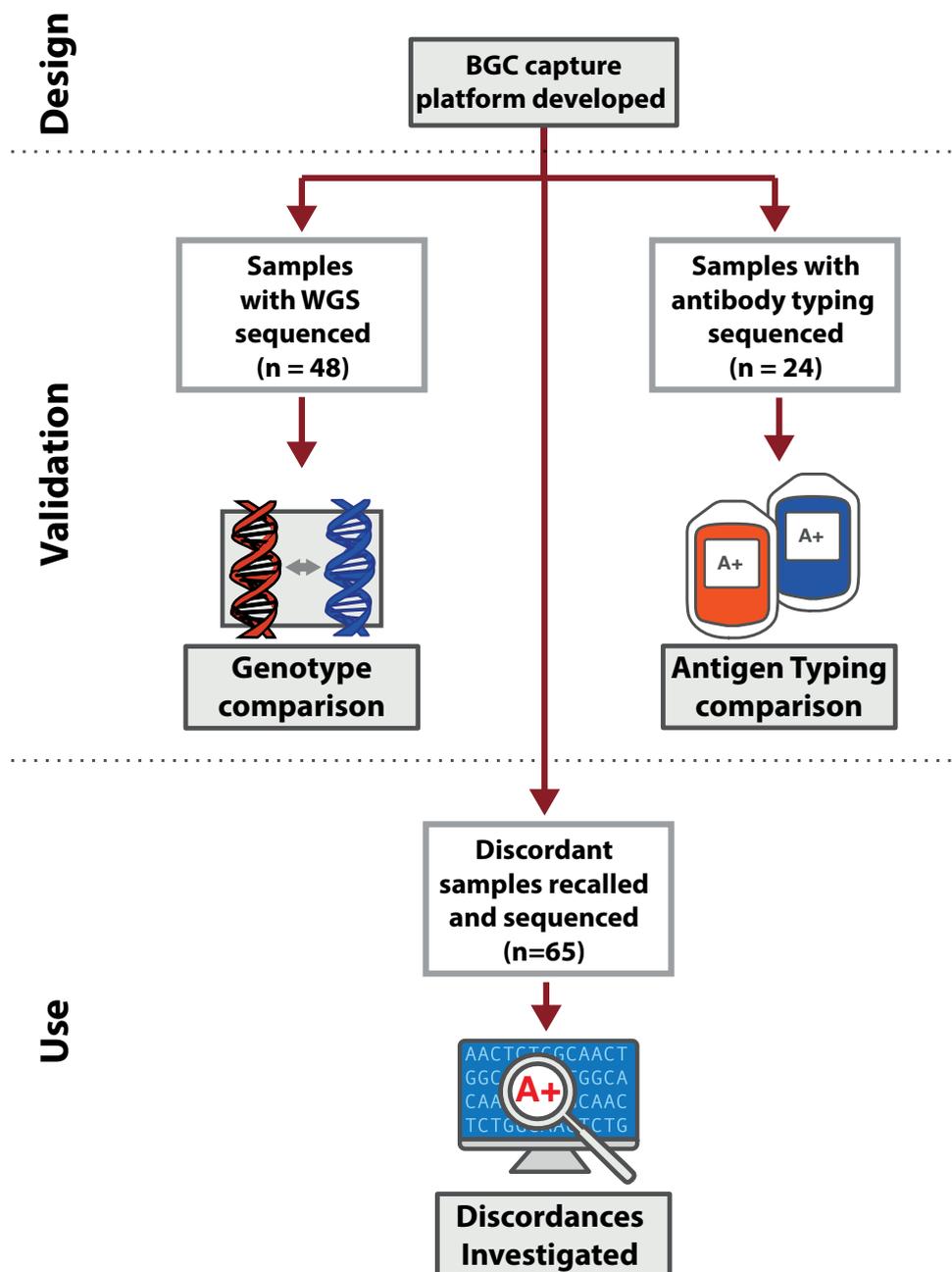


Fig. 5.3 Cartoon representing the overall workflow of this chapter. From top to bottom; **(Design)** Sequencing targets were identified using multiple data sources and combined to produce a unified list of sequencing targets for which capture baits were then synthesised. **(Validation)** Two separate validation experiments were carried out, one in which we compared BGC capture genotype results to those generated by WGS and another in which we compared antigen types inferred using the BGC capture NGS assay to the clinical ones obtained from the electronic donor record. **(Use)** The BGC capture assay was then used for its intended purpose, to resolve discordant results between the UKBBv2 inferred antigen types and clinical antigen types.

5.4 Assay design

The BGC capture NGS assay is designed to sequence all 54 (in 2018) relevant RBC, PLT and NEUT antigen encoding genes identified by the BGC panel of experts. It was decided for 22 genes to capture the entire gene body (exons and introns) and the 1000 bp upstream of the transcript start sites (named complete coverage hereafter). For the remaining 32 genes the consensus coding sequence (CCDS), the 5' and 3' UTRs present in Ensembl and RefSeq and the 1000 bp upstream of the transcript start sites are targeted (named exon coverage hereafter). The decision whether to apply complete or exon only coverage was based on a ranking of the clinical importance of the antigen systems, the complexity of the relationship between gene sequence and antigen expression, and the size in bp of the entire gene body (see Table. 5.2). We also targeted the REs identified in human erythroblasts and known to physically interact through long-range interactions with the promoters of the genes known to be relevant for RBC antigen expression.[101, 84] Additionally, the *HBB* gene is targeted so that the BGC assay results can also be used to determine whether a DNA sample carries the minor allele of variant rs334, which if present in homozygosity is causal of Sickle Cell disease. The final design was reviewed by several BGC experts and in total the targeted ROI for capture and sequencing is 1,094,363 bp.

Table 5.2 BGC capture target genes

HGNC symbol	System / Group	Antigen Type	Coverage selected
<i>ABO</i>	ABO	RBC	Complete
<i>GYP A</i>	MNS	RBC	Complete
<i>GYP B</i>	MNS	RBC	Complete
<i>GYP E</i>	MNS	RBC	Complete
<i>RHD</i>	Rh	RBC	Complete
<i>RHCE</i>	Rh	RBC	Complete
<i>BCAM</i>	Lutheran	RBC	Exon
<i>KLF1</i>	Lutheran	RBC	Complete (gene <10kb)
<i>KEL</i>	KEL	RBC	Exon only
<i>FUT3</i>	Lewis	RBC	Complete (gene <10kb)
<i>ACKR1</i>	Duffy	RBC	Complete and GATA-1 binding site
<i>A4GALT</i>	P1PK	RBC	Exon
<i>SLC14A1</i>	Kidd	RBC	Exon
<i>SLC4A1</i>	Diego	RBC	Exon
<i>ACHE</i>	Cartwright	RBC	Complete (gene <10kb)
<i>CD99</i>	Xg	RBC	Exon
<i>XG</i>	Xg	RBC	Exon
<i>ERMAP</i>	Scianna	RBC	Exon
<i>ART4</i>	Dombrock	RBC	Exon
<i>AQP1</i>	Colton	RBC	Exon
<i>ICAM4</i>	Landsteiner Wiener	RBC	Complete (gene <10kb)
<i>C4A</i>	Chido Rodgers	RBC	Exon
<i>C4B</i>	Chido Rodgers	RBC	Exon
<i>FUT1</i>	H	RBC	Complete (gene <10kb)
<i>XK</i>	Kx	RBC	Exon
<i>GYP C</i>	Gerbich	RBC	Complete
<i>CD55</i>	Cromer	RBC	Exon
<i>CR1</i>	Knops	RBC	Exon
<i>CD44</i>	Indian	RBC	Exon
<i>BSG</i>	Ok	RBC	Exon
<i>CD151</i>	Raph	RBC	Complete (gene <10kb)
<i>SEMA7A</i>	John Milton Hagen	RBC	Exon
<i>GCNT2</i>	I	RBC	Exon
<i>B3GALNT1</i>	Globoside	RBC	Exon
<i>AQP3</i>	Gill	RBC	Complete (gene <10kb)
<i>RHAG</i>	Rh-associated glycoprotein	RBC	Exon
<i>GBGT1</i>	Forsmann	RBC	Exon
<i>ABCG2</i>	Junior	RBC	Exon
<i>ABCB6</i>	LAN	RBC	Complete (gene <10kb)
<i>SMIM1</i>	VEL	RBC	Complete
<i>CD59</i>	CD59	RBC	Exon
<i>SLC29A1</i>	AUG	RBC	Exon
<i>FUT2</i>	Secretor	RBC	Complete (gene <10kb)
<i>ITGB3</i>	HPA-1,4,6,7,8,10,11,14,16,17,19,21,23,26,29	HPA	Exon
<i>GP1BA</i>	HPA-2	HPA	Complete
<i>ITGA2B</i>	HPA-3,9,20,22,24,27,28,Lap ^a	HPA	Exon
<i>ITGA2</i>	HPA-5,13,18,25	HPA	Exon
<i>GP1BB</i>	HPA-12	HPA	Complete (gene <10kb)
<i>CD109</i>	HPA-15	HPA	Exon
<i>FCGR3B</i>	HNA-1	HNA	Complete (gene <10kb)
<i>CD177</i>	HNA-2	HNA	Complete (gene <10kb)
<i>SLC44A2</i>	HNA-3	HNA	Exon
<i>ITGAM</i>	HNA-4	HNA	Exon
<i>ITGAL</i>	HNA-5	HNA	Exon
<i>HBB</i>	-	Sickle Cell	400 bp region covering - NM_000518.4:c.20A>T

5.5 Validation of BGC targeted sequencing assay

Following the selection of targets, the BGC capture NGS assay design was submitted to ROCHE NimbleGen, Inc. (Madison, WI) and DNA baits were produced. We then used these baits to perform two validation experiments.

5.5.1 Comparison to WGS data

DNA samples from 48 individuals enrolled in the NIHR BioResource and for whom 30x WGS was already available were selected at random and sequenced using the BGC capture NGS assay. Sequence reads were aligned to the GRCh37 reference sequencing using BWA 0.7.10. These alignments were used to assess the coverage profile produced by the BGC capture assay. The average read depth for the entire ROI across all samples was 696.54 (range 0 to 2356.290). Overall coverage of the target region was 99.2% at >30x read depth (see Fig. 5.4a). Coverage of all regions relevant for blood cell antigen typing was 99.91% at >30x read depth and 100% at >15x read depth. We also calculated the sequence coverage profile for each base of each target gene for the 48 samples, which were multiplexed in a single sequencing batch (see Fig. 5.4b). This showed that the antigen encoding space of each individual sample was covered adequately.

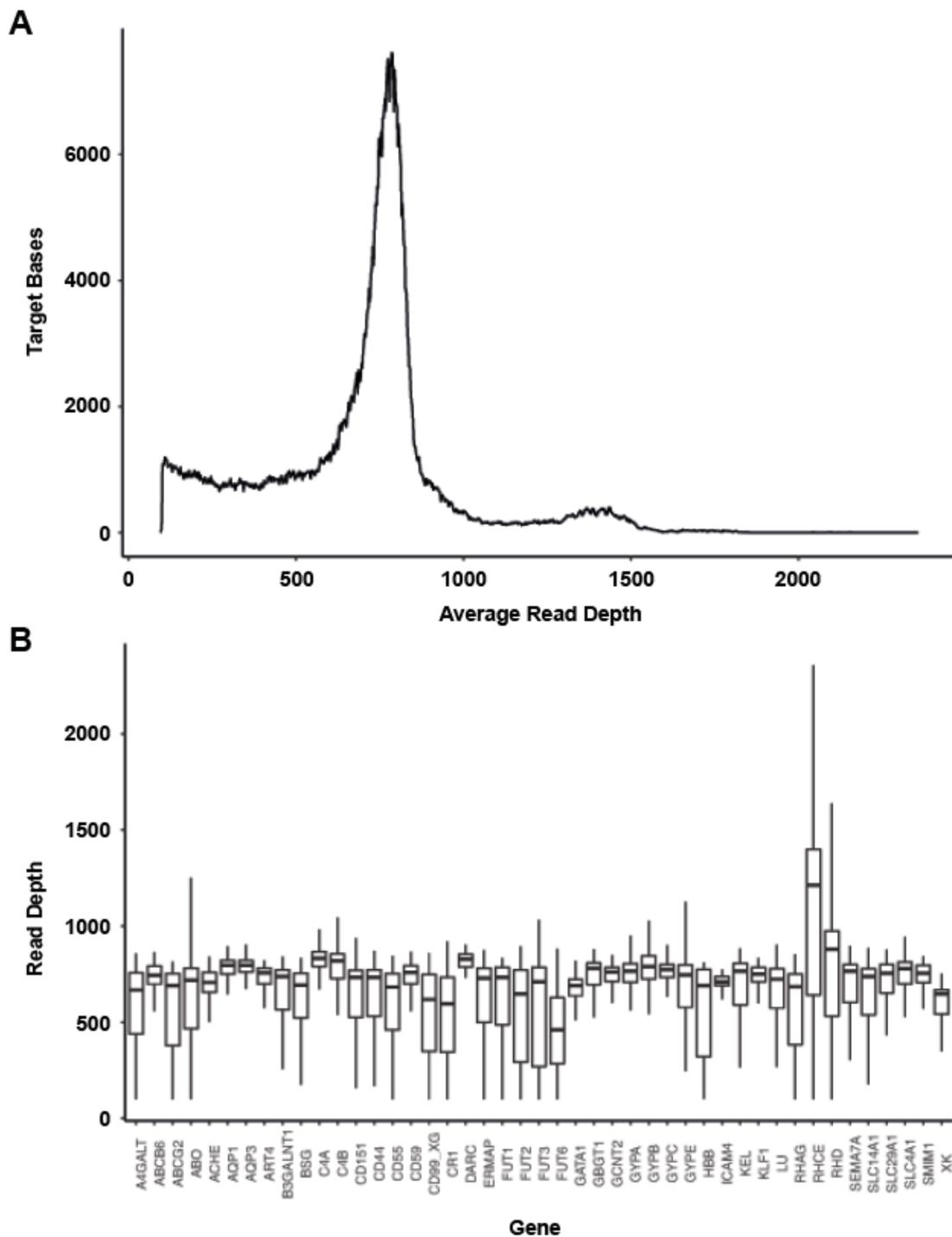


Fig. 5.4 Sequencing coverage of blood cell antigen ROI for the BGC capture NGS assay. (a) Distribution of average read depth across the 48 samples for all targeted bp in the ROI of the BGC assay. (b) Read depth across the 48 samples for each bp in genes relevant for antigen expression on RBCs.

SNVs and indels were called for each of the 48 samples using GATK 3.3 HaplotypeCaller. Sex was then inferred using two statistics based on sequence reads aligned to well-covered target regions (>95% of samples covered at 30x): the ratio between heterozygote and non-reference homozygote genotypes (het/hom) on the X chromosome sites, and the ratio between the median coverage on X and the median coverage on the autosomes (aut/X).[103] The het/hom ratio is computed using heterozygote SNVs with an allele depth between 0.3 and 0.7 to guard against errors. No discordant results were observed when comparing genetically inferred sex for each of the 48 samples versus the self-declared sex, and all were progressed to the next stage of analysis.

Genotype to genotype comparison was then performed between the sequencing results obtained by the BGC capture array and by WGS for each of the 48 samples. Samtools 1.9 was applied to extract reads mapped to the BGC capture ROI from the WGS alignments for each individual. The sequence reads for all samples from the two different sequencing tests were then individually re-aligned to the GRCh37 reference sequence using BWA v0.7.10 and variants were called from subsequent alignments with GATK HaplotypeCaller v3.3 using standard parameters. A stringent filtering step (Mean Mapping Quality < 40.0 || Quality by Depth < 2.0 || Fisher Strand > 100.0 || HaplotypeScore > 13.0 || Mean Quality Rank Sum < -12.5 || Read Position Rank Sum < -8.0) was applied to all variants observed to reduce the chance of erroneous variant identification. Variants passing QC filters for all samples were then merged into technology respective variant call-sets, one named the BGC capture variant call-set and the other the WGS variant call-set. Both call-sets were trimmed by removing variants with <99% overall genotype call rate and then an intersection performed to ensure that each had identical variant contents. This left a total of 3,689 variants (3356 SNVs, 332 indels) for the final comparison experiment.

Concordance between genotypes was 99.65% (172,921 correct in 173,528 comparisons). For 401 (66%) of the 607 discordant genotypes the cause of the error seemed to be due the WGS results, where a sample with a major/minor genotype had been identified as major/major instead - this type of error is expected as the BGC capture NGS assay platform produces much higher read coverage in the ROI compared to WGS (see Table. 5.3). It is therefore a more sensitive test for calling the presence of a minor allele, in particular for regions where coverage of a region by WGS is below 30 reads. The discordance of the remaining 206 genotypes (34%) was most commonly caused by lower quality mapping of reads which in Illumina NGS can occur for numerous reasons that are previously reported.[104] No variants for which discordant genotyping results were observed between the BGC assay and WGS were at variant positions associated with antigen definition according to the ISBT or IPD databases for RBC and HPA antigens, respectively.

Table 5.3 Genotype to genotype comparison between BGC capture and WGS data

		BGC capture			
		MM	Mm	mm	no call
WGS	MM	103753	410	15	371
	Mm	31	55335	151	619
	mm	0	0	13833	1487
	no call	14	28	99	0

After having observed that the BGC capture NGS assay could accurately measure genotypes in the antigen typing space, we next sought to validate the assay for its ability to correctly produce genotypes from which to infer antigen types.

5.5.2 Comparison to clinical antibody typing

DNA samples from 24 NHSBT blood donors enrolled in the NIHR BioResource were selected due completeness of RBC antigen typing data in their electronic donor record and availability of an adequate amount of DNA in the NIHR BioResource sample repository. The 24 DNA samples were retrieved from the repository for analysis by the BGC capture NGS assay. Sequence reads were aligned to the GRCh37 reference sequencing using BWA 0.7.10.

Average read depth over the ROI for the 24 samples was 524.22 (range 0 to 2106). Overall coverage of the ROI was 98.7% at >30x read depth. Coverage of the antigen typing specific ROI was 99.89% at >30x read depth and 100% at >15x read depth. Sex was inferred from genotype using the method described in the previous validation step and the results were concordant with self-declared sex for the 24 samples.

Alignments of the sequence reads were analysed using the bloodTyper exome sequencing pipeline with limited modifications to adjust for the increased read depths of the BGC assay results compared with those from WES. Inferred blood cell antigen types were compared to the clinical ones retrieved from the electronic donor record.

Concordance between BGC capture assay inferred and donor record antigen types was 97.94% in 535 comparisons across 24 RBC antigens for which clinical typing data was available (see Fig. 5.5).

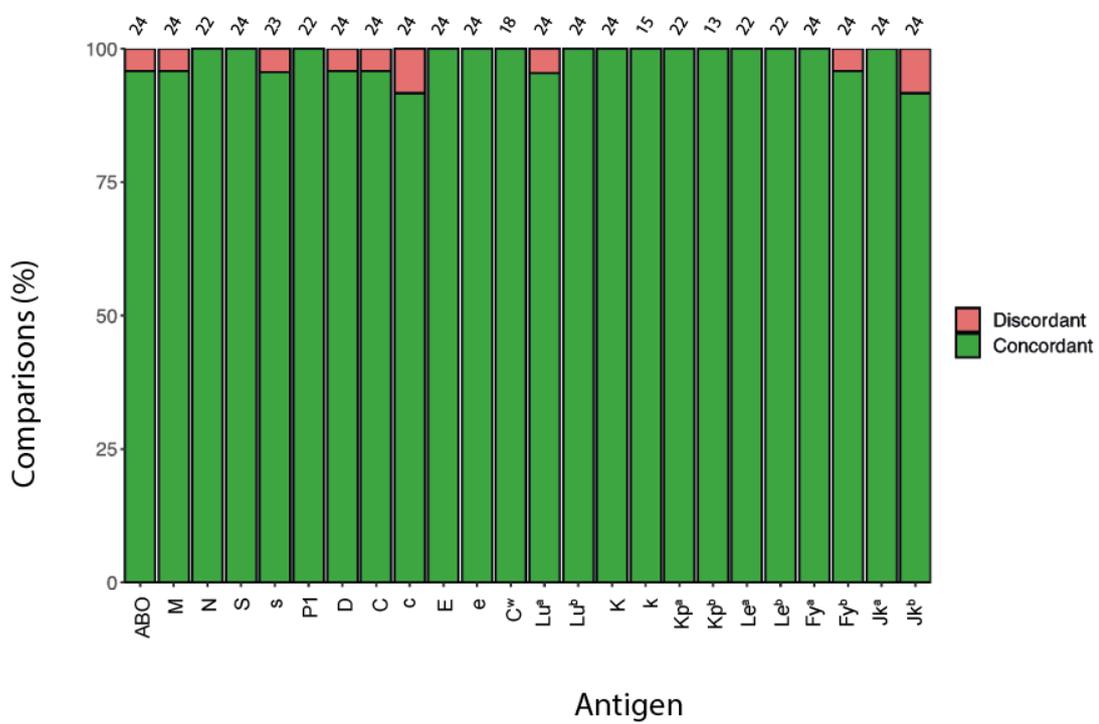


Fig. 5.5 Concordance between BGC capture inferred and donor record antigen types. Concordance per antigen is shown as a percentage of the total number of comparisons, given at the top of each bar, with concordant and discordant results in green and red, respectively.

