Bioinformatic studies of genetic variation at known, novel and candidate blood group loci

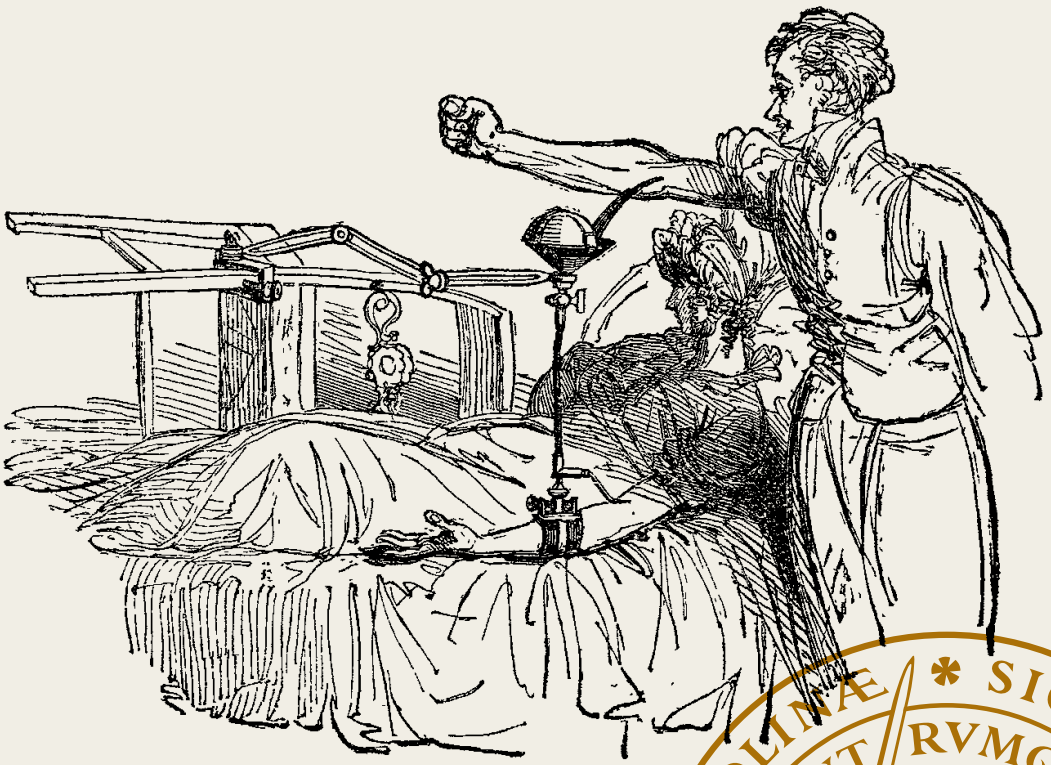
Jöud, Magnus

2018

Document Version:
Publisher's PDF, also known as Version of record

Bioinformatic studies of genetic variation at known, novel and candidate blood group loci

MAGNUS JÖUD
FACULTY OF MEDICINE | LUND UNIVERSITY



Bioinformatic studies of genetic variation at
known, novel and candidate blood group loci

Bioinformatic studies of genetic variation at known, novel and candidate blood group loci

Magnus Jöud

LUNDS
UNIVERSITET

AKADEMISK AVHANDLING

som med vederbörligt tillstånd av Medicinska fakulteten vid Lunds universitet, för avläggande av doktorsexamen i medicinsk vetenskap, kommer att offentligen försvaras i Föreläsningssal 3, Skånes Universitetssjukhus, Lund, måndagen den 11 juni 2018, kl. 13.00.

Fakultetsopponent
Professor Henrik Ullum
Institut for Klinisk Medicin
Københavns Universitet

Organization	Document name
LUND UNIVERSITY	DOCTORAL DISSERTATION
	Date of issue: 2017-05-07
	Sponsoring organization

Author(s)
Magnus Jöud

Title and subtitle
Bioinformatic studies of genetic variation at known, novel and candidate blood group loci

Abstract

Access to compatible blood for transfusion is a prerequisite for modern health care. The compatibility is limited by the presence of antibodies to blood group antigens, polymorphic protein and carbohydrate structures, on the surface of the red blood cell. Blood group antigens arise from genetic variation in the genes underlying their expression. Knowledge of these genes and their variation can facilitate the provision of compatible blood.

The overall aim of the thesis, comprising four papers, was to study the genetic variation at loci underlying human blood group systems and antigens, using bioinformatic methods. In Paper I, the genetic background of the Vel– blood group phenotype was elucidated. In Paper II, the genetic variants regulating the variable expression of the Vel blood group antigen was studied. In Paper III, whole genome sequencing (WGS) data from the 1000 Genomes project were used to create a database of all alleles in known blood group-related genes and to predict the presence of novel blood group antigens. Finally, in Paper IV, human glycosyltransferase genes expressed in erythroid tissue were identified and the potential for candidate carbohydrate-based blood group systems was predicted.

Using SNP array data from Vel-phenotyped blood donors, including members of two families, a 17-base-pair deletion in the previously uncharacterized but evolutionary conserved gene *SMIM1* was found to cause the Vel– blood group phenotype. In Vel+ blood donors from different populations, two polymorphisms in intron 1 of *SMIM1*, rs1175550, and, to a lesser extent, rs143702418, were found to affect the expression of the Vel blood group antigen. In WGS data from the 1000 Genomes project, a large number of previously unreported blood group gene-related alleles were found and compiled into a database, Erythrogene. Among all identified genetic variants, 357 were non-synonymous and predicted to occur on the extracellular portion of blood group-carrying proteins and may represent novel or modified blood group antigens. In the human genome, 244 expressed glycosyltransferase genes were identified, 30 of which were predicted to have properties similar to known genes in carbohydrate-based blood group systems.