Investigation revealed that 7 of the 11 discordances were identified for a single sample with discordant typing results for the ABO, D, M, s, Lu^a, Fy^b, Jk^b antigens. The most likely cause of this large number of discordances in a single sample is an error in data transfer or more likely an error in sample handling in one of the many sample handling steps between blood donation and generation of the BGC assay and WGS genotypes. Sanger sequencing of a selection of variants unique to the sample confirmed that the DNA retrieved from the sample repository was the same as the sample sequenced in this experiment. Without any further way of investigating the discordance between our results and those on donor record, we decided to remove this sample from further analysis. After removal of the aberrant sample, four discordant results remained for four different samples. Three of these concerned C/c antigen typing results. C/c antigen typing by NGS requires a specific algorithm, in brief; The variant controlling C/c expression lies in exon 2 of *RHCE* and homology between *RHD* and *RHCE* means variant calling by NGS in this region can be unreliable. The C genetic configuration of *RHCE* is more homologous to *RHD* exon 2 than the c configuration. Due to increased homology and an artefact of alignment algorithms, NGS sequence reads for *RHCE* exon 2 will map to *RHD* exon 2 for individuals carrying the C encoding variant. bloodTyper exploits this biological/technological artefact to type the C/c antigens by looking for deviations in read depth ratios across *RHD* and *RHCE* exons (see Fig. 5.6).

Following inspection of the average read depth ratios produced by the BGC capture assay across *RHD* and *RHCE* a small adjustment was applied to the bloodTyper algorithm to calibrate it, this resulted in the three discordances for C/c typing being corrected.

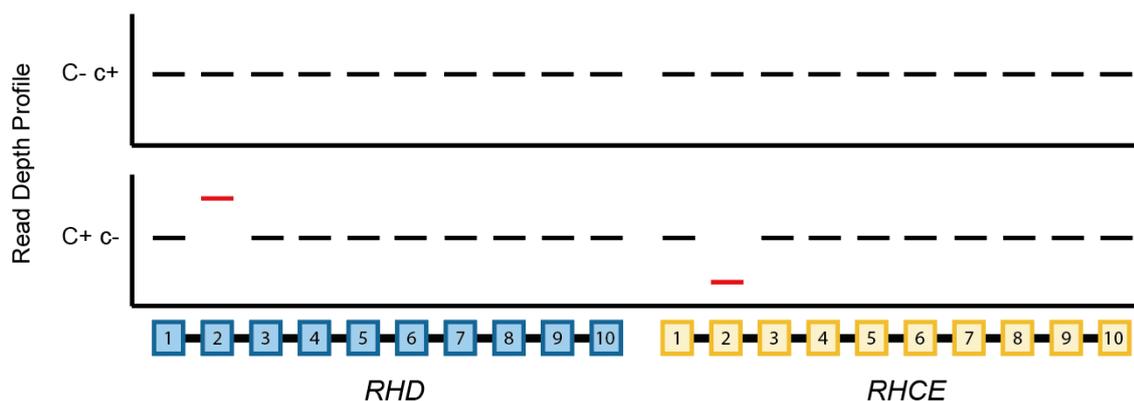
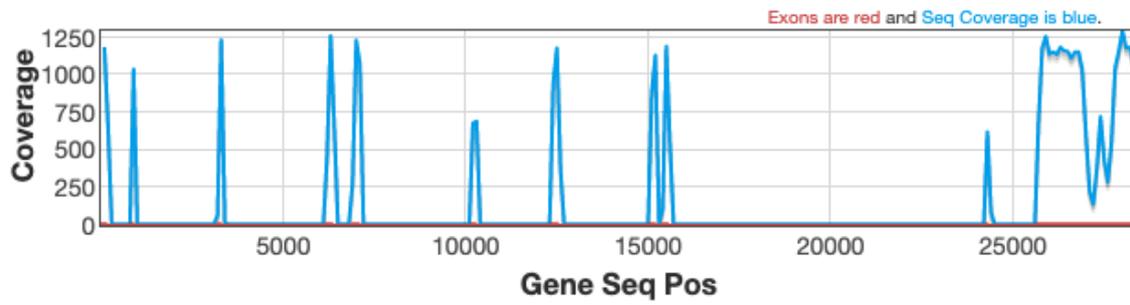


Fig. 5.6 Cartoon showing differences in read depth profile for different C/c antigen phenotypes. **(Top)** For a C-c+ phenotype sample, uniform read depth can be observed across all *RHD* and *RHCE* genes. **(Bottom)** For a C+c- sample, an increase in read depth over *RHD* exon 2 and decrease in read depth over *RHCE* exon 2 can be observed as a result of erroneous mapping of *RHCE* reads to *RHD* due to sequence homology. bloodTyper exploits this artefact to enable C/c antigen typing from NGS data.

The fourth and final discordant result concerns the Jk^b antigen. This sample was Jk^b+ by antibody typing and Jk^b- by genotype. Inspection of the raw sequencing data did not reveal a reason for this discordance. The variant for Jk^{a/b} typing, chr18:43,319,519G>A, was covered by 1161 unique reads (see Fig. 5.7). There is a possibility of allele drop-out with capture sequencing platforms, however it is unlikely in this case as other samples heterozygous for this variant have been detected correctly using the BGC capture assay. There was not enough DNA in the sample repository to perform analysis by an alternative sequencing approach, therefore the discordance for this donor remains unresolved. Expert members of the BGC consider it highly likely that the antibody determined antigen type may have been erroneous.

009 JK SLC14A1 (Predicted Antigens)



Region	10x	20x	30x	Avg. Coverage
gene	4247 / 28394 = 15.0%	4247 / 28394 = 15.0%	4247 / 28394 = 15.0%	152x
exon	4057 / 4058 = 100.0%	4057 / 4058 = 100.0%	4057 / 4058 = 100.0%	1008x

Found Alleles:

Low Coverage <10x per base A G C T del ins

Allele	ISBT	Antigen	Protein	CDS	Gene	Genome
JK*01	JK:1,-2	Jk(a+)	p.Asp280	c.838G	g.15428G	chr18:43,319,519G 1161x

Molecular: JK*01

ISBT Antigens: JK:1, -2

Phenotype: Jk(a+b-)

Fig. 5.7 Screenshot of the BGC capture bloodTyper report for the Jk^b discordant sample. 100% coverage of the *SLC14A1* gene exons can be observed. 0 alternate reads were detected at the Jk^{a/b} antigen defining site.

The results of the two validation experiments show that the BGC capture NGS assay can accurately measure genotypes across the entire ROI and that accurate antigen typing results can be inferred from the genotyping data produced. We proceeded to use the assay for its intended purpose, investigation of the DNA samples for which discordance was observed between UKBBv2 array and clinical antigen typing results.

5.6 Investigation of discordant array samples

In chapter 4 we reported 73 discordant results in 69 samples; for 67 only a single discordant antigen typing result was observed and two samples had two discordant results each. New aliquots of DNA were retrieved from the NIHR BioResource archive for 61 of the 69 samples (for the eight remaining samples the DNA stocks were inadequate). The 61 recalled DNA samples were sequenced using the BGC capture NGS assay and sequence reads were aligned to the GRCh37 reference sequence using BWA 0.7.10. The average read depth across the ROI for all samples was 489.11 (range 0 to 1650.54). Overall average coverage of the ROI was 99.1% at >30x read depth. Coverage of the regions specific to antigen typing was 99.93% at >30x read depth and 100% at >15x read depth for all samples. Genetically inferred sex versus self-reported sex was concordant for all samples.

bloodTyper was used to infer antigen types for each sample and a three-way comparison between UKBBv2 array inferred, BGC capture inferred, and electronic donor record antigen typing results was performed. Manual analysis of the genetic typing result was conducted by three independent experts of the BGC and all clinical antibody typing results from the electronic donor record were reviewed with experts from the typing laboratories of NHSBT and Sanquin.

Using the results of the BGC capture NGS platform we were able to identify the cause for all but 16 of the total 73 discordant results (see Fig. 5.8). Individual reports detailing the analysis of each discordant case are given in supplemental table 1 in the appendix of this thesis.

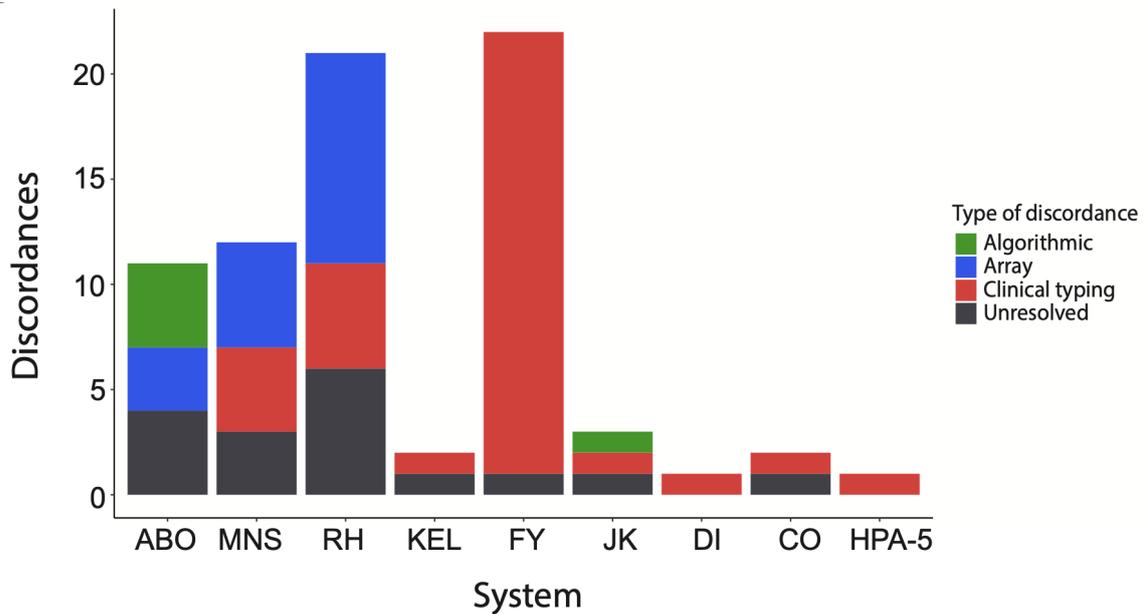


Fig. 5.8 Breakdown of discordant UKBBv2 samples according to reason for discordance and blood group system.

5.6.1 Algorithmic discordances

6 (6.9%) of the discordant results were caused by limitations of the bloodTyper algorithm version used for this analysis.

Four samples were ABO group O according to the electronic donor record and group A according to the results obtained with the UKBBv2 array. The UKBBv2 array detected heterozygous group A phenotype (ABO*A1.01) and group O phenotype (ABO*O.01.01) haplotypes for each of the four samples, leading to an overall group A typing result. Upon sequencing of these samples using the BGC capture assay, we identified additional heterozygous genotypes for the variants NM_020469.2:c.646T>A, NM_020469.2:c.681G>A, NM_020469.2:c.771C>T, and NM_020469.2:c.829G>A in the *ABO* gene.

Alone these variants usually encode an A_{weak} phenotype, however, when observed alongside another O phenotype (ABO*O.01.01) allele, as is the case in these samples, the variants underpin a second O phenotype (ABO*O.09.01/2) resulting in an overall group O typing result. Inspection of raw UKBBv2 array data revealed that all of these variants were accurately called in the affected samples. However, at the time of the UKBBv2 analysis detection of variant antigens, including A_{weak} , was disabled in the bloodTyper array analysis workflow to simplify comparison to donor records. This means that an incorrect result for these samples was reported. Variant antigen typing has since been enabled and these four discordances were resolved.

Additionally, bloodTyper reported Jk^b positive typing results for two samples that were serologically Jk^b negative. In both cases heterozygous genotypes were observed for the variant NM_015865.7:c.342-1g>a which underpins the Jk_{null} phenotype (JK*01N.06). Currently, the bloodTyper algorithm does not use this variant to infer phenotype due to lack of haplotype frequency data - in layman's terms, the algorithm will not decide if Jk^a or Jk^b is nullified for safety reasons. Instead, it issues a warning when the variant is detected that antibody-based confirmation of the Jk typing results is required.

5.6.2 Array discordances

18 (24.7%) of the discordant results were due to array genotyping errors which can be subdivided into three categories: *Incorrect genotype calls (n=5)*: For three ABO and two RH (e) antigen discordances incorrect genotypes were reported by the UKBBv2 array. In these cases, array genotype call confidence was barely above the QC threshold. Inspection of call plots revealed these calls were positioned almost exactly between cluster boundaries. BGC capture antigen typing results were in concordance with donor record typing results for these five samples, and the correct genotype calls were observed. Increasing UKBBv2 array genotype call QC thresholds eliminated these errors by producing no genotype call for affected variants in these samples. bloodTyper would then subsequently not infer antigen status and in practice, these samples would be flagged for re-genotyping or typing via alternative methodology.

Probeset issues (n=5): For these cases, the BGC capture antigen typing results were in concordance with donor record typing results for these samples, and the BGC assay genotyping results did not concur with those reported by the UKBBv2 array. This prompted inspection of the array genotype call plots and revealed that the probes, whilst performing adequately, required further re-design to increase cluster resolution. The M, N, S and s antigens were those affected by this type of error.

Lack of optimum typing variants (n=8): All cases here concerned typing results for the C antigen of the RH system. The BGC capture assay antigen typing results were in concordance with donor record typing results for these samples, and genotypes reported prompted inspection of UKBBv2 array raw genotyping data. Currently, C antigen status is inferred using the variant NM_020485.5:c.307C>T, which directly encodes antigen expression. Many DNA-based technologies fail to accurately type this variant due to extremely high homology between the *RHD* and *RHCE* genes, particularly in exon 2 where this variant is localised. A 109 bp insertion in intron 2 of the *RHCE* gene, located at NC_000001.10:25732083-25732084 (GRCh37), which has been classically used for DNA-based C antigen typing has no working probeset on the UKBBv2 array. Using two confirmatory variants for C typing is

the best strategy to improve accuracy. Improved probesets for typing the 109 bp insertion has been included in the next version of the array.

5.6.3 Error in donor record antigen typing data

33 (45.8%) discordances were explained by erroneous clinical typing results. These errors can be subdivided into two categories:

Individuals with variant antigen expression (n=19): In 19 of these cases, alleles encoding variant antigen expression were detected by the array. Examples of these include the Del (RHD*11, NM_016124:c.885T), K_{mod} (KEL*02M.01, NM_000420.2:c.1088G>A) and Fy^x (FY*02W.01, NM_001122951.2:c.265C>T and NM_001122951.2:c.298G>A) phenotypes. For carriers of these alleles, the chances of false-negative antibody-based clinical typing results are greatly increased. The BGC capture assay antigen types and corresponding genotypes confirmed the results from the UKBBv2 array for these samples. A unit of blood being erroneously typed as negative for a given antigen, may boost antibody levels in a previously sensitised patient. We therefore reported these erroneous clinical typing results to the relevant blood services for follow-up investigation.

Error in clinical typing data (n=14): The remaining 14 discordances in this category involved antigens where current typing reagents have been known to give incorrect results (n=4) or where the BGC capture assay analysis confirmed array genotyping results (n=10). The most likely reason for discordance being erroneous clinical typing data.

5.6.4 Unresolved discordances

16 (21.9%) of the discordant cases remain unresolved. For six of cases, previously unobserved DNA variants which were likely to underpin an antigen negative phenotype were discovered using the BGC capture assay. The absence of these newly identified variants from the ISBT reference tables prevents their use in antigen phenotype inference by bloodTyper and these variants were not included in the UKBBv2 array design. To determine the effect of these six unique variants on antigen expression functional expression studies are required, therefore we regard these cases as unresolved. For the remaining 10 discordances we were unable to resolve the cause of the discordance due to lack of DNA samples for further analysis.

5.7 Discussion

In this chapter we presented the results obtained with the BGC capture assay, a novel targeted sequencing platform designed as a second-line test for resolving discordances between UKBBv2 array-based antigen typing results and the typing results from the electronic donor record. In order to validate the platform, we compared the data produced by the assay against WGS genotype data and against clinical antigen phenotyping data. Excellent concordance was observed between results at both genotype (99.65%) and phenotype (99.05%) resolution, following removal of one erroneous sample.

The BGC capture assay proved to be an indispensable tool for investigating and resolving most of the discordant results. We were able to resolve 78% (57 / 73) of discordances between UKBBv2 array and donor record antigen types. In the unresolved cases, we were able to identify 6 previously unobserved DNA variants, which are likely to underpin altered antigen phenotypes in 6 of the 16 discordant cases. For 10 donors we were not able to use the BGC assay due to a lack of DNA for further analysis.

Most commonly the BGC capture assay was used as an orthogonal confirmatory test either confirming array-based genotyping results or allowing for the identification of falsely reported genotypes. In several cases, the NGS data produced by the BGC capture assay revealed UKBBv2 array probesets that were passing standard QC filters, yet still reporting false 'borderline' genotype calls. This information can now be used to set array genotype calling confidences for those probes or to indicate that they need re-design. This will further improve the typing accuracy for all antigens concerned in future array designs.

Most importantly, in the cases we regard as still 'unresolved' the BGC capture assay has identified novel variants that could explain the genotype/phenotype discordance. Although these variants are not known to the ISBT, we can include them in the next UKBBv2 array design and bloodTyper will report that discordant typing results have previously been observed for samples carrying one of these variants. Furthermore, on the basis of these results research samples will be taken from these donors upon their next donation to allow for further analysis of altered antigen expression caused by these variants. Those new variants which can be linked to an altered antigen phenotype will then be reported to the ISBT for inclusion in the appropriate reference table.

While several targeted sequencing platforms have been reported previously, none of them targets RBC, HPA and HNA antigens simultaneously.[105] Not only does our assay provide excellent coverage of the genes underpinning these antigen systems, but it can also sequence the most recently identified regulatory elements for these genes. Furthermore, no previously published assays have included automated antigen interpretation software, instead, past researches have relied upon manual inspection of sequencing alignments. This type of

analysis is time-intensive and is unsuitable for use as a second-line test to investigate the one percent of samples with discordant genotype-phenotype results (chapter 4) which are being observed when applying the UKBB-v2 array.

During these experiments, we did not sequence any individuals for which antigen expression was underpinned by novel SVs, including complex ones. In other studies, we have applied long-range sequencing to not only accurately identify individuals with the U- phenotype, but also to characterise the exact break point location and nature of the structural variation underlying this clinically important phenotype for patients of African ancestry.[106] Knowledge of these breakpoints has now been encoded into the bloodTyper algorithm enabling accurate antigen typing by assessing NGS read mapping statistics. In future work, we plan to genotype/sequence DNA samples from blood donors with known structural variants on the UKBBv2 array, the BGC capture assay, and the Oxford Nanopore long-read sequencing platform allowing us to comprehensively validate and merge data from all three technologies.

In summary, we have designed and validated the BGC capture NGS assay for resolving samples with discordant results, a crucial tool for continued array development and validation as it allows for a high-resolution investigation of discordant typing results without the requirement to seek fresh blood samples from donors. Currently, this test is more affordable than analysis by WGS and has a significantly lower data footprint with alignments being on average 522 Mb per BGC capture sample compared to 90 Gb per WGS sample. The BGC platform also offers enhanced read coverage of antigen encoding genes in comparison to commercially available WES assays. As the test is combined with automated interpretation software it presents blood supply organisations with the ideal platform for routine investigation of patients with complex alloantibody profiles, blood cell antigens encoded by novel gene haplotypes and for resolving genotype/phenotype discordances where antibody-based phenotype is known. As outlined above an immense data richness that will be generated by typing donors with historical extended typing data will rapidly result in a steep diminishment of the number of discordances.

The BGC capture assay will also be an essential tool in ensuring the continued safe integration of genotyping technologies such as the UKBBv2 array into clinical service where it provides an accurate and affordable orthogonal method for confirmation of genotyping results for antigens where only limited validation has previously been performed due to rarity and lack of serological typing reagents or methods. Furthermore, as improved haplotype reference sequences become available for blood group encoding genes data from the previously discordant sample can be re-analysed in light of new knowledge to improve antigen typing algorithms that are currently confounded by the genotype phase.

Chapter 6

Discussion

As mentioned in the abstract of this thesis, the World Health Organisation reports that globally 118.5 million blood donations are collected each year.[107] This blood is used to provide life-saving transfusion support for millions of individuals with a wide range of medical conditions.

Blood transfusion is widely regarded as a safe procedure and this is due to the common policy of identifying and ensuring compatibility between the ABO and RhD antigens of donor and recipient for each transfusion. Although this policy prevents the majority of adverse Haemolytic Transfusion Reactions, sensitisation to non-self RBC antigens remains an unavoidable consequence of this matching strategy. The incidence of sensitisation after a single transfusion episode has been approximated at 3%.[87, 88] In patients who require chronic transfusion support, the incidence of immunisation can rise to 60%.[89–91, 63, 92] Considering the vast scale of global blood transfusion, this means that each year at least one million individuals will become sensitised as a direct result of the current 'one size fits all' approach and 180,000 of the 300,000 infants born with an inherited haemoglobin disorder will become sensitised due to the transfusion support they are dependant upon. This is of course assuming that the standard of clinical care enjoyed in developed nations is applied globally, which of course it is not.[107]

There is evidence to suggest that the frequency of immunisation events can be reduced by increasing the level of matching between donor and recipient blood groups, and some blood supply organisations have implemented this policy for patients who are most at risk of immunisation events.[96, 108, 97] In order to implement such a policy for all transfusion recipients, a large portion of donors needs to be typed for all clinically relevant red blood cell (RBC) antigens. Many automated antibody- and DNA-based donor typing assays have been developed, with studies reporting that 99.8% (5661/5672) of complex blood requests could be served using just 43,066 donors genotyped for only a limited number of

antigens.[94, 95, 36, 38] However, these tests are cost-prohibitive, do not type all clinically relevant antigens and have thus not been adopted by global blood supply organisations. This is exemplified by the fact that 85% of blood donors in England have no typing data clinically relevant to RBC antigens such as those of the Duffy, Kidd, MNS, and Lutheran systems and 94% of English donors in England have no typing data for rare antigens (information from look-up in NHSBT's PULSE database - 2019). Even though many advancements in the donor typing field have been made, the lack of typed donors persists and broadly most patients are still not benefiting from these developments.

In recent years the rapid fall in the cost of genome-wide DNA typing combined with the development of advanced open-source bioinformatic analysis tools has brought about the era of genomic precision medicine. Countries such as the UK have already successfully integrated Whole Genome Sequencing (WGS) into routine clinical practice and the NHS Executive have secured Illumina WGS capacity for 0.5 M DNA samples for the 2020-2023 period.[60]

The first studies investigating the use of WGS in Transfusion Medicine reported that the technology could be used to better classify and type blood donors and patients, enabling us to provide better-matched blood to patients that require regular transfusion, and to those who are at risk of adverse antibody-related events.[66, 65] However, a deeper understanding of the genetics underpinning blood group antigens that genomic data allowed was accompanied by the lack of international standards on how to define the variation observed and how to link it to antigen expression.

In **chapter 3** of this thesis, we addressed these problems by conducting the first data-driven review of blood group gene reference transcripts. This work resulted in the establishment of fixed Locus Reference Genomic records for each of the RBC antigen encoding genes and has allowed the ISBT to link the antigen defining variants in their reference tables to dbSNP identifiers and therefore coordinates in the human reference genome. We demonstrated, by analysis of WGS data, how this international standard for interpretation of blood group antigens from genetic data can be used to develop automated analysis tools such as bloodTyper and to extract clinically relevant information from the genome sequencing data of patients with rare diseases and cancer.

Although this represents an important step in introducing genomics-based matching of blood, our analysis of genetic variation in the *KEL* locus of just 13,037 individuals resulted in the discovery of 50-fold more unique haplotype sequences ($n=1,697$) than can be produced by considering only the variation recorded in the ISBT reference tables ($n=35$). In order to ensure the safe integration of genomics data into clinical practice, the transfusion medicine community must take the next step and use high coverage WGS datasets such as those

made available by the NIHR BioResource and the 100KGP to produce the fully phased haplotype reference maps needed to improve the accuracy of genetically inferred antigen types. Through continued international collaboration, we must ensure that this haplotype reference resource covers all ethnicities.

We showed that WGS can produce highly accurate RBC and HPA antigen typing results for patients, with 10,070 (99.5%) concordant results in 10,115 antigen typing results comparisons; however, WGS is currently too expensive to be applied to vast numbers of blood donors. This leaves us in a similar situation as other tests; we have excellent typing data for patients but have only limited typing data on the donors required to support them.

We address this dilemma in **chapter 4** by reporting the development of the UKBBv2 genotyping array for donor genotyping. This affordable assay is capable of identifying almost all clinically relevant RBC blood group antigens and also generated typing for HPA and HLA antigens. We validated the test by genotyping 7,473 donors and observed 99.92% concordance between clinical and array antigen typing results in 103,326 comparisons across 44 clinically relevant RBC, HLA and HPA antigens. Using the 1.2 million antigen genotyping results produced we were able to achieve a 2.6-fold increase in the number of matches identified in the same donors to support 3,146 historical complex patient cases with multiple RBC alloantibodies. Demonstrating the immediate clinical value of the UKBBv2 array, the data produced in this pre-clinical validation study was used to identify five compatible donors for a Dutch patient with bone marrow failure requiring regular transfusions for whom no compatible donors could be identified in the Netherlands.

To continue improving the design of the UKBBv2 donor genotyping array and accompanying bloodTyper interpretation software, it is essential to understand the cause of discordances between genotype inferred- and antibody determined antigen types. To this end, we developed the BGC capture assay, an affordable targeted NGS sequencing test that can be deployed by blood supply organisations as a second-line test for investigation of discordances observed between UKBBv2 array inferred and clinically recorded antigen types. In **chapter 5** we used this assay to resolve 78% (57 / 73) of discordances between UKBBv2 array and donor record antigen types reported in chapter 4. In the remaining cases, we were able to identify six previously unobserved DNA variants, which are likely to underpin altered antigen phenotypes in 6 of the 16 discordant cases. For 10 discordances we were not able to use the BGC sequencing assay due to lack of DNA for further analysis.

The fact that only 0.7% of donors used for validation in this study were of non-European ancestry is a limiting factor of this work. Due to the low frequency of RH variant antigens in Europeans and the high frequency in Sickle Cell patients of African ancestry, further validation must be performed before this test can be used to clinically type these patients

and the donors required to support them for variant antigens. Similarly, patients with other haemoglobinopathies such as Thalassemia are frequently of Southern Mediterranean, Persian, Arabian, and South-East Asian ancestry - populations in which blood group antigen frequencies are also significantly different to those observed European ancestry individuals. It is these patients who stand to benefit most from the extended matching that the technology reported here enables, and it is, therefore, crucial to generate validation data on a large number of individuals from these populations. The complete concordance between 835 antigen typing results for the 57 non-European individuals and identification of two erroneous serological typing results due to variant RhD antigen expression reported here, while not adequate for validation purposes, provides support for further studies using samples from donors and patients recruited by ancestry and specific antigen type.

In summary, the international standards, analysis software, donor genotyping assays, and validation data presented in this thesis provides blood supply organisations with the opportunity of implementing a policy of genomics-based precision transfusion medicine. Variants of the UKBBv2 assay presented here will be used by the FinnGen Biobank, Million Veteran Program, Taiwan Precision Medicine Initiative, and the impending UK 5 Million studies. This means that in the near future full blood cell antigen types will become available for millions of patients and healthy individuals in the population. The transfusion medicine community and global blood supply organisations should now take advantage of the developments in clinical genomics by investing in the regulatory and computational infrastructure required to make use of the dense antigen typing data which will become freely available. The BGC collaboration of seven national blood supply organisations and the New York Blood Centre will provide an exemplar of how genome-informed precision matching of blood and platelets can be delivered at scale by obtaining regulatory approval for the UKBBv2 array typing test.

6.1 Translating this work

The immediate next step is to assess how accurately the UKBBv2 array can type the blood group antigens of individuals of non-European ancestry. At the time of writing, the BGC has already assembled a panel of 3,137 DNA samples from Dutch and American donors of African ancestry and had these samples genotyped on the UKBBv2 array. Preliminary comparisons have shown 99.73% concordance in 68,193 antigen typing comparisons, all 183 discordances concern antigens within the RH and MNS blood group systems which are encoded by previously unobserved structural variation of the genome. On the basis of these early results, the blood typing array content has been redesigned to incorporate probes for accurate detection of structural variation in the relevant genes. In-depth analysis of this

data, including serological confirmation of variant RH antigens, will be performed over the coming months.

With the aim of integrating array genotyping technology into the global blood supply chain, the BGC is now actively engaged in the process of seeking FDA regulatory approval of a 50,000 variant, 384 sample format donor typing only Axiom array. In order to achieve this on a global scale, the BGC has expanded its membership to now include the national blood services of Australia, Canada, Finland, New Zealand, and South Africa and has planned a two-phase accreditation study. In this study STRIDES NIHR BioResource samples from 60,000 NHSBT donors will be genotyped at NHSBT Colindale using Thermo Fisher's GENETITAN-MC instrument. In addition, samples from another 10,000 donors made available by the other BGC members will be aggregated at NHSBT Cambridge. These 10,000 samples together with 4,000 STRIDES samples will be genotyped at the clinical laboratories of NHSBT, NYBC and Sanquin. Sample collection, DNA extraction and genotyping will be performed in a clinically accredited setting. The data on 70,000 blood donors will be analysed in Cambridge.

In Phase 1, the first 7,000 donor samples will be typed in a 'pre-clinical' validation exercise with the aim of further validate the genotyping platform itself alongside the processes which have been put in place to bring genotyping into the blood service laboratories. The results of this phase will also allow the BGC to take a final position for which antigens approval will be sought from regulators for 'diagnostic test status' vs 'screening test status'.

In Phase 2, the remaining 7,000, donor samples will be typed and the data will be used for obtaining regulatory approval from the FDA and the European regulatory agencies. Sample processing, genotyping, and data analysis will be performed in a 'hands off' and regulated manner - with antigen typing done 'on-board' the GENETITAN instruments. Typing results, antigen concordance data, and validation documentation will be presented to the regulatory authorities.

In 2021 Thermo Fisher will also place a GENETITAN within the Australian Red Cross Blood Service. It is expected that the four blood services with GENETITAN experience may become 'first clients' if the array becomes accredited. In the UK the selection of the array to be used for genotyping the 5M participants in the EDDRUK study will depend on the outcome of a European tendering process.

References

- [1] Thomas F. Baskett. James Blundell: the first transfusion of human blood., 2002. ISSN 03009572.
- [2] K. Landsteiner. Ueber Agglutinationserscheinungen normalen menschlichen Blutes. 1901. *Wiener Klinische Wochenschrift*, 2001. ISSN 00435325.
- [3] K. Landsteiner and Philip Levine. A New Agglutinable Factor Differentiating Individual Human Bloods. *Proceedings of the Society for Experimental Biology and Medicine*, 1927. ISSN 15353699. doi: 10.3181/00379727-24-3483.
- [4] Karl Landsteiner and Alexander S. Wiener. An Agglutinable Factor in Human Blood Recognized by Immune Sera for Rhesus Blood. *Proceedings of the Society for Experimental Biology and Medicine*, 1940. ISSN 15353699. doi: 10.3181/00379727-43-11151.
- [5] R. R.A. COOMBS, A. E. MOURANT, and R. R. RACE. A new test for the detection of weak and incomplete Rh agglutinins. *British journal of experimental pathology*, 1945. ISSN 00071021.
- [6] Geoff Daniels. *Human Blood Groups: 3rd edition*. 2013. ISBN 9781444333244. doi: 10.1002/9781118493595.
- [7] Blood Group AlleleTerminology ISBT. Red Cell Immunogenetics and Blood Group Terminology, 2019. URL <https://www.isbtweb.org/working-parties/red-cell-immunogenetics-and-blood-group-terminology/>.
- [8] G. Köhler and C. Milstein. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature*, 1975. ISSN 00280836. doi: 10.1038/256495a0.
- [9] S. KOSKIMIES. Human Lymphoblastoid Cell Line Producing Specific Antibody against Rh-Antigen D. *Scandinavian Journal of Immunology*, 1980. ISSN 13653083. doi: 10.1111/j.1365-3083.1980.tb00210.x.
- [10] A. W. BOYLSTON, B. GARDNER, R. L. ANDERSON, and N. C. HUGHES-JONES. Production of Human IgM Anti-D in Tissue Culture by EB-Virus-transformed Lymphocytes. *Scandinavian Journal of Immunology*, 1980. ISSN 13653083. doi: 10.1111/j.1365-3083.1980.tb00077.x.
- [11] Mark D. Melamed, John Gordon, Steven J. Ley, David Edgar, and Nevin C. Hughes-Jones. Senescence of a human lymphoblastoid clone producing anti-Rhesus(D). *European Journal of Immunology*, 1985. ISSN 15214141. doi: 10.1002/eji.1830150720.

- [12] K.M. Thompson and N.C. Hughes-Jones. Production and characterization of monoclonal anti-Rh. *Baillière's Clinical Haematology*, 3(2):243–253, 4 1990. ISSN 09503536. doi: 10.1016/S0950-3536(05)80049-6. URL <https://linkinghub.elsevier.com/retrieve/pii/S0950353605800496>.
- [13] Jill R. Storry and Martin L. Olsson. Genetic basis of blood group diversity, 2004. ISSN 00071048.
- [14] Geoff Daniels. The molecular genetics of blood group polymorphism, 2009. ISSN 03406717.
- [15] W. T. MORGAN. A contribution to human biochemical genetics; the chemical basis of blood-group specificity. *Proceedings of the Royal Society of London. Series B, Containing papers of a Biological character. Royal Society (Great Britain)*, 1960. ISSN 00804649. doi: 10.1098/rspb.1960.0002.
- [16] W. M. Watkins. Biochemistry and Genetics of the ABO, Lewis, and P blood group systems., 1980. ISSN 0065275X.
- [17] Marion E. Reid, Christine Lomas-Francis, and Martin L. Olsson. *The Blood Group Antigen*. 2012. ISBN 9780124158498. doi: 10.1016/C2011-0-69689-9.
- [18] Soohee Lee. Molecular basis of Kell blood group phenotypes, 1997. ISSN 00429007.
- [19] Soohee Lee, David C.W. Russo, Alexander P. Reiner, Jeffrey H. Lee, Michael Y. Sy, Marilyn J. Telen, W. John Judd, Philippe Simon, Maria J. Rodrigues, Teresa Chabert, Joyce Poole, Snezana Jovanovic-Srzentic, Cyril Levene, Vered Yahalom, and Colvin M. Redman. Molecular Defects Underlying the Kell Null Phenotype. *Journal of Biological Chemistry*, 2001. ISSN 00219258. doi: 10.1074/jbc.M103433200.
- [20] Lung Chih Yu, Yuh Ching Twu, Ching Yi Chang, and Marie Lin. Molecular basis of the Kell-null phenotype: A mutation at the splice site of human KEL gene abolishes the expression of Kell blood group antigens. *Journal of Biological Chemistry*, 2001. ISSN 00219258. doi: 10.1074/jbc.M009879200.
- [21] William J. Lane, Connie M. Westhoff, Nicholas S. Gleadall, Maria Aguad, Robin Smeland-Wagman, Sunitha Vege, Daimon P. Simmons, Helen H. Mah, Matthew S. Lebo, Klaudia Walter, Nicole Soranzo, Emanuele Di Angelantonio, John Danesh, David J. Roberts, Nick A. Watkins, Willem H. Ouwehand, Adam S. Butterworth, Richard M. Kaufman, Heidi L. Rehm, Leslie E. Silberstein, Robert C. Green, David W. Bates, Carrie Blout, Kurt D. Christensen, Allison L. Cirino, Carolyn Y. Ho, Joel B. Krier, Lisa S. Lehmann, Calum A. MacRae, Cynthia C. Morton, Denise L. Perry, Christine E. Seidman, Shamil R. Sunyaev, Jason L. Vassy, Erica Schonman, Tiffany Nguyen, Eleanor Steffens, Wendi Nicole Betting, Samuel J. Aronson, Ozge Ceyhan-Birsoy, Kalotina Machini, Heather M. McLaughlin, Danielle R. Azzariti, Ellen A. Tsai, Jennifer Blumenthal-Barby, Lindsay Z. Feuerman, Amy L. McGuire, Kaitlyn Lee, Jill O. Robinson, Melody J. Slashinski, Pamela M. Diamond, Kelly Davis, Peter A. Ubel, Peter Kraft, J. Scott Roberts, Judy E. Garber, Tina Hambuch, Michael F. Murray, Isaac Kohane, and Sek Won Kong. Automated typing of red blood cell and platelet antigens: a whole-genome sequencing study. *The Lancet Haematology*, 5(6): e241–e251, 2018. ISSN 23523026. doi: 10.1016/S2352-3026(18)30053-X.

- [22] B. Veldhuisen, C. E. Van Der Schoot, and M. De Haas. Blood group genotyping: From patient to high-throughput donor screening. *Vox Sanguinis*, 97(3):198–206, 2009. ISSN 00429007. doi: 10.1111/j.1423-0410.2009.01209.x.
- [23] Shoko Nishihara, Hisashi Narimatsu, Hiroko Iwasaki, Shin Yazawa, Suguru Akamatsu, Takao Ando, Taiko Seno, and Ikuyo Narimatsu. Molecular genetic analysis of the human Lewis histo-blood group system. *Journal of Biological Chemistry*, 1994. ISSN 00219258.
- [24] Yasuo Fukumori, Shiro Ohnoki, Hirotohi Shibata, Hideo Yamaguchi, and Hiroaki Nishimukai. Genotyping of ABO blood groups by PCR and RFLP analysis of 5 nucleotide positions. *International Journal Of Legal Medicine*, 1995. ISSN 09379827. doi: 10.1007/BF01428401.
- [25] R. W.A.M. Kuijpers, N. M. Faber, H. T.M. Cuypers, W. H. Ouwehand, and A. E.G.K. Von dem Borne. NH₂-terminal globular domain of human platelet glycoprotein Ib α has a methionine145/threonine145 amino acid polymorphism, which is associated with the HPA-2 (Ko) alloantigens. *Journal of Clinical Investigation*, 1992. ISSN 00219738. doi: 10.1172/JCI115596.
- [26] Mark Poulter, Tim J. Kemp, and Ben Carritt. DNA-based rhesus typing: Simultaneous determination of RHC and RHD status using the polymerase chain reaction. *Vox Sanguinis*, 1996. ISSN 00429007. doi: 10.1111/j.1423-0410.1996.tb01316.x.
- [27] P. A. Maaskant-Van Wijk, B. H.W. Faas, J. A.M. De Ruijter, M. A.M. Overbeeke, A. E.G.Kr Von Dem Borne, D. J. Van Rhenen, and C. E. Van Der Schoot. Genotyping of RHD by multiplex polymerase chain reaction analysis of six RHD-specific exons. *Transfusion*, 1998. ISSN 00411132. doi: 10.1046/j.1537-2995.1998.38111299056309.x.
- [28] Franz F. Wagner, Rita Bittner, Eduard K. Petershofen, Andrea Doescher, and Thomas H. Müller. Cost-efficient sequence-specific priming-polymerase chain reaction screening for blood donors with rare phenotypes. *Transfusion*, 2008. ISSN 00411132. doi: 10.1111/j.1537-2995.2008.01682.x.
- [29] T. Ficko, V. Galvani, R. Ruprecht, T. Dovc, and Primož Rožman. Real-time PCR genotyping of human platelet alloantigens HPA-1, HPA-2, HPA-3 and HPA-5 is superior to the standard PCR-SSP method. *Transfusion Medicine*, 2004. ISSN 09587578. doi: 10.1111/j.1365-3148.2004.00538.x.
- [30] Yan Yun Wu and Gyorgy Csako. Rapid and/or high-throughput genotyping for human red blood cell, platelet and leukocyte antigens, and forensic applications. *Clinica Chimica Acta*, 363(1-2):165–176, 1 2006. ISSN 00098981. doi: 10.1016/j.cccn.2005.07.010. URL <https://linkinghub.elsevier.com/retrieve/pii/S0009898105004353>.
- [31] Martine G.H.M. Grootkerk-Tax, Aicha Ait Soussan, Masja De Haas, Petra A. Maaskant-van Wijk, and C. Ellen Van Der Schoot. Evaluation of prenatal RHD typing strategies on cell-free fetal DNA from maternal plasma. *Transfusion*, 2006. ISSN 00411132. doi: 10.1111/j.1537-2995.2006.01044.x.

- [32] H. Polin, M. Danzer, J. Pröll, K. Hofer, U. Heilinger, A. Zopf, and C. Gabriel. Introduction of a real-time based blood group genotyping approach. *Vox Sanguinis*, 2008. ISSN 00429007. doi: 10.1111/j.1423-0410.2008.01067.x.
- [33] Henk S.P. Garritsen, Alex Xiu-Cheng Fan, Nicole Bosse, Horst Hannig, Reinhard Kelsch, Hartmut Kroll, Wolfgang Holzgreve, and Xiao Yan Zhong. Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry for genotyping of human platelet-specific antigens. *Transfusion*, 49(2):252–258, 2 2009. ISSN 00411132. doi: 10.1111/j.1537-2995.2008.01953.x. URL <http://doi.wiley.com/10.1111/j.1537-2995.2008.01953.x>.
- [34] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009. ISSN 13674803. doi: 10.1093/bioinformatics/btp352.
- [35] Ghazala Hashmi, Tasmia Shariff, Michael Seul, Prabhakar Vissavajhala, Kim Hue-Roye, Dalisay Charles-Pierre, Christine Lomas-Francis, Asok Chaudhuri, and Marion E. Reid. A flexible array format for large-scale, rapid blood group DNA typing. *Transfusion*, 2005. ISSN 00411132. doi: 10.1111/j.1537-2995.2005.04362.x.
- [36] Sigrid H.W. Beiboer, Tinka Wieringa-Jelsma, Petra A. Maaskant-Van Wijk, C. Ellen Van Der Schoot, Rob Van Zwieten, Dirk Roos, Johan T. Den Dunnen, and Masja De Haas. Rapid genotyping of blood group antigens by multiplex polymerase chain reaction and DNA microarray hybridization. *Transfusion*, 45(5):667–679, 2005. ISSN 00411132. doi: 10.1111/j.1537-2995.2005.04319.x.
- [37] Lopez Martinez M., Chinnapapagari S.K.R., Olsson M.L., Nogues N., Scott M.L., 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 J., Azkarate M., Vesga M.A., Camacho E., Cheroutre G., Link A., Jinoch P., Svobodova I., Martinez A., De Haas M., Flegel W.A., and Avent N.D. Mass-scale extensive genotyping of 3000 RBC samples by bloodchip v 1.0, 2009.
- [38] Willy A. Flegel, Jerome L. Gottschall, and Gregory A. Denomme. Implementing mass-scale red cell genotyping at a blood center. *Transfusion*, 55(11):2610–2615, 2015. ISSN 15372995. doi: 10.1111/trf.13168.
- [39] Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William Fitzhugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie Levine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P. Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd,

Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C. Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Robert H. Waterston, Richard K. Wilson, Ladeana W. Hillier, John D. McPherson, Marco A. Marra, Elaine R. Mardis, Lucinda A. Fulton, Asif T. Chinwalla, Kimberlie H. Pepin, Warren R. Gish, Stephanie L. Chissoe, Michael C. Wendl, Kim D. Delehaunty, Tracie L. Miner, Andrew Delehaunty, Jason B. Kramer, Lisa L. Cook, Robert S. Fulton, Douglas L. Johnson, Patrick J. Minx, Sandra W. Clifton, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward Uberbacher, Marvin Frazier, Richard A. Gibbs, Donna M. Muzny, Steven E. Scherer, John B. Bouck, Erica J. Sodergren, Kim C. Worley, Catherine M. Rives, James H. Gorrell, Michael L. Metzker, Susan L. Naylor, Raju S. Kucherlapati, David L. Nelson, George M. Weinstock, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, Jean Weissenbach, Roland Heilig, William Saurin, Francois Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Catherine Robert, Patrick Wincker, André Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, Douglas R. Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Mei Lee Hong, Joann Dubois, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Ronald W. Davis, Nancy A. Federspiel, A. Pia Abola, Michael J. Proctor, Bruce A. Roe, Feng Chen, Huaqin Pan, Juliane Ramser, Hans Lehrach, Richard Reinhardt, W. Richard McCombie, Melissa De La Bastide, Neilay Dedhia, Helmut Blöcker, Klaus Hornischer, Gabriele Nordsiek, Richa Agarwala, L. Aravind, Jeffrey A. Bailey, Alex Bateman, Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G. Brown, Christopher B. Burge, Lorenzo Cerutti, Hsiu Chuan Chen, Deanna Church, Michele Clamp, Richard R. Copley, Tobias Doerks, Sean R. Eddy, Evan E. Eichler, Terrence S. Furey, James Galagan, James G.R. Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L. Steven Johnson, Thomas A. Jones, Simon Kasif, Arek Kasprzyk, Scot Kennedy, W. James Kent, Paul Kitts, Eugene V. Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M. Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V. Moran, Nicola Mulder, Victor J. Pollara, Chris P. Ponting, Greg Schuler, Jörg Schultz, Guy Slater, Arian F.A. Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I. Wolf, Kenneth H. Wolfe, Shiaw Pyng Yang, Ru Fang Yeh, Francis Collins, Mark S. Guyer, Jane Peterson, Adam Felsenfeld, Kris A. Wetterstrand, Richard M. Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R. Cox, Maynard V. Olson, Rajinder Kaul, Christopher Raymond, Nobuyoshi Shimizu, Kazuhiko Kawasaki, Shinsei Minoshima, Glen A. Evans, Maria Athanasiou, Roger Schultz, Aristides Patrinos, and Michael J. Morgan. Initial sequencing and analysis of the human genome. *Nature*, 2001. ISSN 00280836. doi: 10.1038/35057062.

- [40] John W. Belmont, Paul Hardenbol, Thomas D. Willis, Fuli Yu, Huanming Yang, Lan Yang Ch'Ang, Wei Huang, Bin Liu, Yan Shen, Paul Kwong Hang Tam, Lap Chee Tsui, Mary Miu Yee Waye, Jeffrey Tze Fei Wong, Changqing Zeng, Qingrun Zhang, Mark S. Chee, Luana M. Galver, Semyon Kruglyak, Sarah S. Murray, Arnold R. Oliphant, Alexandre Montpetit, Fanny Chagnon, Vincent Ferretti, Martin Leboeuf, Michael S. Phillips, Andrei Verner, Shenghui Duan, Denise L. Lind, Raymond D.

- Miller, John Rice, Nancy L. Saccone, Patricia Taillon-Miller, Ming Xiao, Akihiro Sekine, Koki Sorimachi, Yoichi Tanaka, Tatsuhiko Tsunoda, Eiji Yoshino, David R. Bentley, Sarah Hunt, Don Powell, Houcan Zhang, Ichiro Matsuda, Yoshimitsu Fukushima, Darryl R. Macer, Eiko Suda, Charles Rotimi, Clement A. Adebamowo, Toyin Aniagwu, Patricia A. Marshall, Olayemi Matthew, Chibuzor Nkwodimmah, Charmaine D.M. Royal, Mark F. Leppert, Missy Dixon, Fiona Cunningham, Ardavan Kanani, Gudmundur A. Thorisson, Peter E. Chen, David J. Cutler, Carl S. Kashuk, Peter Donnelly, Jonathan Marchini, Gilean A.T. McVean, Simon R. Myers, Lon R. Cardon, Andrew Morris, Bruce S. Weir, James C. Mullikin, Michael Feolo, Mark J. Daly, Renzong Qiu, Alastair Kent, Georgia M. Dunston, Kazuto Kato, Norio Niikawa, Jessica Watkin, Richard A. Gibbs, Erica Sodergren, George M. Weinstock, Richard K. Wilson, Lucinda L. Fulton, Jane Rogers, Bruce W. Birren, Hua Han, Hongguang Wang, Martin Godbout, John C. Wallenburg, Paul L'Archevêque, Guy Bellemare, Kazuo Todani, Takashi Fujita, Satoshi Tanaka, Arthur L. Holden, Francis S. Collins, Lisa D. Brooks, Jean E. McEwen, Mark S. Guyer, Elke Jordan, Jane L. Peterson, Jack Spiegel, Lawrence M. Sung, Lynn F. Zacharia, Karen Kennedy, Michael G. Dunn, Richard Seabrook, Mark Shillito, Barbara Skene, John G. Stewart, David L. Valle, Ellen Wright Clayton, Lynn B. Jorde, Aravinda Chakravarti, Mildred K. Cho, Troy Duster, Morris W. Foster, Maria Jasperse, Bartha M. Knoppers, Pui Yan Kwok, Julio Licinio, Jeffrey C. Long, Pilar Ossorio, Vivian Ota Wang, Charles N. Rotimi, Patricia Spallone, Sharon F. Terry, Eric S. Lander, Eric H. Lai, Deborah A. Nickerson, Gonçalo R. Abecasis, David Altshuler, Michael Boehnke, Panos Deloukas, Julie A. Douglas, Stacey B. Gabriel, Richard R. Hudson, Thomas J. Hudson, Leonid Kruglyak, Yusuke Nakamura, Robert L. Nussbaum, Stephen F. Schaffner, Stephen T. Sherry, Lincoln D. Stein, and Toshihiro Tanaka. The international HapMap project. *Nature*, 426(6968):789–796, 2003. ISSN 00280836. doi: 10.1038/nature02168.
- [41] Paul R. Burton, David G. Clayton, Lon R. Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P. Kwiatkowski, Mark I. McCarthy, Willem H. Ouwehand, Nilesh J. Samani, John A. Todd, Peter Donnelly, Jeffrey C. Barrett, Dan Davison, Doug Easton, David Evans, Hin Tak Leung, Jonathan L. Marchini, Andrew P. Morris, Chris C.A. Spencer, Martin D. Tobin, Antony P. Attwood, James P. Boorman, Barbara Cant, Ursula Everson, Judith M. Hussey, Jennifer D. Jolley, Alexandra S. Knight, Kerstin Koch, Elizabeth Meech, Sarah Nutland, Christopher V. Prowse, Helen E. Stevens, Niall C. Taylor, Graham R. Walters, Neil M. Walker, Nicholas A. Watkins, Thilo Winzer, Richard W. Jones, Wendy L. McArdle, Susan M. Ring, David P. Strachan, Marcus Pembrey, Gerome Breen, David St. Clair, Sian Caesar, Katherine Gordon-Smith, Lisa Jones, Christine Fraser, Elaine K. Green, Detelina Grozeva, Marian L. Hamshere, Peter A. Holmans, Ian R. Jones, George Kirov, Valentina Moskvina, Ivan Nikolov, Michael C. O'Donovan, Michael J. Owen, David A. Collier, Amanda Elkin, Anne Farmer, Richard Williamson, Peter McGuffin, Allan H. Young, I. Nicol Ferrier, Stephen G. Ball, Anthony J. Balmforth, Jennifer H. Barrett, D. Timothy Bishop, Mark M. Iles, Azhar Maqbool, Nadira Yuldasheva, Alistair S. Hall, Peter S. Braund, Richard J. Dixon, Massimo Mangino, Suzanne Stevens, John R. Thompson, Francesca Bredin, Mark Tremelling, Miles Parkes, Hazel Drummond, Charles W. Lees, Elaine R. Nimmo, Jack Satsangi, Sheila A. Fisher, Alastair Forbes, Cathryn M. Lewis, Clive M. Onnie, Natalie J. Prescott, Jeremy Sanderson, Christopher G. Mathew, Jamie Barbour, M. Khalid Mohiuddin, Catherine E. Todhunter, John C. Mansfield, Tariq Ahmad, Fraser R. Cummings, Derek P. Jewell, John Webster, Morris J. Brown, G. Mark Lath-

- rop, John Connell, Anna Dominiczak, Carolina A. Braga Marcano, Beverley Burke, Richard Dobson, Johannie Gungadoo, Kate L. Lee, Patricia B. Munroe, Stephen J. Newhouse, Abiodun Onipinla, Chris Wallace, Mingzhan Xue, Mark Caulfield, Martin Farrall, Anne Barton, Ian N. Bruce, Hannah Donovan, Steve Eyre, Paul D. Gilbert, Samantha L. Hider, Anne M. Hinks, Sally L. John, Catherine Potter, Alan J. Silman, Deborah P.M. Symmons, Wendy Thomson, Jane Worthington, David B. Dunger, Barry Widmer, Timothy M. Frayling, Rachel M. Freathy, Hana Lango, John R.B. Perry, Beverley M. Shields, Michael N. Weedon, Andrew T. Hattersley, Graham A. Hitman, Mark Walker, Kate S. Elliott, Christopher J. Groves, Cecilia M. Lindgren, Nigel W. Rayner, Nicholas J. Timpson, Eleftheria Zeggini, Melanie Newport, Giorgio Sirugo, Emily Lyons, Fredrik Vannberg, Adrian V.S. Hill, Linda A. Bradbury, Claire Farrar, Jennifer J. Pointon, Paul Wordsworth, Matthew A. Brown, Jayne A. Franklyn, Joanne M. Heward, Matthew J. Simmonds, Stephen C.L. Gough, Sheila Seal, Michael R. Stratton, Nazneen Rahman, Sclerosis Maria Ban, An Goris, Stephen J. Sawcer, Alastair Compston, David Conway, Muminatou Jallow, Kirk A. Rockett, Suzannah J. Bumpstead, Amy Chaney, Kate Downes, Mohammed J.R. Ghorri, Rhian Gwilliam, Sarah E. Hunt, Michael Inouye, Andrew Keniry, Emma King, Ralph McGinnis, Simon Potter, Rathi Ravindrarajah, Pamela Whittaker, Claire Widden, David Withers, Niall J. Cardin, Teresa Ferreira, Joanne Pereira-Gale, Ingileif B. Hallgrimsdóttir, Bryan N. Howie, Chris C.A. Spencer, Zhan Su, Yik Ying Teo, Damjan Vukcevic, David Bentley, and Alistair Compston. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 2007. ISSN 14764687. doi: 10.1038/nature05911.
- [42] Nick Craddock, Matthew E. Hurles, Niall Cardin, Richard D. Pearson, Vincent Plagnol, Samuel Robson, Damjan Vukcevic, Chris Barnes, Donald F. Conrad, Eleni Giannoulitou, Chris Holmes, Jonathan L. Marchini, Kathy Stirrups, Martin D. Tobin, Louise V. Wain, Chris Yau, Jan Aerts, Tariq Ahmad, T. Daniel Andrews, Hazel Arbury, Anthony Attwood, Adam Auton, Stephen G. Ball, Anthony J. Balmforth, Jeffrey C. Barrett, Inês Barroso, Anne Barton, Amanda J. Bennett, Sanjeev Bhaskar, Katarzyna Blaszczyk, John Bowes, Oliver J. Brand, Peter S. Braund, Francesca Bredin, Gerome Breen, Morris J. Brown, Ian N. Bruce, Jaswinder Bull, Oliver S. Burren, John Burton, Jake Byrnes, Sian Caesar, Chris M. Clee, Alison J. Coffey, John M.C. Connell, Jason D. Cooper, Anna F. Dominiczak, Kate Downes, Hazel E. Drummond, Darshna Dudakia, Andrew Dunham, Bernadette Ebbs, Diana Eccles, Sarah Edkins, Cathryn Edwards, Anna Elliot, Paul Emery, David M. Evans, Gareth Evans, Steve Eyre, Anne Farmer, I. Nicol Ferrier, Lars Feuk, Tomas Fitzgerald, Edward Flynn, Alistair Forbes, Liz Forty, Jayne A. Franklyn, Rachel M. Freathy, Polly Gibbs, Paul Gilbert, Omer Gokumen, Katherine Gordon-Smith, Emma Gray, Elaine Green, Chris J. Groves, Detelina Grozeva, Rhian Gwilliam, Anita Hall, Naomi Hammond, Matt Hardy, Pile Harrison, Neelam Hassanali, Husam Hebaishi, Sarah Hines, Anne Hinks, Graham A. Hitman, Lynne Hocking, Eleanor Howard, Philip Howard, Joanna M.M. Howson, Debbie Hughes, Sarah Hunt, John D. Isaacs, Mahim Jain, Derek P. Jewell, Toby Johnson, Jennifer D. Jolley, Ian R. Jones, Lisa A. Jones, George Kirov, Cordelia F. Langford, Hana Lango-Allen, G. Mark Lathrop, James Lee, Kate L. Lee, Charlie Lees, Kevin Lewis, Cecilia M. Lindgren, Meeta Maisuria-Armer, Julian Maller, John Mansfield, Paul Martin, Dunecan C.O. Massey, Wendy L. McArdle, Peter McGuffin, Kirsten E. McLay, Alex Mentzer, Michael L. Mimmack, Ann E. Morgan, Andrew P. Morris, Craig Mowat, Simon Myers, William Newman, Elaine R. Nimmo, Michael C. O'Donovan, Abiodun

- Onipinla, Ifejinelo Onyiah, Nigel R. Ovington, Michael J. Owen, Kimmo Palin, Kirstie Parnell, David Pernet, John R.B. Perry, Anne Phillips, Dalila Pinto, Natalie J. Prescott, Inga Prokopenko, Michael A. Quail, Suzanne Rafelt, Nigel W. Rayner, Richard Redon, David M. Reid, Anthony Renwick, Susan M. Ring, Neil Robertson, Ellie Russell, David St Clair, Jennifer G. Sambrook, Jeremy D. Sanderson, Helen Schuilenburg, Carol E. Scott, Richard Scott, Sheila Seal, Sue Shaw-Hawkins, Beverley M. Shields, Matthew J. Simmonds, Debbie J. Smyth, Elilan Somaskantharajah, Katarina Spanova, Sophia Steer, Jonathan Stephens, Helen E. Stevens, Millicent A. Stone, Zhan Su, Deborah P.M. Symmons, John R. Thompson, Wendy Thomson, Mary E. Travers, Clare Turnbull, Armand Valsesia, Mark Walker, Neil M. Walker, Chris Wallace, Margaret Warren-Perry, Nicholas A. Watkins, John Webster, Michael N. Weedon, Anthony G. Wilson, Matthew Woodburn, B. Paul Wordsworth, Allan H. Young, Eleftheria Zeggini, Nigel P. Carter, Timothy M. Frayling, Charles Lee, Gil McVean, Patricia B. Munroe, Aarno Palotie, Stephen J. Sawcer, Stephen W. Scherer, David P. Strachan, Chris Tyler-Smith, Matthew A. Brown, Paul R. Burton, Mark J. Caulfield, Alastair Compston, Martin Farrall, Stephen C.L. Gough, Alistair S. Hall, Andrew T. Hattersley, Adrian V.S. Hill, Christopher G. Mathew, Marcus Pembrey, Jack Satsangi, Michael R. Stratton, Jane Worthington, Panos Deloukas, Audrey Duncanson, Dominic P. Kwiatkowski, Mark I. McCarthy, Willem H. Ouwehand, Miles Parkes, Nazneen Rahman, John A. Todd, Nilesh J. Samani, and Peter Donnelly. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 2010. ISSN 00280836. doi: 10.1038/nature08979.
- [43] Nicole Soranzo, Tim D. Spector, Massimo Mangino, Brigitte Kühnel, Augusto Rendon, Alexander Teumer, Christina Willenborg, Benjamin Wright, Li Chen, Mingyao Li, Perttu Salo, Benjamin F. Voight, Philippa Burns, Roman A. Laskowski, Yali Xue, Stephan Menzel, David Altshuler, John R. Bradley, Suzannah Bumpstead, Mary Susan Burnett, Joseph Devaney, Angela Döring, Roberto Elosua, Stephen E. Epstein, Wendy Erber, Mario Falchi, Stephen F. Garner, Mohammed J.R. Ghorri, Alison H. Goodall, Rhian Gwilliam, Hakon H. Hakonarson, Alistair S. Hall, Naomi Hammond, Christian Hengstenberg, Thomas Illig, Inke R. König, Christopher W. Knouff, Ruth McPherson, Olle Melander, Vincent Mooser, Matthias Nauck, Markku S. Nieminen, Christopher J. O'Donnell, Leena Peltonen, Simon C. Potter, Holger Prokisch, Daniel J. Rader, Catherine M. Rice, Robert Roberts, Veikko Salomaa, Jennifer Sambrook, Stefan Schreiber, Heribert Schunkert, Stephen M. Schwartz, Jovana Serbanovic-Canic, Juha Sinisalo, David S. Siscovick, Klaus Stark, Ida Surakka, Jonathan Stephens, John R. Thompson, Uwe Völker, Henry Völzke, Nicholas A. Watkins, George A. Wells, H. Erich Wichmann, David A. Van Heel, Chris Tyler-Smith, Swee Lay Thein, Sekar Kathiresan, Markus Perola, Muredach P. Reilly, Alexandre F.R. Stewart, Jeanette Erdmann, Nilesh J. Samani, Christa Meisinger, Andreas Greinacher, Panos Deloukas, Willem H. Ouwehand, and Christian Gieger. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nature Genetics*, 2009. ISSN 10614036. doi: 10.1038/ng.467.
- [44] Nicole Soranzo, Augusto Rendon, Christian Gieger, Chris I. Jones, Nicholas A. Watkins, Stephan Menzel, Angela Döring, Jonathan Stephens, Holger Prokisch, Wendy Erber, Simon C. Potter, Sarah L. Bray, Philippa Burns, Jennifer Jolley, Mario Falchi, Brigitte Kühnel, Jeanette Erdmann, Heribert Schunkert, Nilesh J. Samani, Thomas Illig, Stephen F. Garner, Angela Rankin, Christa Meisinger, John R. Bradley, Swee Lay

- Thein, Alison H. Goodall, Tim D. Spector, Panos Deloukas, and Willem H. Ouwehand. Anovel variant on chromosome 7q22.3 associated with mean platelet volume, counts, and function. *Blood*, 2009. ISSN 00064971. doi: 10.1182/blood-2008-10-184234.
- [45] Christian Gieger, Aparna Radhakrishnan, Ana Cvejic, Weihong Tang, Eleonora Porcu, Giorgio Pistis, Jovana Serbanovic-Canic, Ulrich Elling, Alison H. Goodall, Yann Labrune, Lorna M. Lopez, Reedik Mägi, Stuart Meacham, Yukinori Okada, Nicola Pirastu, Rossella Sorice, Alexander Teumer, Katrin Voss, Weihua Zhang, Ramiro Ramirez-Solis, Joshua C. Bis, David Ellinghaus, Martin Gögele, Jouke Jan Hottenga, Claudia Langenberg, Peter Kovacs, Paul F. O'reilly, So Youn Shin, Tõnu Esko, Jaana Hartiala, Stavroula Kanoni, Federico Murgia, Afshin Parsa, Jonathan Stephens, Pim Van Der Harst, C. Ellen Van Der Schoot, Hooman Allayee, Antony Attwood, Beverley Balkau, François Bastardot, Saonli Basu, Sebastian E. Baumeister, Ginevra Biino, Lorenzo Bomba, Amélie Bonnefond, François Cambien, John C. Chambers, Francesco Cucca, Pio Dadamo, Gail Davies, Rudolf A. De Boer, Eco J.C. De Geus, Angela Döring, Paul Elliott, Jeanette Erdmann, David M. Evans, Mario Falchi, Wei Feng, Aaron R. Folsom, Ian H. Frazer, Quince D. Gibson, Nicole L. Glazer, Chris Hammond, Anna Liisa Hartikainen, Susan R. Heckbert, Christian Hengstenberg, Micha Hersch, Thomas Illig, Ruth J.F. Loos, Jennifer Jolley, Kay Tee Khaw, Brigitte Kühnel, Marie Christine Kyrtsonis, Vasiliki Lagou, Heather Lloyd-Jones, Thomas Lumley, Massimo Mangino, Andrea Maschio, Irene Mateo Leach, Barbara Mcknight, Yasin Memari, Braxton D. Mitchell, Grant W. Montgomery, Yusuke Nakamura, Matthias Nauck, Gerjan Navis, Ute Nöthlings, Ilja M. Nolte, David J. Porteous, Anneli Pouta, Peter P. Pramstaller, Janne Pullat, Susan M. Ring, Jerome I. Rotter, Daniela Ruggiero, Aimo Ruukonen, Cinzia Sala, Nilesh J. Samani, Jennifer Sambrook, David Schlessinger, Stefan Schreiber, Heribert Schunkert, James Scott, Nicholas L. Smith, Harold Snieder, John M. Starr, Michael Stumvoll, Atsushi Takahashi, W. H. Wilson Tang, Kent Taylor, Albert Tenesa, Swee Lay Thein, Anke Tönjes, Manuela Uda, Sheila Ulivi, Dirk J. Van Veldhuisen, Peter M. Visscher, Uwe Völker, H. Erich Wichmann, Kerri L. Wiggins, Gonneke Willemsen, Tsun Po Yang, Jing Hua Zhao, Paavo Zitting, John R. Bradley, George V. Dedoussis, Paolo Gasparini, Stanley L. Hazen, Andres Metspalu, Mario Pirastu, Alan R. Shuldiner, L. Joost Van Pelt, Jaap Jan Zwaginga, Dorret I. Boomsma, Ian J. Deary, Andre Franke, Philippe Froguel, Santhi K. Ganesh, Marjo Riitta Jarvelin, Nicholas G. Martin, Christa Meisinger, Bruce M. Psaty, Timothy D. Spector, Nicholas J. Wareham, Jan Willem N. Akkerman, Marina Ciullo, Panos Deloukas, Andreas Greinacher, Steve Jupe, Naoyuki Kamatani, Jyoti Khadake, Jaspal S. Kooner, Josef Penninger, Inga Prokopenko, Derek Stemple, Daniela Toniolo, Lorenz Wernisch, Serena Sanna, Andrew A. Hicks, Augusto Rendon, Manuel A. Ferreira, Willem H. Ouwehand, and Nicole Soranzo. New gene functions in megakaryopoiesis and platelet formation. *Nature*, 480(7376):201–208, 2011. ISSN 00280836. doi: 10.1038/nature10659. URL <http://dx.doi.org/10.1038/nature10659>.
- [46] Pim Van Der Harst, Weihua Zhang, Irene Mateo Leach, Augusto Rendon, Niek Verweij, Joban Sehmi, Dirk S. Paul, Ulrich Elling, Hooman Allayee, Xinzhong Li, Aparna Radhakrishnan, Sian Tsung Tan, Katrin Voss, Christian X. Weichenberger, Cornelis A. Albers, Abtehale Al-Hussani, Folkert W. Asselbergs, Marina Ciullo, Fabrice Danjou, Christian Dina, Tõnu Esko, David M. Evans, Lude Franke, Martin Gögele, Jaana Hartiala, Micha Hersch, Hilma Holm, Jouke Jan Hottenga, Stavroula Kanoni, Marcus E. Kleber, Vasiliki Lagou, Claudia Langenberg, Lorna M. Lopez, Leo Pekka Lyytikäinen,

- Olle Melander, Federico Murgia, Ilja M. Nolte, Paul F. O'Reilly, Sandosh Padmanabhan, Afshin Parsa, Nicola Pirastu, Eleonora Porcu, Laura Portas, Inga Prokopenko, Janina S. Ried, So Youn Shin, Clara S. Tang, Alexander Teumer, Michela Traglia, Sheila Ulivi, Harm Jan Westra, Jian Yang, Jing Hua Zhao, Franco Anni, Abdel Abdellaoui, Antony Attwood, Beverley Balkau, Stefania Bandinelli, François Bastardot, Beben Benyamin, Bernhard O. Boehm, William O. Cookson, Debashish Das, Paul I.W. De Bakker, Rudolf A. De Boer, Eco J.C. De Geus, Marleen H. De Moor, Maria Dimitriou, Francisco S. Domingues, Angela Döring, Gunnar Engström, Gudmundur Ingi Eyjolfsson, Luigi Ferrucci, Krista Fischer, Renzo Galanello, Stephen F. Garner, Bernd Genser, Quince D. Gibson, Giorgia Grotto, Daniel Fannar Gudbjartsson, Sarah E. Harris, Anna Liisa Hartikainen, Claire E. Hastie, Bo Hedblad, Thomas Illig, Jennifer Jolley, Mika Kähönen, Ido P. Kema, John P. Kemp, Liming Liang, Heather Lloyd-Jones, Ruth J.F. Loos, Stuart Meacham, Sarah E. Medland, Christa Meisinger, Yasin Memari, Evelin Mihailov, Kathy Miller, Miriam F. Moffatt, Matthias Nauck, Maria Novatchkova, Teresa Nutile, Isleifur Olafsson, Pall T. Onundarson, Debora Parracciani, Brenda W. Penninx, Lucia Perseu, Antonio Piga, Giorgio Pistis, Anneli Pouta, Ursula Puc, Olli Raitakari, Susan M. Ring, Antonietta Robino, Daniela Ruggiero, Aimo Ruokonen, Aude Saint-Pierre, Cinzia Sala, Andres Salumets, Jennifer Sambrook, Hein Schepers, Carsten Oliver Schmidt, Herman H.W. Silljé, Rob Sladek, Johannes H. Smit, John M. Starr, Jonathan Stephens, Patrick Sulem, Toshiko Tanaka, Unnur Thorsteinsdottir, Vinicius Tragante, Wiek H. Van Gilst, L. Joost Van Pelt, Dirk J. Van Veldhuisen, Uwe Völker, John B. Whitfield, Gonneke Willemsen, Bernhard R. Winkelmann, Gerald Wirnsberger, Ale Algra, Francesco Cucca, Adamo Pio D'Adamo, John Danesh, Ian J. Deary, Anna F. Dominiczak, Paul Elliott, Paolo Fortina, Philippe Froguel, Paolo Gasparini, Andreas Greinacher, Stanley L. Hazen, Marjo Riitta Jarvelin, Kay Tee Khaw, Terho Lehtimäki, Winfried Maerz, Nicholas G. Martin, Andres Metspalu, Braxton D. Mitchell, Grant W. Montgomery, Carmel Moore, Gerjan Navis, Mario Pirastu, Peter P. Pramstaller, Ramiro Ramirez-Solis, Eric Schadt, James Scott, Alan R. Shuldiner, George Davey Smith, J. Gustav Smith, Harold Snieder, Rossella Sorice, Tim D. Spector, Kari Stefansson, Michael Stumvoll, W. H. Wilson Tang, Daniela Toniolo, Anke Tönjes, Peter M. Visscher, Peter Vollenweider, Nicholas J. Wareham, Bruce H.R. Wolffenbuttel, Dorret I. Boomsma, Jacques S. Beckmann, George V. Dedoussis, Panos Deloukas, Manuel A. Ferreira, Serena Sanna, Manuela Uda, Andrew A. Hicks, Josef Martin Penninger, Christian Gieger, Jaspal S. Kooner, Willem H. Ouwehand, Nicole Soranzo, and John C. Chambers. Seventy-five genetic loci influencing the human red blood cell. *Nature*, 2012. ISSN 00280836. doi: 10.1038/nature11677.
- [47] William J. Astle, Heather Elding, Tao Jiang, Dave Allen, Dace Ruklisa, Alice L. Mann, Daniel Mead, Heleen Bouman, Fernando Riveros-Mckay, Myrto A. Kostadima, John J. Lambourne, Suthesh Sivapalaratnam, Kate Downes, Kousik Kundu, Lorenzo Bomba, Kim Berentsen, John R. Bradley, Louise C. Daugherty, Olivier Delaneau, Kathleen Freson, Stephen F. Garner, Luigi Grassi, Jose Guerrero, Matthias Haimel, Eva M. Janssen-Megens, Anita Kaan, Mihir Kamat, Bowon Kim, Amit Mandoli, Jonathan Marchini, Joost H.A. Martens, Stuart Meacham, Karyn Megy, Jared O'Connell, Romina Petersen, Nilofar Sharifi, Simon M. Sheard, James R. Staley, Salih Tuna, Martijn van der Ent, Klaudia Walter, Shuang Yin Wang, Eleanor Wheeler, Steven P. Wilder, Valentina Iotchkova, Carmel Moore, Jennifer Sambrook, Hendrik G. Stunnenberg, Emanuele Di Angelantonio, Stephen Kaptoge, Taco W. Kuijpers, Enrique Carrillo-de Santa-Pau, David Juan, Daniel Rico, Alfonso Valencia, Lu Chen, Bing

- Ge, Louella Vasquez, Tony Kwan, Diego Garrido-Martín, Stephen Watt, Ying Yang, Roderic Guigo, Stephan Beck, Dirk S. Paul, Tomi Pastinen, David Bujold, Guillaume Bourque, Mattia Frontini, John Danesh, David J. Roberts, Willem H. Ouwehand, Adam S. Butterworth, and Nicole Soranzo. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*, 167(5):1415–1429, 2016. ISSN 10974172. doi: 10.1016/j.cell.2016.10.042.
- [48] Lloyd T. Elliott, Kevin Sharp, Fidel Alfaro-Almagro, Sinan Shi, Karla L. Miller, Gwenaëlle Douaud, Jonathan Marchini, and Stephen M. Smith. Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature*, 2018. ISSN 14764687. doi: 10.1038/s41586-018-0571-7.
- [49] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O’Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018. ISSN 14764687. doi: 10.1038/s41586-018-0579-z. URL <https://doi.org/10.1038/s41586-018-0579-z>.
- [50] Emanuele Di Angelantonio, Simon G. Thompson, Stephen Kaptoge, Carmel Moore, Matthew Walker, Jane Armitage, Willem H. Ouwehand, David J. Roberts, John Danesh, Jenny Donovan, Ian Ford, Rachel Henry, Beverley J. Hunt, Bridget Le Huray, Susan Mehenny, Gail Mifflin, Jane Green, Mike Stredder, Nicholas A. Watkins, Alan McDermott, Clive Ronaldson, Claire Thomson, Zoe Tolkien, Lorna Williamson, David Allen, Jennifer Sambrook, Tracey Hammerton, David Bruce, Fizzah Choudry, Cedric Ghevaert, Kirstie Johnston, Anne Kelly, Adam King, Alfred Mo, Lizanne Page, Penny Richardson, Peter Senior, Yagnesh Umrana, Henna Wong, Gavin Murphy, Adrian C. Newland, Keith Wheatley, Michael Greaves, Marc Turner, Tahir Aziz, Richard Brain, Christine Davies, Ruth Turner, Paula Wakeman, Alison Dent, Alan Wakeman, Ben Anthony, Desmond Bland, Will Parrondo, Helen Vincent, Candy Weatherill, Andrea Forsyth, Carol Butterfield, Tracey Wright, Karen Ellis, Pat Poynton, Carolyn Brooks, Emma Martin, Lara Littler, Lindsay Williams, Donna Blair, Karen Ackerley, Lynn Woods, Sophie Stanley, Gemma Walsh, Gayle Franklin, Cheryl Howath, Sarah Sharpe, Deborah Smith, Lauren Botham, Caroline Williams, Claire Alexander, Gareth Sowerbutts, Diane Furnival, Michael Thake, Shilpa Patel, Carolyn Roost, Sandra Sowerby, Mary Joy Appleton, Eileen Bays, Geoff Bowyer, Steven Clarkson, Stuart Halson, Kate Holmes, Gareth Humphries, Lee Parvin-Cooper, Jason Towler, Joanne Addy, Patricia Barrass, Louise Stennett, Susan Burton, Hannah Dingwall, Victoria Clarke, Maria Potton, Thomas Bolton, Michael Daynes, Sarah Spackman, Abudu Momodu, James Fenton, Omer Muhammed, Nicholas Oates, Tim Peakman, Christine Ryan, Kristian Spreckley, Craig Stubbins, Joanna Williams, James Brennan, Cedric Mochon, Samantha Taylor, Kimberley Warren, Jonathan Mant, and John Danesh. Efficiency and safety of varying the frequency of whole blood donation (INTERVAL): a randomised trial of 45000 donors. *The Lancet*, 390(10110):2360–2371, 2017. ISSN 1474547X. doi: 10.1016/S0140-6736(17)31928-1.
- [51] Emanuele Di Angelantonio. Comparison of alternative strategies to assess haemoglobin levels in whole blood donors (COMPARE study). *ISRCTN registry*, 2017. doi: ISRCTN90871183. URL <https://doi.org/10.1186/ISRCTN90871183>.

- [52] David R. Bentley, Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, Dirk J. Evers, Colin L. Barnes, Helen R. Bignell, Jonathan M. Boutell, Jason Bryant, Richard J. Carter, R. Keira Cheetham, Anthony J. Cox, Darren J. Ellis, Michael R. Flatbush, Niall A. Gormley, Sean J. Humphray, Leslie J. Irving, Mirian S. Karbelashvili, Scott M. Kirk, Heng Li, Xiaohai Liu, Klaus S. Maisinger, Lisa J. Murray, Bojan Obradovic, Tobias Ost, Michael L. Parkinson, Mark R. Pratt, Isabelle M.J. Rasolonjatovo, Mark T. Reed, Roberto Rigatti, Chiara Rodighiero, Mark T. Ross, Andrea Sabot, Subramanian V. Sankar, Aylwyn Scally, Gary P. Schroth, Mark E. Smith, Vincent P. Smith, Anastassia Spiridou, Peta E. Torrance, Svilen S. Tzonev, Eric H. Vermaas, Klaudia Walter, Xiaolin Wu, Lu Zhang, Mohammed D. Alam, Carole Anastasi, Ify C. Aniebo, David M.D. Bailey, Iain R. Bancarz, Saibal Banerjee, Selena G. Barbour, Primo A. Baybayan, Vincent A. Benoit, Kevin F. Benson, Claire Bevis, Phillip J. Black, Asha Boodhun, Joe S. Brennan, John A. Bridgham, Rob C. Brown, Andrew A. Brown, Dale H. Buermann, Abass A. Bundu, James C. Burrows, Nigel P. Carter, Nestor Castillo, Maria Chiara E. Catenazzi, Simon Chang, R. Neil Cooley, Natasha R. Crake, Olubunmi O. Dada, Konstantinos D. Diakoumakos, Belen Dominguez-Fernandez, David J. Earnshaw, Ugonna C. Egbujor, David W. Elmore, Sergey S. Etchin, Mark R. Ewan, Milan Fedurco, Louise J. Fraser, Karin V. Fuentes Fajardo, W. Scott Furey, David George, Kimberley J. Gietzen, Colin P. Goddard, George S. Golda, Philip A. Granieri, David E. Green, David L. Gustafson, Nancy F. Hansen, Kevin Harnish, Christian D. Haudenschild, Narinder I. Heyer, Matthew M. Hims, Johnny T. Ho, Adrian M. Horgan, Katya Hoschler, Steve Hurwitz, Denis V. Ivanov, Maria Q. Johnson, Terena James, T. A. Huw Jones, Gyoung Dong Kang, Tzvetana H. Kerelska, Alan D. Kersey, Irina Khrebtukova, Alex P. Kindwall, Zoya Kingsbury, Paula I. Kokko-Gonzales, Anil Kumar, Marc A. Laurent, Cynthia T. Lawley, Sarah E. Lee, Xavier Lee, Arnold K. Liao, Jennifer A. Loch, Mitch Lok, Shujun Luo, Radhika M. Mammen, John W. Martin, Patrick G. McCauley, Paul McNitt, Parul Mehta, Keith W. Moon, Joe W. Mullens, Taksina Newington, Zemin Ning, Bee Ling Ng, Sonia M. Novo, Michael J. O'Neill, Mark A. Osborne, Andrew Osnowski, Omead Ostadan, Lambros L. Paraschos, Lea Pickering, Andrew C. Pike, Alger C. Pike, D. Chris Pinkard, Daniel P. Pliskin, Joe Podhasky, Victor J. Quijano, Come Raczy, Vicki H. Rae, Stephen R. Rawlings, Ana Chiva Rodriguez, Phyllida M. Roe, John Rogers, Maria C. Rogert Bacigalupo, Nikolai Romanov, Anthony Romieu, Rithy K. Roth, Natalie J. Rourke, Silke T. Ruediger, Eli Rusman, Raquel M. Sanches-Kuiper, Martin R. Schenker, Josefina M. Seoane, Richard J. Shaw, Mitch K. Shiver, Steven W. Short, Ning L. Sizto, Johannes P. Sluis, Melanie A. Smith, Jean Ernest Sohna Sohna, Eric J. Spence, Kim Stevens, Neil Sutton, Lukasz Szajkowski, Carolyn L. Tregidgo, Gerardo Turcatti, Stephanie Vandevondele, Yuli Verhovsky, Selene M. Virk, Suzanne Wakelin, Gregory C. Walcott, Jingwen Wang, Graham J. Worsley, Juying Yan, Ling Yau, Mike Zuerlein, Jane Rogers, James C. Mullikin, Matthew E. Hurles, Nick J. McCooke, John S. West, Frank L. Oaks, Peter L. Lundberg, David Klenerman, Richard Durbin, and Anthony J. Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 2008. ISSN 00280836. doi: 10.1038/nature07517.
- [53] Adam Auton, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Peter Donnelly, Evan E. Eichler, Paul Flicek, Stacey B. Gabriel, Richard A. Gibbs, Eric D. Green, Matthew E. Hurles, Bartha M. Knoppers, Jan O. Korbel, Eric S. Lander, Charles Lee, Hans Lehrach,

Elaine R. Mardis, Gabor T. Marth, Gil A. McVean, Deborah A. Nickerson, Jeanette P. Schmidt, Stephen T. Sherry, Jun Wang, Richard K. Wilson, Eric Boerwinkle, Harsha Doddapaneni, Yi Han, Viktoriya Korchina, Christie Kovar, Sandra Lee, Donna Muzny, Jeffrey G. Reid, Yiming Zhu, Yuqi Chang, Qiang Feng, Xiaodong Fang, Xiaosen Guo, Min Jian, Hui Jiang, Xin Jin, Tianming Lan, Guoqing Li, Jingxiang Li, Yingrui Li, Shengmao Liu, Xiao Liu, Yao Lu, Xuedi Ma, Meifang Tang, Bo Wang, Guangbiao Wang, Honglong Wu, Renhua Wu, Xun Xu, Ye Yin, Dandan Zhang, Wenwei Zhang, Jiao Zhao, Meiru Zhao, Xiaole Zheng, Namrata Gupta, Neda Gharani, Lorraine H. Toji, Norman P. Gerry, Alissa M. Resch, Jonathan Barker, Laura Clarke, Laurent Gil, Sarah E. Hunt, Gavin Kelman, Eugene Kulesha, Rasko Leinonen, William M. McLaren, Rajesh Radhakrishnan, Asier Roa, Dmitriy Smirnov, Richard E. Smith, Ian Streeter, Anja Thormann, Iliana Toneva, Brendan Vaughan, Xiangqun Zheng-Bradley, Russell Grocock, Sean Humphray, Terena James, Zoya Kingsbury, Ralf Sudbrak, Marcus W. Albrecht, Vyacheslav S. Amstislavskiy, Tatiana A. Borodina, Matthias Lienhard, Florian Mertes, Marc Sultan, Bernd Timmermann, Marie Laure Yaspo, Lucinda Fulton, Victor Ananiev, Zinaida Belaia, Dimitriy Beloslyudtsev, Nathan Bouk, Chao Chen, Deanna Church, Robert Cohen, Charles Cook, John Garner, Timothy Hefferon, Mikhail Kimelman, Chunlei Liu, John Lopez, Peter Meric, Chris O'Sullivan, Yuri Ostapchuk, Lon Phan, Sergiy Ponomarov, Valerie Schneider, Eugene Shekhtman, Karl Sirotkin, Douglas Slotta, Hua Zhang, Senduran Balasubramaniam, John Burton, Petr Danecek, Thomas M. Keane, Anja Kolb-Kokocinski, Shane McCarthy, James Stalker, Michael Quail, Christopher J. Davies, Jeremy Gollub, Teresa Webster, Brant Wong, Yiping Zhan, Christopher L. Campbell, Yu Kong, Anthony Marcketta, Fuli Yu, Lilian Antunes, Matthew Bainbridge, Aniko Sabo, Zhuoyi Huang, Lachlan J.M. Coin, Lin Fang, Qibin Li, Zhenyu Li, Haoxiang Lin, Binghang Liu, Ruibang Luo, Haojing Shao, Yinlong Xie, Chen Ye, Chang Yu, Fan Zhang, Hancheng Zheng, Hongmei Zhu, Can Alkan, Elif Dal, Fatma Kahveci, Erik P. Garrison, Deniz Kural, Wan Ping Lee, Wen Fung Leong, Michael Stromberg, Alistair N. Ward, Jiantao Wu, Mengyao Zhang, Mark J. Daly, Mark A. DePristo, Robert E. Handsaker, Eric Banks, Gaurav Bhatia, Guillermo Del Angel, Giulio Genovese, Heng Li, Seva Kashin, Steven A. McCarroll, James C. Nemes, Ryan E. Poplin, Seungtae C. Yoon, Jayon Lihm, Vladimir Makarov, Srikanth Gottipati, Alon Keinan, Juan L. Rodriguez-Flores, Tobias Rausch, Markus H. Fritz, Adrian M. Stütz, Kathryn Beal, Avik Datta, Javier Herrero, Graham R.S. Ritchie, Daniel Zerbino, Pardis C. Sabeti, Ilya Shlyakhter, Stephen F. Schaffner, Joseph Vitti, David N. Cooper, Edward V. Ball, Peter D. Stenson, Bret Barnes, Markus Bauer, R. Keira Cheetham, Anthony Cox, Michael Eberle, Scott Kahn, Lisa Murray, John Peden, Richard Shaw, Eimear E. Kenny, Mark A. Batzer, Miriam K. Konkel, Jerilyn A. Walker, Daniel G. MacArthur, Monkol Lek, Ralf Herwig, Li Ding, Daniel C. Koboldt, David Larson, Kai Ye, Simon Gravel, Anand Swaroop, Emily Chew, Tuuli Lappalainen, Yaniv Erlich, Melissa Gymrek, Thomas Frederick Willems, Jared T. Simpson, Mark D. Shriver, Jeffrey A. Rosenfeld, Carlos D. Bustamante, Stephen B. Montgomery, Francisco M. De La Vega, Jake K. Byrnes, Andrew W. Carroll, Marianne K. DeGorter, Phil Lacroute, Brian K. Maples, Alicia R. Martin, Andres Moreno-Estrada, Suyash S. Shringarpure, Fouad Zakharia, Eran Halperin, Yael Baran, Eliza Cerveira, Jaeho Hwang, Ankit Malhotra, Dariusz Plewczynski, Kamen Radew, Mallory Romanovitch, Chengsheng Zhang, Fiona C.L. Hyland, David W. Craig, Alexis Christoforides, Nils Homer, Tyler Izatt, Ahmet A. Kurdoglu, Shripad A. Sinari, Kevin Squire, Chunlin Xiao, Jonathan Sebat, Danny

- Antaki, Madhusudan Gujral, Amina Noor, Kenny Ye, Esteban G. Burchard, Ryan D. Hernandez, Christopher R. Gignoux, David Haussler, Sol J. Katzman, W. James Kent, Bryan Howie, Andres Ruiz-Linares, Emmanouil T. Dermitzakis, Scott E. Devine, Hyun Min Kang, Jeffrey M. Kidd, Tom Blackwell, Sean Caron, Wei Chen, Sarah Emery, Lars Fritsche, Christian Fuchsberger, Goo Jun, Bingshan Li, Robert Lyons, Chris Scheller, Carlo Sidore, Shiya Song, Elzbieta Sliwerska, Daniel Taliun, Adrian Tan, Ryan Welch, Mary Kate Wing, Xiaowei Zhan, Philip Awadalla, Alan Hodgkinson, Yun Li, Xinghua Shi, Andrew Quitadamo, Gerton Lunter, Jonathan L. Marchini, Simon Myers, Claire Churchhouse, Olivier Delaneau, Anjali Gupta-Hinch, Warren Kretzschmar, Zamin Iqbal, Iain Mathieson, Androniki Menelaou, Andy Rimmer, Dionysia K. Xifara, Taras K. Oleksyk, Yunxin Fu, Xiaoming Liu, Momiao Xiong, Lynn Jorde, David Witherspoon, Jinchuan Xing, Brian L. Browning, Sharon R. Browning, Fereydoun Hormozdiari, Peter H. Sudmant, Ekta Khurana, Chris Tyler-Smith, Cornelis A. Albers, Qasim Ayub, Yuan Chen, Vincenza Colonna, Luke Jostins, Klaudia Walter, Yali Xue, Mark B. Gerstein, Alexej Abyzov, Suganthi Balasubramanian, Jieming Chen, Declan Clarke, Yao Fu, Arif O. Harmanci, Mike Jin, Donghoon Lee, Jeremy Liu, Xinmeng Jasmine Mu, Jing Zhang, Yan Zhang, Chris Hartl, Khalid Shakir, Jeremiah Degenhardt, Sascha Meiers, Benjamin Raeder, Francesco Paolo Casale, Oliver Stegle, Eric Wubbo Lameijer, Ira Hall, Vineet Bafna, Jacob Michaelson, Eugene J. Gardner, Ryan E. Mills, Gargi Dayama, Ken Chen, Xian Fan, Zechen Chong, Tenghui Chen, Mark J. Chaisson, John Huddleston, Maika Malig, Bradley J. Nelson, Nicholas F. Parrish, Ben Blackburne, Sarah J. Lindsay, Zemin Ning, Yujun Zhang, Hugo Lam, Cristina Sisu, Danny Challis, Uday S. Evani, James Lu, Uma Nagaswamy, Jin Yu, Wangshen Li, Lukas Habegger, Haiyuan Yu, Fiona Cunningham, Ian Dunham, Kasper Lage, Jakob Berg Jaspersen, Heiko Horn, Donghoon Kim, Rob Desalle, Apurva Narechania, Melissa A. Wilson Sayres, Fernando L. Mendez, G. David Poznik, Peter A. Underhill, David Mittelman, Ruby Banerjee, Maria Cerezo, Thomas W. Fitzgerald, Sandra Louzada, Andrea Massaia, Fengtang Yang, Divya Kalra, Walker Hale, Xu Dan, Kathleen C. Barnes, Christine Beiswanger, Hongyu Cai, Hongzhi Cao, Brenna Henn, Danielle Jones, Jane S. Kaye, Alastair Kent, Angeliki Kerasidou, Rasika Mathias, Pilar N. Ossorio, Michael Parker, Charles N. Rotimi, Charmaine D. Royal, Karla Sandoval, Yeyang Su, Zhongming Tian, Sarah Tishkoff, Marc Via, Yuhong Wang, Huanming Yang, Ling Yang, Jiayong Zhu, Walter Bodmer, Gabriel Bedoya, Zhiming Cai, Yang Gao, Jiayou Chu, Leena Peltonen, Andres Garcia-Montero, Alberto Orfao, Julie Dutil, Juan C. Martinez-Cruzado, Rasika A. Mathias, Anselm Hennis, Harold Watson, Colin McKenzie, Firdausi Qadri, Regina LaRocque, Xiaoyan Deng, Danny Asogun, Onikepe Folarin, Christian Happi, Omonwunmi Omoniwa, Matt Stremlau, Ridhi Tariyal, Muminatou Jallow, Fatoumatta Sisay Joof, Tumani Corrah, Kirk Rockett, Dominic Kwiatkowski, Jaspal Kooner, Tran Tinh Hien, Sarah J. Dunstan, Nguyen ThuyHang, Richard Fonnier, Robert Garry, Lansana Kanneh, Lina Moses, John Schieffelin, Donald S. Grant, Carla Gallo, Giovanni Poletti, Danish Saleheen, Asif Rasheed, Lisa D. Brooks, Adam L. Felsenfeld, Jean E. McEwen, Yekaterina Vaydylevich, Audrey Duncanson, Michael Dunn, and Jeffery A. Schloss. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015. ISSN 14764687. doi: 10.1038/nature15393.
- [54] Monkol Lek, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O'Donnell-Luria, James S. Ware, Andrew J. Hill, Beryl B. Cummings, Taru Tukiainen, Daniel P. Birnbaum, Jack A. Kosmicki, Laramie E. Dun-

- can, Karol Estrada, Fengmei Zhao, James Zou, Emma Pierce-Hoffman, Joanne Berghout, David N. Cooper, Nicole Deflaux, Mark DePristo, Ron Do, Jason Flannick, Menachem Fromer, Laura Gauthier, Jackie Goldstein, Namrata Gupta, Daniel Howrigan, Adam Kiezun, Mitja I. Kurki, Ami Levy Moonshine, Pradeep Natarajan, Lorena Orozco, Gina M. Peloso, Ryan Poplin, Manuel A. Rivas, Valentin Ruano-Rubio, Samuel A. Rose, Douglas M. Ruderfer, Khalid Shakir, Peter D. Stenson, Christine Stevens, Brett P. Thomas, Grace Tiao, Maria T. Tusie-Luna, Ben Weisburd, Hong Hee Won, Dongmei Yu, David M. Altshuler, Diego Ardissino, Michael Boehnke, John Danesh, Stacey Donnelly, Roberto Elosua, Jose C. Florez, Stacey B. Gabriel, Gad Getz, Stephen J. Glatt, Christina M. Hultman, Sekar Kathiresan, Markku Laakso, Steven McCarroll, Mark I. McCarthy, Dermot McGovern, Ruth McPherson, Benjamin M. Neale, Aarno Palotie, Shaun M. Purcell, Danish Saleheen, Jeremiah M. Scharf, Pamela Sklar, Patrick F. Sullivan, Jaakko Tuomilehto, Ming T. Tsuang, Hugh C. Watkins, James G. Wilson, Mark J. Daly, Daniel G. MacArthur, H. E. Abboud, G. Abecasis, C. A. Aguilar-Salinas, O. Arellano-Campos, G. Atzmon, I. Aukrust, C. L. Barr, G. I. Bell, S. Bergen, L. Bjørkhaug, J. Blangero, D. W. Bowden, C. L. Budman, N. P. Burt, F. Centeno-Cruz, J. C. Chambers, K. Chambert, R. Clarke, R. Collins, G. Coppola, E. J. Córdova, M. L. Cortes, N. J. Cox, R. Duggirala, M. Farrall, J. C. Fernandez-Lopez, P. Fontanillas, T. M. Frayling, N. B. Freimer, C. Fuchsberger, H. García-Ortiz, A. Goel, M. J. Gómez-Vázquez, M. E. González-Villalpando, C. González-Villalpando, M. A. Grados, L. Groop, C. A. Haiman, C. L. Hanis, A. T. Hattersley, B. E. Henderson, J. C. Hopewell, A. Huerta-Chagoya, S. Islas-Andrade, S. B. Jacobs, S. Jalilzadeh, C. P. Jenkinson, J. Moran, S. Jiménez-Morale, A. Kähler, R. A. King, G. Kirov, J. S. Kooner, T. Kyriakou, J. Y. Lee, D. M. Lehman, G. Lyon, W. MacMahon, P. K. Magnusson, A. Mahajan, J. Marrugat, A. Martínez-Hernández, C. A. Mathews, G. McVean, J. B. Meigs, T. Meitinger, E. Mendoza-Caamal, J. M. Mercader, K. L. Mohlke, H. Moreno-Macías, A. P. Morris, L. A. Najmi, P. R. Njølstad, M. C. O'Donovan, M. L. Ordóñez-Sánchez, M. J. Owen, T. Park, D. L. Pauls, D. Posthuma, C. Revilla-Monsalve, L. Riba, S. Ripke, R. Rodríguez-Guillén, M. Rodríguez-Torres, P. Sandor, M. Seielstad, R. Sladek, X. Soberón, T. D. Spector, S. E. Tai, T. M. Teslovich, G. Walford, L. R. Wilkens, and A. L. Williams. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, 2016. ISSN 14764687. doi: 10.1038/nature19057. URL <http://www.ncbi.nlm.nih.gov/pubmed/27535533> <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5018207>.
- [55] Sohyun Hwang, Eiru Kim, Insuk Lee, and Edward M. Marcotte. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports*, 5, 2015. ISSN 20452322. doi: 10.1038/srep17875.
- [56] Nuala Moran. 10,000 rare-disease genomes sequenced. *Nature Biotechnology*, 2014. ISSN 1087-0156. doi: 10.1038/nbt0114-7.
- [57] Daniel Taliun, Daniel Harris, Michael Kessler, Jedidiah Carlson, Zachary Szpiech, Raul Torres, Sarah Taliun, André Corvelo, Stephanie Gogarten, Hyun Min Kang, Achilleas Pitsillides, Jonathon LeFaive, Seung-been Lee, Xiaowen Tian, Brian Browning, Sayantan Das, Anne-Katrin Emde, Wayne Clarke, Douglas Loesch, Amol Shetty, Thomas Blackwell, Quenna Wong, François Aguet, Christine Albert, Alvaro Alonso, Kristin Ardlie, Stella Aslibekyan, Paul Auer, John Barnard, R. Graham Barr, Lewis Becker, Rebecca Beer, Emelia Benjamin, Lawrence Bielak, John Blangero, Michael

- Boehnke, Donald Bowden, Jennifer Brody, Esteban Burchard, Brian Cade, James Casella, Brandon Chalazan, Yii-Der Ida Chen, Michael Cho, Seung Hoan Choi, Mina Chung, Clary Clish, Adolfo Correa, Joanne Curran, Brian Custer, Dawood Darbar, Michelle Daya, Mariza de Andrade, Dawn DeMeo, Susan Dutcher, Patrick Ellinor, Leslie Emery, Diane Fatkin, Lukas Forer, Myriam Fornage, Nora Franceschini, Christian Fuchsberger, Stephanie Fullerton, Soren Germer, Mark Gladwin, Daniel Gottlieb, Xiuqing Guo, Michael Hall, Jiang He, Nancy Heard-Costa, Susan Heckbert, Marguerite Irvin, Jill Johnsen, Andrew Johnson, Sharon Kardia, Tanika Kelly, Shannon Kelly, Eimear Kenny, Douglas Kiel, Robert Klemmer, Barbara Konkle, Charles Kooperberg, Anna Köttgen, Leslie Lange, Jessica Lasky-Su, Daniel Levy, Xihong Lin, Keng-Han Lin, Chunyu Liu, Ruth Loos, Lori Garman, Robert Gerszten, Steven Lubitz, Kathryn Lunetta, Angel Mak, Ani Manichaikul, Alisa Manning, Rasika Mathias, David McManus, Stephen McGarvey, James Meigs, Deborah Meyers, Julie Mikulla, Mollie Minear, Braxton Mitchell, Sanghamitra Mohanty, May Montasser, Courtney Montgomery, Alanna Morrison, Joanne Murabito, Andrea Natale, Pradeep Natarajan, Sarah Nelson, Kari North, Jeffrey O'Connell, Nicholette Palmer, Nathan Pankratz, Gina Peloso, Patricia Peyser, Wendy Post, Bruce Psaty, D.C. Rao, Susan Redline, Alexander Reiner, Dan Roden, Jerome Rotter, Ingo Ruczinski, Chloé Sarnowski, Sebastian Schoenherr, Jeong-Sun Seo, Sudha Seshadri, Vivien Sheehan, M. Benjamin Shoemaker, Albert Smith, Nicholas Smith, Jennifer Smith, Nona Sotoodehnia, Adrienne Stilp, Weihong Tang, Kent Taylor, Marilyn Telen, Timothy Thornton, Russell Tracy, David Berg, Ramachandran Vasani, Karine Viaud-Martinez, Scott Vrieze, Daniel Weeks, Bruce Weir, Scott Weiss, Lu-Chen Weng, Cristen Willer, Yingze Zhang, Xutong Zhao, Donna Arnett, Allison Ashley-Koch, Kathleen Barnes, Eric Boerwinkle, Stacey Gabriel, Richard Gibbs, Kenneth Rice, Stephen Rich, Edwin Silverman, Pankaj Qasba, Weiniu Gan, George Papanicolaou, Deborah Nickerson, Sharon Browning, Michael Zody, Sebastian Zöllner, James Wilson, L Adrienne Cupples, Cathy Laurie, Cashell Jaquish, Ryan Hernandez, Timothy O'Connor, and Gonçalo Abecasis. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *bioRxiv*, 2019. doi: 10.1101/563866.
- [58] Ryan L. Collins, Harrison Brand, Konrad J. Karczewski, Xuefang Zhao, Jessica Alföldi, Laurent C. Francioli, Amit V. Khera, Chelsea Lowther, Laura D. Gauthier, Harold Wang, Nicholas A. Watts, Matthew Solomonson, Anne O'Donnell-Luria, Alexander Baumann, Ruchi Munshi, Mark Walker, Christopher W. Whelan, Yongqing Huang, Ted Brookings, Ted Sharpe, Matthew R. Stone, Elise Valkanas, Jack Fu, Grace Tiao, Kristen M. Laricchia, Valentin Ruano-Rubio, Christine Stevens, Namrata Gupta, Caroline Cusick, Lauren Margolin, Jessica Alföldi, Irina M. Armean, Eric Banks, Louis Bergelson, Kristian Cibulskis, Ryan L. Collins, Kristen M. Connolly, Miguel Covarrubias, Beryl Cummings, Mark J. Daly, Stacey Donnelly, Yossi Farjoun, Steven Ferreira, Laurent Francioli, Stacey Gabriel, Laura D. Gauthier, Jeff Gentry, Namrata Gupta, Thibault Jeandet, Diane Kaplan, Konrad J. Karczewski, Kristen M. Laricchia, Christopher Llanwarne, Eric V. Minikel, Ruchi Munshi, Benjamin M. Neale, Sam Novod, Anne H. O'Donnell-Luria, Nikelle Petrillo, Timothy Poterba, David Roazen, Valentin Ruano-Rubio, Andrea Saltzman, Kaitlin E. Samocha, Molly Schleicher, Cotton Seed, Matthew Solomonson, Jose Soto, Grace Tiao, Kathleen Tibbetts, Charlotte Tolonen, Christopher Vittal, Gordon Wade, Arcturus Wang, Qingbo Wang, James S. Ware, Nicholas A. Watts, Ben Weisburd, Nicola Whiffin, Carlos A. Aguilar Salinas, Tariq Ahmad, Christine M. Albert, Diego Ardissono, Gil Atzmon, John Barnard, Lau-

- rent Beaugerie, Emelia J. Benjamin, Michael Boehnke, Lori L. Bonnycastle, Erwin P. Bottinger, Donald W. Bowden, Matthew J. Bown, John C. Chambers, Juliana C. Chan, Daniel Chasman, Judy Cho, Mina K. Chung, Bruce Cohen, Adolfo Correa, Dana Dabelea, Mark J. Daly, Dawood Darbar, Ravindranath Duggirala, Josée Dupuis, Patrick T. Ellinor, Roberto Elosua, Jeanette Erdmann, Tõnu Esko, Martti Färkkilä, Jose Florez, Andre Franke, Gad Getz, Benjamin Glaser, Stephen J. Glatt, David Goldstein, Clicerio Gonzalez, Leif Groop, Christopher Haiman, Craig Hanis, Matthew Harms, Mikko Hiltunen, Matti M. Holi, Christina M. Hultman, Mikko Kallela, Jaakko Kaprio, Sekar Kathiresan, Bong Jo Kim, Young Jin Kim, George Kirov, Jaspal Kooner, Seppo Koskinen, Harlan M. Krumholz, Subra Kugathasan, Soo Heon Kwak, Markku Laakso, Terho Lehtimäki, Ruth J.F. Loos, Steven A. Lubitz, Ronald C.W. Ma, Daniel G. MacArthur, Jaume Marrugat, Kari M. Mattila, Steven McCarroll, Mark I. McCarthy, Dermot McGovern, Ruth McPherson, James B. Meigs, Olle Melander, Andres Metspalu, Benjamin M. Neale, Peter M. Nilsson, Michael C. O'Donovan, Dost Ongur, Lorena Orozco, Michael J. Owen, Colin N.A. Palmer, Aarno Palotie, Kyong Soo Park, Carlos Pato, Ann E. Pulver, Nazneen Rahman, Anne M. Remes, John D. Rioux, Samuli Ripatti, Dan M. Roden, Danish Saleheen, Veikko Salomaa, Nilesh J. Samani, Jeremiah Scharf, Heribert Schunkert, Moore B. Shoemaker, Pamela Sklar, Hilikka Soininen, Harry Sokol, Tim Spector, Patrick F. Sullivan, Jaana Suvisaari, E. Shyong Tai, Yik Ying Teo, Tuomi Tiinamäijä, Ming Tsuang, Dan Turner, Teresa Tusie-Luna, Erkki Vartiainen, James S. Ware, Hugh Watkins, Rinse K. Weersma, Maija Wessman, James G. Wilson, Ramnik J. Xavier, Kent D. Taylor, Henry J. Lin, Stephen S. Rich, Wendy S. Post, Yii Der Ida Chen, Jerome I. Rotter, Chad Nusbaum, Anthony Philippakis, Eric Lander, Stacey Gabriel, Benjamin M. Neale, Sekar Kathiresan, Mark J. Daly, Eric Banks, Daniel G. MacArthur, and Michael E. Talkowski. A structural variation reference for medical and population genetics. *Nature*, 2020. ISSN 14764687. doi: 10.1038/s41586-020-2287-8.
- [59] Konrad J. Karczewski, Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, Kristen M. Laricchia, Andrea Ganna, Daniel P. Birnbaum, Laura D. Gauthier, Harrison Brand, Matthew Solomonson, Nicholas A. Watts, Daniel Rhodes, Moriel Singer-Berk, Eleina M. England, Eleanor G. Seaby, Jack A. Kosmicki, Raymond K. Walters, Katherine Tashman, Yossi Farjoun, Eric Banks, Timothy Poterba, Arcturus Wang, Cotton Seed, Nicola Whiffin, Jessica X. Chong, Kaitlin E. Samocha, Emma Pierce-Hoffman, Zachary Zappala, Anne H. O'Donnell-Luria, Eric Vallabh Minikel, Ben Weisburd, Monkol Lek, James S. Ware, Christopher Vittal, Irina M. Armean, Louis Bergelson, Kristian Cibulskis, Kristen M. Connolly, Miguel Covarrubias, Stacey Donnelly, Steven Ferreira, Stacey Gabriel, Jeff Gentry, Namrata Gupta, Thibault Jeandet, Diane Kaplan, Christopher Llanwarne, Ruchi Munshi, Sam Novod, Nikelle Petrillo, David Roazen, Valentin Ruano-Rubio, Andrea Saltzman, Molly Schleicher, Jose Soto, Kathleen Tibbetts, Charlotte Tolonen, Gordon Wade, Michael E. Talkowski, Carlos A. Aguilar Salinas, Tariq Ahmad, Christine M. Albert, Diego Ardissino, Gil Atzmon, John Barnard, Laurent Beaugerie, Emelia J. Benjamin, Michael Boehnke, Lori L. Bonnycastle, Erwin P. Bottinger, Donald W. Bowden, Matthew J. Bown, John C. Chambers, Juliana C. Chan, Daniel Chasman, Judy Cho, Mina K. Chung, Bruce Cohen, Adolfo Correa, Dana Dabelea, Mark J. Daly, Dawood Darbar, Ravindranath Duggirala, Josée Dupuis, Patrick T. Ellinor, Roberto Elosua, Jeanette Erdmann, Tõnu Esko, Martti Färkkilä, Jose Florez, Andre Franke, Gad Getz, Benjamin Glaser, Stephen J. Glatt, David Goldstein, Clicerio

- Gonzalez, Leif Groop, Christopher Haiman, Craig Hanis, Matthew Harms, Mikko Hiltunen, Matti M. Holi, Christina M. Hultman, Mikko Kallela, Jaakko Kaprio, Sekar Kathiresan, Bong Jo Kim, Young Jin Kim, George Kirov, Jaspal Kooner, Seppo Koskinen, Harlan M. Krumholz, Subra Kugathasan, Soo Heon Kwak, Markku Laakso, Terho Lehtimäki, Ruth J.F. Loos, Steven A. Lubitz, Ronald C.W. Ma, Daniel G. MacArthur, Jaume Marrugat, Kari M. Mattila, Steven McCarroll, Mark I. McCarthy, Dermot McGovern, Ruth McPherson, James B. Meigs, Olle Melander, Andres Metspalu, Benjamin M. Neale, Peter M. Nilsson, Michael C. O'Donovan, Dost Ongur, Lorena Orozco, Michael J. Owen, Colin N.A. Palmer, Aarno Palotie, Kyong Soo Park, Carlos Pato, Ann E. Pulver, Nazneen Rahman, Anne M. Remes, John D. Rioux, Samuli Ripatti, Dan M. Roden, Danish Saleheen, Veikko Salomaa, Nilesh J. Samani, Jeremiah Scharf, Heribert Schunkert, Moore B. Shoemaker, Pamela Sklar, Hilkka Soininen, Harry Sokol, Tim Spector, Patrick F. Sullivan, Jaana Suvisaari, E. Shyong Tai, Yik Ying Teo, Tuomi Tiinamaija, Ming Tsuang, Dan Turner, Teresa Tusie-Luna, Erkki Vartiainen, James S. Ware, Hugh Watkins, Rinse K. Weersma, Maija Wessman, James G. Wilson, Ramnik J. Xavier, Benjamin M. Neale, Mark J. Daly, and Daniel G. MacArthur. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 2020. ISSN 14764687. doi: 10.1038/s41586-020-2308-7.
- [60] Ernest Turro, William J. Astle, Karyn Megy, Stefan Gräf, Daniel Greene, Olga Shamardina, Hana Lango Allen, Alba Sanchis-Juan, Mattia Frontini, Chantal Thys, Jonathan Stephens, Rutendo Mapeta, Oliver S. Burren, Kate Downes, Matthias Haimel, Salih Tuna, Sri V.V. Deevi, Timothy J. Aitman, David L. Bennett, Paul Calleja, Keren Carss, Mark J. Caulfield, Patrick F. Chinnery, Peter H. Dixon, Daniel P. Gale, Roger James, Ania Koziell, Michael A. Laffan, Adam P. Levine, Eamonn R. Maher, Hugh S. Markus, Joannella Morales, Nicholas W. Morrell, Andrew D. Mumford, Elizabeth Ormondroyd, Stuart Rankin, Augusto Rendon, Sylvia Richardson, Irene Roberts, Noemi B.A. Roy, Moin A. Saleem, Kenneth G.C. Smith, Hannah Stark, Rhea Y.Y. Tan, Andreas C. Themistocleous, Adrian J. Thrasher, Hugh Watkins, Andrew R. Webster, Martin R. Wilkins, Catherine Williamson, James Whitworth, Sean Humphray, David R. Bentley, Stephen Abbs, Lara Abulhoul, Julian Adlard, Munaza Ahmed, Timothy J. Aitman, Hana Alachkar, David J. Allsup, Jeff Almeida-King, Philip Ancliff, Richard Antrobus, Ruth Armstrong, Gavin Arno, Sofie Ashford, William J. Astle, Anthony Attwood, Paul Aurora, Christian Babbs, Chiara Bacchelli, Tamam Bakchoul, Siddharth Banka, Tadbir Bariana, Julian Barwell, Joana Batista, Helen E. Baxendale, Phil L. Beales, David L. Bennett, David R. Bentley, Agnieszka Bierzynska, Tina Biss, Maria A.K. Bitner-Glindzicz, Graeme C. Black, Marta Bleda, Iulia Blesneac, Detlef Bockenhauer, Harm Bogaard, Christian J. Bourne, Sara Boyce, John R. Bradley, Eugene Bragin, Gerome Breen, Paul Brennan, Carole Brewer, Matthew Brown, Andrew C. Browning, Michael J. Browning, Rachel J. Buchan, Matthew S. Buckland, Teofila Bueser, Carmen Bugarin Diz, John Burn, Siobhan O. Burns, Oliver S. Burren, Nigel Burrows, Paul Calleja, Carolyn Campbell, Gerald Carr-White, Keren Carss, Ruth Casey, Mark J. Caulfield, Jenny Chambers, John Chambers, Melanie M.Y. Chan, Calvin Cheah, Floria Cheng, Patrick F. Chinnery, Manali Chitre, Martin T. Christian, Colin Church, Jill Clayton-Smith, Maureen Cleary, Naomi Clements Brod, Gerry Coghlan, Elizabeth Colby, Trevor R.P. Cole, Janine Collins, Peter W. Collins, Camilla Colombo, Cecilia J. Compton, Robin Condliffe, Stuart Cook, H. Terence Cook, Nichola Cooper, Paul A. A. Corris, Abigail Furnell, Fiona Cunningham, Nicola S. Curry, Antony J. Cutler, Matthew J. Daniels, Mehul Dattani, Louise C. Daugherty, John Davis, Anthony

De Soyza, Sri V.V. Deevi, Timothy Dent, Charu Deshpande, Eleanor F. Dewhurst, Peter H. Dixon, Sofia Douzgou, Kate Downes, Anna M. Drazyk, Elizabeth Drewe, Daniel Duarte, Tina Dutt, J. David M. Edgar, Karen Edwards, William Egner, Melanie N. Ekani, Perry Elliott, Wendy N. Erber, Marie Erwood, Maria C. Estiu, Dafydd Gareth Evans, Gillian Evans, Tamara Everington, Mélanie Eyries, Hiva Fassihi, Remi Favier, Jack Findhammer, Debra Fletcher, Frances A. Flinter, R. Andres Floto, Tom Fowler, James Fox, Amy J. Frary, Courtney E. French, Kathleen Freson, Mattia Frontini, Daniel P. Gale, Henning Gall, Vijeya Ganesan, Michael Gattens, Claire Geoghegan, Terence S.A. Gerighty, Ali G. Gharavi, Stefano Ghio, Hossein Ardeschir Ghofrani, J. Simon R. Gibbs, Kate Gibson, Kimberly C. Gilmour, Barbara Girerd, Nicholas S. Gleadall, Sarah Goddard, David B. Goldstein, Keith Gomez, Pavels Gordins, David Gosal, Stefan Gräf, Jodie Graham, Luigi Grassi, Daniel Greene, Lynn Greenhalgh, Andreas Greinacher, Paolo Gresele, Philip Griffiths, Sofia Grigoriadou, Russell J. Grocock, Detelina Grozeva, Mark Gurnell, Scott Hackett, Charaka Hadinnapola, William M. Hague, Rosie Hague, Matthias Haimel, Matthew Hall, Helen L. Hanson, Eshika Haque, Kirsty Harkness, Andrew R. Harper, Claire L L. Harris, Daniel Hart, Ahamad Hassan, Grant Hayman, Alex Henderson, Archana Herwadkar, Jonathan Hoffman, Simon Holden, Rita Horvath, Henry Houlden, Arjan C C. Houweling, Luke S. Howard, Fengyuan Hu, Gavin Hudson, Joseph Hughes, Aarnoud P. Huissoon, Marc Humbert, Sean Humphray, Sarah Hunter, Matthew Hurles, Melita Irving, Louise Izatt, Roger James, Sally A. Johnson, Stephen Jolles, Jennifer Jolley, Dragana Josifova, Neringa Jurkute, Tim Karten, Johannes Karten, Mary A. Kasanicki, Hanadi Kazkaz, Rashid Kazmi, Peter Kelleher, Anne M. Kelly, Wilf Kelsall, Carly Kempster, David G. Kiely, Nathalie Kingston, Robert Klima, Nils Koelling, Myrto Kostadima, Gabor Kovacs, Ania Koziell, Roman Kreuzhuber, Taco W. Kuijpers, Ajith Kumar, Dinakantha Kumararatne, Manju A. Kurian, Michael A. Laffan, Fiona Laloo, Michele Lambert, Hana Lango Allen, Allan Lawrie, D. Mark Layton, Nick Lench, Claire Lentaigne, Tracy Lester, Adam P. Levine, Rachel Linger, Hilary Longhurst, Lorena E. Lorenzo, Eleni Louka, Paul A. Lyons, Rajiv D. Machado, Robert V. MacKenzie Ross, Bella Madan, Eamonn R. Maher, Jesmeen Maimaris, Samantha Malka, Sarah Mangles, Rutendo Mapeta, Kevin J. Marchbank, Stephen Marks, Hugh S. Markus, Hanns Ulrich Marschall, Andrew Marshall, Jennifer Martin, Mary Mathias, Emma Matthews, Heather Maxwell, Paul McAlinden, Mark I. McCarthy, Harriet McKinney, Aoife McMahan, Stuart Meacham, Adam J. Mead, Ignacio Medina Castello, Karyn Megy, Sarju G G. Mehta, Michel Michaelides, Carolyn Millar, Shehla N. Mohammed, Shahin Moledina, David Montani, Anthony T. Moore, Joannella Morales, Nicholas W. Morrell, Monika Mozere, Keith W. Muir, Andrew D. Mumford, Andrea H. Nemeth, William G. Newman, Michael Newnham, Sadia Noorani, Paquita Nurden, Jennifer O'Sullivan, Samya Obaji, Chris Odhams, Steven Okoli, Andrea Olschewski, Horst Olschewski, Kai Ren Ong, S. Helen Oram, Elizabeth Ormondroyd, Willem H. Ouwehand, Claire Palles, Sofia Papadia, Soo Mi Park, David Parry, Smita Patel, Joan Paterson, Andrew Peacock, Simon H H. Pearce, John Peden, Kathelijne Peerlinck, Christopher J. Penkett, Joanna Pepke-Zaba, Romina Petersen, Clarissa Pilkington, Kenneth E.S. Poole, Radhika Prathalingam, Bethan Psaila, Angela Pyle, Richard Quinton, Shamima Rahman, Stuart Rankin, Anupama Rao, F. Lucy Raymond, Paula J. Rayner-Matthews, Christine Rees, Augusto Rendon, Tara Renton, Christopher J. Rhodes, Andrew S.C. Rice, Sylvia Richardson, Alex Richter, Leema Robert, Irene Roberts, Anthony Rogers, Sarah J. Rose, Robert Ross-Russell, Catherine Roughley, Noemi B. A Roy, Deborah M. Ruddy,

- Omid Sadeghi-Alavijeh, Moin A. Saleem, Nilesh Samani, Crina Samarghitean, Alba Sanchis-Juan, Ravishankar B. Sargur, Robert N. Sarkany, Simon Satchell, Sinisa Savic, John A. Sayer, Genevieve Sayer, Laura Scelsi, Andrew M. Schaefer, Sol Schulman, Richard Scott, Marie Scully, Claire Searle, Werner Seeger, Arjune Sen, W. A. Carrock Sewell, Denis Seyres, Neil Shah, Olga Shamardina, Susan E. Shapiro, Adam C. Shaw, Patrick J. Short, Keith Sibson, Lucy Side, Ilenia Simeoni, Michael A. A. Simpson, Matthew C. Sims, Suthesh Sivapalaratnam, Damian Smedley, Katherine R. Smith, Kenneth G.C. Smith, Katie Snape, Nicole Soranzo, Florent Soubrier, Laura Southgate, Olivera Spasic-Boskovic, Simon Staines, Emily Staples, Hannah Stark, Jonathan Stephens, Charles Steward, Kathleen E. Stirrups, Alex Stuckey, Jay Suntharalingam, Emilia M. Swietlik, Petros Syrris, R. Campbell Tait, Kate Talks, Rhea Y.Y. Tan, Katie Tate, John M. Taylor, Jenny C. Taylor, James E. Thaventhiran, Andreas C. Themistocleous, Ellen Thomas, David Thomas, Moira J. Thomas, Patrick Thomas, Kate Thomson, Adrian J. Thrasher, Glen Threadgold, Chantal Thys, Tobias Tilly, Marc Tischkowitz, Catherine Titterton, John A. Todd, Cheng Hock Toh, Bas Tolhuis, Ian P. Tomlinson, Mark Toshner, Matthew Traylor, Carmen Treacy, Paul Treadaway, Richard Trembath, Salih Tuna, Wojciech Turek, Ernest Turro, Philip Twiss, Tom Vale, Chris Van Geet, Natalie van Zuydam, Maarten Vandekuilen, Anthony M. Vandersteen, Marta Vazquez-Lopez, Julie von Ziegenweidt, Anton Vonk Noordegraaf, Annette Wagner, Quinten Waisfisz, Suellen M. Walker, Neil Walker, Klaudia Walter, James S. Ware, Hugh Watkins, Christopher Watt, Andrew R. Webster, Lucy Wedderburn, Wei Wei, Steven B. Welch, Julie Wessels, Sarah K. Westbury, John Paul Westwood, John Wharton, Deborah Whitehorn, James Whitworth, Andrew O. M. Wilkie, Martin R. Wilkins, Catherine Williamson, Brian T. Wilson, Edwin K.S. Wong, Nicholas Wood, Yvette Wood, Christopher Geoffrey Woods, Emma R R. Woodward, Stephen J. Wort, Austen Worth, Michael Wright, Katherine Yates, Patrick F.K. Yong, Timothy Young, Ping Yu, Patrick Yu-Wai-Man, Eliska Zlamalova, Nathalie Kingston, Neil Walker, John R. Bradley, Sofie Ashford, Christopher J. Penkett, Kathleen Freson, Kathleen E. Stirrups, F. Lucy Raymond, and Willem H. Ouwehand. Whole-genome sequencing of patients with rare diseases in a national health system. *Nature*, 2020. ISSN 14764687. doi: 10.1038/s41586-020-2434-2.
- [61] S. Stabentheiner, M. Danzer, N. Niklas, S. Atzmüller, J. Pröll, C. Hackl, H. Polin, K. Hofer, and C. Gabriel. Overcoming methodical limits of standard RHD genotyping by next-generation sequencing. *Vox Sanguinis*, 2011. ISSN 00429007. doi: 10.1111/j.1423-0410.2010.01444.x.
- [62] Marcia R. Dezan, Ingrid Helena Ribeiro, Valéria B. Oliveira, Juliana B. Vieira, Francisco C. Gomes, Lucas A.M. Franco, Leonardo Varuzza, Roberto Ribeiro, Karen Ziza Chinoca, José Eduardo Levi, José Eduardo Krieger, Alexandre Costa Pereira, Sandra F.M. Gualandro, Vanderson G. Rocha, Alfredo Mendrone-Junior, Ester Cerdeira Sabino, and Carla Luana Dinardo. RHD and RHCE genotyping by next-generation sequencing is an effective strategy to identify molecular variants within sickle cell disease patients. *Blood Cells, Molecules, and Diseases*, 65 (January):8–15, 2017. ISSN 10960961. doi: 10.1016/j.bcmd.2017.03.014. URL <http://dx.doi.org/10.1016/j.bcmd.2017.03.014>.
- [63] Stella T. Chou, Jonathan M. Flanagan, Sunitha Vege, Naomi L.C. Luban, R. Clark Brown, Russell E. Ware, and Connie M. Westhoff. Whole-exome se-

- quencing for RH genotyping and alloimmunization risk in children with sickle cell anemia. *Blood Advances*, 1(18):1414–1422, 2017. ISSN 24739537. doi: 10.1182/bloodadvances.2017007898. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5727856/pdf/advances007898.pdf>.
- [64] Manuel Giollo, Giovanni Minervini, Marta Scalzotto, Emanuela Leonardi, Carlo Ferrari, and Silvio C. E. Tosatto. BOOGIE: Predicting Blood Groups from High Throughput Sequencing Data. *PLOS ONE*, 10(4):e0124579, 4 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0124579. URL <https://dx.plos.org/10.1371/journal.pone.0124579>.
- [65] William J. Lane, Connie M. Westhoff, Jon Michael Uy, Maria Aguad, Robin Smeland-Wagman, Richard M. Kaufman, Heidi L. Rehm, Robert C. Green, and Leslie E. Silberstein. Comprehensive red blood cell and platelet antigen prediction from whole genome sequencing: Proof of principle. *Transfusion*, 56(3):743–754, 2016. ISSN 15372995. doi: 10.1111/trf.13416.
- [66] Mattias Möller, Magnus Jöud, Jill R. Storry, and Martin L. Olsson. ErythroGene: a database for in-depth analysis of the extensive variation in 36 blood group systems in the 1000 Genomes Project. *Blood Advances*, 2016. ISSN 2473-9529. doi: 10.1182/bloodadvances.2016001867.
- [67] Applied Biosystems. Axiom Analysis Suite 3.1 User Guide, 2017. URL <https://www.thermofisher.com/us/en/home/life-science/microarray-analysis/microarray-analysis-instruments-software-services/microarray-analysis-software/axiom-analysis-suite.html>.
- [68] Lourdes Serrano Cardona and Encarnación Muñoz Mata. Paraninfo Digital. *Early Human Development*, 83(1):1–11, 2013. ISSN 03783782. doi: 10.1016/j.earlhumdev.2006.05.022. URL <http://dx.doi.org/10.1016/j.earlhumdev.2015.09.003>
<http://dx.doi.org/10.1016/j.earlhumdev.2014.01.002>
<http://www.sciencedirect.com/science/article/pii/S2341287914000763>
<http://dx.doi.org/10.1016/>
- [69] Alexander Dilthey, Stephen Leslie, Loukas Moutsianas, Judong Shen, Charles Cox, Matthew R. Nelson, and Gil McVean. Multi-Population Classical HLA Type Imputation. *PLoS Computational Biology*, 9(2), 2013. ISSN 1553734X. doi: 10.1371/journal.pcbi.1002877.
- [70] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009. ISSN 13674803. doi: 10.1093/bioinformatics/btp324.
- [71] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, and Richard Durbin. The variant call format and VCFtools. *Bioinformatics*, 2011. ISSN 13674803. doi: 10.1093/bioinformatics/btr330.
- [72] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernysky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. The genome analysis toolkit: A MapReduce framework for

- analyzing next-generation DNA sequencing data. *Genome Research*, 2010. ISSN 10889051. doi: 10.1101/gr.107524.110.
- [73] Xiaoyu Chen, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony J. Cox, Semyon Kruglyak, and Christopher T. Saunders. Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 2016. ISSN 14602059. doi: 10.1093/bioinformatics/btv710.
- [74] Eric Roller, Sergii Ivakhno, Steve Lee, Thomas Royce, and Stephen Tanner. Canvas: Versatile and scalable detection of copy number variants. *Bioinformatics*, 2016. ISSN 14602059. doi: 10.1093/bioinformatics/btw163.
- [75] Applied Biosystems. Array Power Tools (APT).
- [76] NIHR. NIHR BioResource Webpage, 2020. URL <https://bioresource.nihr.ac.uk/>.
- [77] COMPARE. COMPARE Study Webpage. URL <http://www.donorhealth-btru.nihr.ac.uk/studies/compare-study/>.
- [78] Tiffany C. Timmer, Rosa de Groot, Karin Habets, Eva Maria Merz, Femmeke J. Prinsze, Femke Atsma, Wim L.A.M. de Kort, and Katja van den Hurk. Donor InSight: characteristics and representativeness of a Dutch cohort study on blood and plasma donors. *Vox Sanguinis*, 114(2):117–128, 2019. ISSN 14230410. doi: 10.1111/vox.12731.
- [79] Jason L. Vassy, Denise M. Lautenbach, Heather M. McLaughlin, Sek W. Kong, Kurt D. Christensen, Joel Krier, Isaac S. Kohane, Lindsay Z. Feuerman, Jennifer Blumenthal-Barby, J. S. Roberts, Lisa S. Lehmann, Carolyn Y. Ho, Peter A. Ubel, Calum A. MacRae, Christine E. Seidman, Michael F. Murray, Amy L. McGuire, Heidi L. Rehm, Robert C. Green, David W. Bates, Alexis D. Carere, Allison Cirino, Lauren Connor, Jake Duggan, William J. Lane, Christina Liu, Rachel Miller, Cynthia C. Morton, Shamil Sunyaev, Sandy Aronson, Ozge Ceyhan-Birsoy, Siva Gowrisankar, Matthew S. Lebo, Ignat Leschiner, Kalotina Machini, Danielle R. Metterville, Sarita Panchang, Jill Oliver Robinson, Melody J. Slashinski, Stewart C. Alexander, Kelly Davis, Peter Kraft, Judy E. Garber, Tina Hambuch, and In Hee Lee. The MedSeq Project: A randomized trial of integrating whole genome sequencing into clinical medicine. *Trials*, 2014. ISSN 17456215. doi: 10.1186/1745-6215-15-85.
- [80] Luigi Grassi, Osagie G. Izuogu, Natasha A.N. Jorge, Denis Seyres, Mariona Bustamante, Frances Burden, Samantha Farrow, Neda Farahi, Fergal J. Martin, Adam Frankish, Jonathan M. Mudge, Myrto Kostadima, Romina Petersen, John J. Lambourne, Sophia Rowston, Enca Martin-Rendon, Laura Clarke, Kate Downes, Xavier Estivill, Paul Flicek, Joost H.A. Martens, Marie-Laure Yaspo, Hendrik G. Stunnenberg, Willem H. Ouwehand, Fabio Passetti, Ernest Turro, and Mattia Frontini. Cell type specific novel lncRNAs and circRNAs in the BLUEPRINT haematopoietic transcriptomes atlas. *Haematologica*, 2020. ISSN 0390-6078. doi: 10.3324/haematol.2019.238147.
- [81] Carmel Moore, Jennifer Sambrook, Matthew Walker, Zoe Tolkien, Stephen Kaptoge, David Allen, Susan Mehenny, Jonathan Mant, Emanuele D. Angelantonio, Simon G. Thompson, Willem Ouwehand, David J. Roberts, and John Danesh. The INTERVAL

trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: Study protocol for a randomised controlled trial. *Trials*, 15(1):1–11, 2014. ISSN 17456215. doi: 10.1186/1745-6215-15-363.

- [82] Locus Reference Genomic. URL <https://www.lrg-sequence.org/>.
- [83] Hendrik G. Stunnenberg, Martin Hirst, Sergio Abrignani, David Adams, Melanie de Almeida, Lucia Altucci, Viren Amin, Ido Amit, Stylianos E. Antonarakis, Samuel Aparicio, Takahiro Arima, Laura Arrigoni, Rob Arts, Vahid Asnafi, Manel Esteller, Jae-Bum Bae, Kevin Bassler, Stephan Beck, Benjamin Berkman, Bradley E. Bernstein, Mikhail Bilenky, Adrian Bird, Christoph Bock, Bernhard Boehm, Guillaume Bourque, Charles E. Breeze, Benedikt Brors, David Bujold, Oliver Burren, Marion J. Bussemakers, Adam Butterworth, Elias Campo, Enrique Carrillo-de Santa-Pau, Lisa Chadwick, Kui Ming Chan, Wei Chen, Tom H. Cheung, Luca Chiapperino, Nak Hyen Choi, Ho-Ryun Chung, Laura Clarke, Joseph M. Connors, Philippe Cronet, John Danesh, Manolis Dermitzakis, Gerard Drewes, Pawel Durek, Stephanie Dyke, Tomasz Dylag, Connie J. Eaves, Peter Ebert, Roland Eils, Jürgen Eils, Catherine A. Ennis, Tariq Enver, Elise A. Feingold, Bärbel Felder, Anne Ferguson-Smith, Jude Fitzgibbon, Paul Flicek, Roger S.Y. Foo, Peter Fraser, Mattia Frontini, Eileen Furlong, Sitanshu Gakkhar, Nina Gasparoni, Gilles Gasparoni, Daniel H. Geschwind, Petar Glažar, Thomas Graf, Frank Grosveld, Xin-Yuan Guan, Roderic Guigo, Ivo G. Gut, Alf Hamann, Bok-Ghee Han, R. Alan Harris, Simon Heath, Kristian Helin, Jan G. Hengstler, Alireza Heravi-Moussavi, Karl Herrup, Steven Hill, Jason A. Hilton, Benjamin C. Hitz, Bernhard Horsthemke, Ming Hu, Joo-Yeon Hwang, Nancy Y. Ip, Takashi Ito, Biola-Maria Javierre, Sasa Jenko, Thomas Jenuwein, Yann Joly, Steven J.M. Jones, Yae Kanai, Hee Gyung Kang, Aly Karsan, Alexandra K. Kiemer, Song Cheol Kim, Bong-Jo Kim, Hyeon-Hoe Kim, Hiroshi Kimura, Sarah Kinkley, Filippos Klironomos, In-Uk Koh, Myrto Kostadima, Christopher Kressler, Roman Kreuzhuber, Anshul Kundaje, Ralf Küppers, Carolyn Larabell, Paul Lasko, Mark Lathrop, Daniel H.S. Lee, Suman Lee, Hans Lehrach, Elsa Leitão, Thomas Lengauer, Åke Lernmark, R. David Leslie, Gilberto K.K. Leung, Danny Leung, Markus Loeffler, Yussanne Ma, Antonello Mai, Thomas Manke, Eric R. Marcotte, Marco A. Marra, Joost H.A. Martens, Jose Ignacio Martin-Subero, Karen Maschke, Christoph Merten, Aleksandar Milosavljevic, Saverio Minucci, Totai Mitsuyama, Richard A. Moore, Fabian Müller, Andrew J. Mungall, Mihai G. Netea, Karl Nordström, Irene Norstedt, Hiroaki Okae, Vitor Onuchic, Francis Ouellette, Willem Ouwehand, Massimiliano Pagani, Vera Pancaldi, Thomas Pap, Tomi Pastinen, Ronak Patel, Dirk S. Paul, Michael J. Pazin, Pier Giuseppe Pelicci, Anthony G. Phillips, Julia Polansky, Bo Porse, J. Andrew Pospisilik, Shyam Prabhakar, Dena C. Procaccini, Andreas Radbruch, Nikolaus Rajewsky, Vardham Rakyan, Wolf Reik, Bing Ren, David Richardson, Andreas Richter, Daniel Rico, David J. Roberts, Philip Rosenstiel, Mark Rothstein, Abdulrahman Salhab, Hiroyuki Sasaki, John S. Satterlee, Sascha Sauer, Claudia Schacht, Florian Schmidt, Gerd Schmitz, Stefan Schreiber, Christopher Schröder, Dirk Schübeler, Joachim L. Schultze, Ronald P. Schulyer, Marcel Schulz, Martin Seifert, Katsuhiko Shirahige, Reiner Siebert, Thomas Sierocinski, Laura Siminoff, Anupam Sinha, Nicole Soranzo, Salvatore Spicuglia, Mikhail Spivakov, Christian Steidl, J. Seth Strattan, Michael Stratton, Peter Südbek, Hao Sun, Narumi Suzuki, Yutaka Suzuki, Amos Tanay, David Torrents, Frederick L. Tyson, Thomas Ulas, Sebastian Ullrich, Toshikazu Ushijima, Alfonso Valencia, Edo Vellenga, Martin Vingron, Chris Wallace,

- Stefan Wallner, Jörn Walter, Huating Wang, Stephanie Weber, Nina Weiler, Andreas Weller, Andrew Weng, Steven Wilder, Sam M. Wiseman, Angela R. Wu, Zhenguo Wu, Jieyi Xiong, Yasuhiro Yamashita, Xinyi Yang, Desmond Y. Yap, Kevin Y. Yip, Stephen Yip, Jae-II Yoo, Daniel Zerbino, and Gideon Zipprich. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell*, 167(5):1145–1149, 11 2016. ISSN 00928674. doi: 10.1016/j.cell.2016.11.007. URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867416315288>.
- [84] Romina Petersen, John J. Lambourne, Biola M. Javierre, Luigi Grassi, Roman Kreuzhuber, Dace Ruklisa, Isabel M. Rosa, Ana R. Tomé, Heather Elding, Johanna P. Van Geffen, Tao Jiang, Samantha Farrow, Jonathan Cairns, Abeer M. Al-Subaie, Sofie Ashford, Antony Attwood, Joana Batista, Heleen Bouman, Frances Burden, Fizzah A. Choudry, Laura Clarke, Paul Flicek, Stephen F. Garner, Matthias Haimel, Carly Kempster, Vasileios Ladopoulos, An Sofie Lenaerts, Paulina M. Materek, Harriet McKinney, Stuart Meacham, Daniel Mead, Magdolna Nagy, Christopher J. Penkett, Augusto Rendon, Denis Seyres, Benjamin Sun, Salih Tuna, Marie Elise Van Der Weide, Steven W. Wingett, Joost H. Martens, Oliver Stegle, Sylvia Richardson, Ludovic Vallier, David J. Roberts, Kathleen Freson, Lorenz Wernisch, Hendrik G. Stunnenberg, John Danesh, Peter Fraser, Nicole Soranzo, Adam S. Butterworth, Johan W. Heemskerk, Ernest Turro, Mikhail Spivakov, Willem H. Ouwehand, William J. Astle, Kate Downes, Myrto Kostadima, and Mattia Frontini. Platelet function is modified by common sequence variation in megakaryocyte super enhancers. *Nature Communications*, 2017. ISSN 20411723. doi: 10.1038/ncomms16058.
- [85] European Blood Alliance. EBA Annual Report, 2015. URL https://www.europeanbloodalliance.eu/wp-content/uploads/2016/05/EBA_annual_report_2015.pdf.
- [86] Michael F. Murphy, Derwood H. Pamphilon, and Nancy M. Heddle. *Practical Transfusion Medicine*. Wiley, fifth edition, 2013. ISBN 9780470670514. doi: 10.1002/9781118520093.
- [87] Leo van de Watering, Jo Hermans, Marian Witvliet, Michel Versteegh, and Anneke Brand. HLA and RBC immunization after filtered and buffy coat-depleted blood transfusion in cardiac surgery: a randomized controlled trial. *Transfusion*, 43(6): 765–771, 6 2003. ISSN 0041-1132. doi: 10.1046/j.1537-2995.2003.00390.x. URL <http://doi.wiley.com/10.1046/j.1537-2995.2003.00390.x>.
- [88] John M. Higgins and Steven R. Sloan. Stochastic modeling of human RBC alloimmunization: Evidence for a distinct population of immunologic responders. *Blood*, 112(6):2546–2553, 2008. ISSN 00064971. doi: 10.1182/blood-2008-03-146415.
- [89] Scott T. Miller, Hae Young Kim, Debra L. Weiner, Carrie G. Wager, Dianne Gallagher, Lori A. Styles, Carlton D. Dampier, and Susan D. Roseff. Red blood cell alloimmunization in sickle cell disease: Prevalence in 2010. *Transfusion*, 53(4):704–709, 2013. ISSN 00411132. doi: 10.1111/j.1537-2995.2012.03796.x.
- [90] Stella T. Chou, Robert I. Liem, and Alexis A. Thompson. Challenges of alloimmunization in patients with haemoglobinopathies, 2012. ISSN 00071048.

- [91] Stella T. Chou, Tannoa Jackson, Sunitha Vege, Kim Smith-Whitley, David F. Friedman, and Connie M. Westhoff. High prevalence of red blood cell alloimmunization in sickle cell disease despite transfusion from Rh-matched minority donors. *Blood*, 122(6): 1062–1071, 2013. ISSN 00064971. doi: 10.1182/blood-2013-03-490623.
- [92] Stella T. Chou, Perry Evans, Sunitha Vege, Sarita L. Coleman, David F. Friedman, Margaret Keller, and Connie M. Westhoff. RH genotype matching for transfusion support in sickle cell disease. *Blood*, 132(11):1198–1207, 2018. ISSN 15280020. doi: 10.1182/blood-2018-05-851360. URL <http://www.bloodjournal.org/lookup/doi/10.1182/blood-2018-05-851360>.
- [93] Seema Kacker, Paul M. Ness, William J. Savage, Kevin D. Frick, R. Sue Shirey, Karen E. King, and Aaron A.R. Tobian. Cost-effectiveness of prospective red blood cell antigen matching to prevent alloimmunization among sickle cell patients. *Transfusion*, 54(1):86–97, 2014. ISSN 00411132. doi: 10.1111/trf.12250.
- [94] Peter Bugert, Simon McBride, Graham Smith, Alex Dugrillon, Harald Klüter, Willem H. Ouweland, and Paul Metcalfe. Microarray-based genotyping for blood groups: Comparison of gene array and 5-nuclease assay techniques with human platelet antigen as a model. *Transfusion*, 45(5):654–659, 2005. ISSN 00411132. doi: 10.1111/j.1537-2995.2005.04318.x.
- [95] G. A. Denomme and M. Van Oene. High-throughput multiplex single-nucleotide polymorphism analysis for red cell and platelet antigen genotypes. *Transfusion*, 45(5): 660–666, 2005. ISSN 00411132. doi: 10.1111/j.1537-2995.2005.04365.x.
- [96] Michele Lasalle-Williams, Rachelle Nuss, Tuan Le, Laura Cole, Kathy Hassell, James R. Murphy, and Daniel R. Ambruso. Extended red blood cell antigen matching for transfusions in sickle cell disease: A review of a 14-year experience from a single center (CME). *Transfusion*, 51(8):1732–1739, 2011. ISSN 00411132. doi: 10.1111/j.1537-2995.2010.03045.x.
- [97] Ross M. Fasano, Erin K. Meyer, Jane Branscomb, Mia S. White, Robert W. Gibson, and James R. Eckman. Impact of Red Blood Cell Antigen Matching on Alloimmunization and Transfusion Complications in Patients with Sickle Cell Disease: A Systematic Review, 2019. ISSN 15329496.
- [98] Jean Louis H. Kerkhoffs, Jeroen C.J. Eikenboom, Leo M.G. Van De Watering, Rinie J. Van Wordragen-Vlaswinkel, Pierre W. Wijermans, and Anneke Brand. The clinical impact of platelet refractoriness: Correlation with bleeding and survival. *Transfusion*, 48(9):1959–1965, 2008. ISSN 00411132. doi: 10.1111/j.1537-2995.2008.01799.x.
- [99] Y. Jia, W. Li, N. Liu, K. Zhang, Z. Gong, D. Li, L. Wang, D. Wang, Y. Jing, J. Wang, and X. Shan. Prevalence of platelet-specific antibodies and efficacy of crossmatch-compatible platelet transfusions in refractory patients. *Transfusion Medicine*, 24(6): 406–410, 2014. ISSN 13653148. doi: 10.1111/tme.12157.
- [100] William J. Lane, Maria Aguad, Robin Smeland-Wagman, Sunitha Vege, Helen H. Mah, Abigail Joseph, Carrie L. Blout, Tiffany T. Nguyen, Matthew S. Lebo, Manpreet Sidhu, Christine Lomas-Francis, Richard M. Kaufman, Robert C. Green, Connie M. Westhoff, David W. Bates, Carrie Blout, Kurt D. Christensen, Allison L. Cirino, Robert C. Green,

- Carolyn Y. Ho, Joel B. Krier, William J. Lane, Lisa S. Lehmann, Calum A. MacRae, Cynthia C. Morton, Denise L. Perry, Christine E. Seidman, Shamil R. Sunyaev, Jason L. Vassy, Erica Schonman, Tiffany Nguyen, Eleanor Steffens, Wendi Nicole Betting, Samuel J. Aronson, Ozge Ceyhan-Birsoy, Kalotina Machini, Heather M. McLaughlin, Danielle R. Azzariti, Heidi L. Rehm, Ellen A. Tsai, Jennifer Blumenthal-Barby, Lindsay Z. Feuerman, Amy L. McGuire, Kaitlyn Lee, Jill O. Robinson, Melody J. Slashinski, Pamela M. Diamond, Kelly Davis, Peter A. Ubel, Peter Kraft, J. Scott Roberts, Judy E. Garber, Tina Hambuch, Michael F. Murray, Isaac Kohane, and Sek Won Kong. A whole genome approach for discovering the genetic basis of blood group antigens: independent confirmation for P1 and Xga. *Transfusion*, 59(3):908–915, 2019. ISSN 15372995. doi: 10.1111/trf.15089.
- [101] Biola M. Javierre, Sven Sewitz, Jonathan Cairns, Steven W. Wingett, Csilla Várnai, Michiel J. Thiecke, Paula Freire-Pritchett, Mikhail Spivakov, Peter Fraser, Oliver S. Burren, Antony J. Cutler, John A. Todd, Chris Wallace, Steven P. Wilder, Roman Kreuzhuber, Myrto Kostadima, Daniel R. Zerbino, Oliver Stegle, Roman Kreuzhuber, Frances Burden, Samantha Farrow, Karola Rehnström, Kate Downes, Luigi Grassi, Myrto Kostadima, Willem H. Ouwehand, Mattia Frontini, Roman Kreuzhuber, Frances Burden, Samantha Farrow, Karola Rehnström, Kate Downes, Myrto Kostadima, Willem H. Ouwehand, Mattia Frontini, Steven M. Hill, Fan Wang, Hendrik G. Stunnenberg, Willem H. Ouwehand, Mattia Frontini, Willem H. Ouwehand, Joost H. Martens, Bowon Kim, Nilofar Sharifi, Eva M. Janssen-Megens, Marie Laure Yaspo, Matthias Linser, Alexander Kovacsovics, Laura Clarke, David Richardson, Avik Datta, and Paul Flicek. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, 2016. ISSN 10974172. doi: 10.1016/j.cell.2016.09.037.
- [102] William J. Lane, Maria Aguad, Robin Smeland-Wagman, Sunitha Vege, Helen H. Mah, Abigail Joseph, Carrie L. Blout, Tiffany T. Nguyen, Matthew S. Lebo, Manpreet Sidhu, Christine Lomas-Francis, Richard M. Kaufman, Robert C. Green, Connie M. Westhoff, David W. Bates, Carrie Blout, Kurt D. Christensen, Allison L. Cirino, Robert C. Green, Carolyn Y. Ho, Joel B. Krier, William J. Lane, Lisa S. Lehmann, Calum A. MacRae, Cynthia C. Morton, Denise L. Perry, Christine E. Seidman, Shamil R. Sunyaev, Jason L. Vassy, Erica Schonman, Tiffany Nguyen, Eleanor Steffens, Wendi Nicole Betting, Samuel J. Aronson, Ozge Ceyhan-Birsoy, Kalotina Machini, Heather M. McLaughlin, Danielle R. Azzariti, Heidi L. Rehm, Ellen A. Tsai, Jennifer Blumenthal-Barby, Lindsay Z. Feuerman, Amy L. McGuire, Kaitlyn Lee, Jill O. Robinson, Melody J. Slashinski, Pamela M. Diamond, Kelly Davis, Peter A. Ubel, Peter Kraft, J. Scott Roberts, Judy E. Garber, Tina Hambuch, Michael F. Murray, Isaac Kohane, and Sek Won Kong. A whole genome approach for discovering the genetic basis of blood group antigens: independent confirmation for P1 and Xga. *Transfusion*, 59(3):908–915, 2019. ISSN 15372995. doi: 10.1111/trf.15089.
- [103] Ilenia Simeoni, Jonathan C. Stephens, Fengyuan Hu, Sri V.V. Deevi, Karyn Megy, Tadbir K. Bariana, Claire Lentaigne, Sol Schulman, Suthesh Sivapalaratnam, Minka J.A. Vries, Sarah K. Westbury, Daniel Greene, Sofia Papadia, Marie Christine Alessi, Antony P. Attwood, Matthias Ballmaier, Gareth Baynam, Emilse Bermejo, Marta Bertoli, Paul F. Bray, Loredana Bury, Marco Cattaneo, Peter Collins, Louise C. Daugherty, Rémi Favier, Deborah L. French, Bruce Furie, Michael Gattens, Manuela

- Germeshausen, Cedric Ghevaert, Anne C. Goodeve, Jose A. Guerrero, Daniel J. Hampshire, Daniel P. Hart, Johan W.M. Heemskerck, Yvonne M.C. Henskens, Marian Hill, Nancy Hogg, Jennifer D. Jolley, Walter H. Kahr, Anne M. Kelly, Ron Kerr, Myrto Kostadima, Shinji Kunishima, Michele P. Lambert, Ri Liesner, José A. López, Rutendo P. Mapeta, Mary Mathias, Carolyn M. Millar, Amit Nathwani, Marguerite Neerman-Arbez, Alan T. Nurden, Paquita Nurden, Maha Othman, Kathelijne Peerlinck, David J. Perry, Pawan Poudel, Pieter Reitsma, Matthew T. Rondina, Peter A. Smethurst, William Stevenson, Artur Szkotak, Salih Tuna, Christel Van Geet, Deborah Whitehorn, David A. Wilcox, Bin Zhang, Shoshana Revel-Vilk, Paolo Gresele, Daniel B. Bellissimo, Christopher J. Penkett, Michael A. Laffan, Andrew D. Mumford, Augusto Rendon, Keith Gomez, Kathleen Freson, Willem H. Ouwehand, and Ernest Turro. A high-throughput sequencing test for diagnosing inherited bleeding, thrombotic, and platelet disorders. *Blood*, 127(23):2791–2803, 2016. ISSN 15280020. doi: 10.1182/blood-2015-12-688267.
- [104] Franziska Pfeiffer, Carsten Gröber, Michael Blank, Kristian Händler, Marc Beyer, Joachim L. Schultze, and Günter Mayer. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific Reports*, 2018. ISSN 20452322. doi: 10.1038/s41598-018-29325-6.
- [105] Agnieszka Orzińska, Katarzyna Guz, Michał Mikula, Maria Kulecka, Anna Kluska, Aneta Balabas, Monika Pelc-Kłopotowska, Jerzy Ostrowski, and Ewa Brojer. A preliminary evaluation of Next-generation sequencing as a screening tool for targeted genotyping of erythrocyte and platelet antigens in blood donors. *Blood Transfusion*, 2018. ISSN 17232007. doi: 10.2450/2017.0253-16.
- [106] William J. Lane, Nicholas S. Gleadall, Judith Aeschlimann, Sunitha Vege, Alba Sanchis-Juan, Jonathan Stephens, Jensyn Cone Sullivan, Helen H. Mah, Maria Aguard, Robin Smeland-Wagman, Matthew S. Lebo, Prathik K. Vijay Kumar, Richard M. Kaufman, Robert C. Green, Willem H. Ouwehand, and Connie M. Westhoff. Multiple GYPB gene deletions associated with the U phenotype in those of African ancestry. *Transfusion*, 2020. ISSN 15372995. doi: 10.1111/trf.15839.
- [107] World Health Organization. Blood safety and donation, 2008. URL <http://www.who.int/mediacentre/factsheets/fs279/en/index.html>.
- [108] A. Belsito, D. Costa, S. Signoriello, C. Fiorito, I. Tartaglione, M. Casale, S. Perrotta, K. Magnussen, and C. Napoli. Clinical outcome of transfusions with extended red blood cell matching in β -thalassemia patients: A single-center experience. *Transfusion and Apheresis Science*, 58(1):65–71, 2019. ISSN 18781683. doi: 10.1016/j.transci.2018.11.006.