The use of bioinformatic methods in the search of genetic variation underlying blood group systems and antigens was successful. The benefits of utilizing publicly available genotyping data in studies of blood groups are highlighted.

Keywords
Bioinformatics, blood groups, glycosyltransferases, transfusion medicine

Classification system and/or index terms (if any)

Supplementary bibliographical information	Language
	English
ISSN and key title: 1652-8220	ISBN: 978-91-7619-658-8
Recipient's notes	Number of pages: 84 / Price
	Security classification

Bioinformatic studies of genetic variation at known, novel and candidate blood group loci

Magnus Jöud

LUND
UNIVERSITY

Front cover: The Gravitator instrument designed by transfusion pioneer James Blundell (1790–1877) for the purpose of transfusing blood between humans. From Blundell J. Observations of transfusion of blood. *Lancet*. 1829;12:321–324.

Back cover: Alignment of exome sequencing reads from the 1000 Genomes project participant HG00128 to *SMIM1* exon 3. The 17-base-pair deletion defining the Vel– blood group phenotype can be seen in heterozygous state as black horizontal lines.

Magnus Jöud, M.D.
Division of Hematology and Transfusion Medicine
Department of Laboratory Medicine
Faculty of Medicine, Lund University
Sweden
E-mail: magnus.joud@med.lu.se

Supervisor: Professor Martin L Olsson, M.D., Ph.D.
Co-supervisor: Dr Ann-Sofie Liedberg, M.D., Ph.D.

ISBN 978-91-7619-658-8
ISSN 1652-8220
Lund University, Faculty of Medicine Doctoral Dissertation Series 2018:92

Printed in Sweden by Media-Tryck, Lund University
Lund 2018

In memory of Hans Uhrnell (1921–2016)

Contents

Abbreviations 11

List of papers 13

Introduction 15
Blood transfusions – a historical background 15
Blood groups 17
Protein blood groups 23
Carbohydrate blood groups 25
Genetic basis of blood groups 27

Aims 33

Methods 35
Data sources 35
Statistical analysis 38
Software implementation 41

Results and discussion 43
Elucidation of the genetic background of Vel (Paper I) 43
Effects of genetic variation in *SMIM1* (Paper II) 46
Genetic variation at blood group loci in the 1000 Genomes project (Paper III) 50
Prediction of candidate carbohydrate blood group loci (Paper IV) 52

General discussion 55

Conclusions 57

Populärvetenskaplig sammanfattning 59
Bakgrund 59
Avhandlingens delarbeten 60

Acknowledgements 63

References 65

Abbreviations

1000G	1000 Genomes project
bp	base pair
CDG	congenital disorders of glycosylation
CDS	coding DNA sequence
ChIP-Seq	chromatin immunoprecipitation followed by massively parallel sequencing
cffDNA	cell-free fetal DNA
EMSA	electrophoretic mobility shift assays
ENCODE	Encyclopedia of DNA Elements
eQTL	expression quantitative trait locus
GT	glycosyltransferase
GWAS	genome-wide association study
HDFN	hemolytic disease of the fetus and newborn
HLA	human leukocyte antigen
HSCT	hematopoietic stem cell transplantation
HTR	hemolytic transfusion reaction
ISBT	International Society of Blood Transfusion
indel	insertion and/or deletion
kb	kilo base pairs
LD	linkage disequilibrium
LE	linkage equilibrium
LRG	Locus Reference Genomic
MAF	minor allele frequency
MALDI-TOF	matrix-assisted laser desorption/ionization time-of-flight
MCHC	mean corpuscular hemoglobin concentration
NGS	next-generation sequencing
PBM	patient blood management
PCR	polymerase chain reaction
PCR-ASP	PCR with allele-specific primers

PCR-RFLP	restriction fragment length polymorphism analysis of PCR-amplified fragments
PCR-SSP	PCR with sequence-specific primers
PLS	passenger lymphocyte syndrome
PRCA	pure red cell aplasia
qPCR	quantitative PCR
RBC	red blood cell
RNA-Seq	RNA sequencing
SMIM1	Small membrane integral protein 1
SNP	single nucleotide polymorphism
SNV	single nucleotide variant
SO	Sequence Ontology
VCF	variant call format
WGS	whole-genome sequencing

List of papers

This thesis is based on the following four publications. They will henceforth be referred to by their roman numerals (I–IV). An asterisk (*) denotes equal first author contribution.

I. Storry JR*, **Jöud M***, Christophersen MK, Thuresson B, Åkerström B, Sojka BN, Nilsson B, Olsson ML. Homozygosity for a null allele of *SMIM1* defines the Vel-negative blood group phenotype. *Nature Genetics*. 2013; 45(5):537–541.

II. Christophersen MK, **Jöud M**, Ajore R, Vege S, Ljungdahl KW, Westhoff CM, Olsson ML, Storry JR, Nilsson B. *SMIM1* variants rs1175550 and rs143702418 independently modulate Vel blood group antigen expression. *Scientific Reports*. 2017;7:40451.

III. Möller M, **Jöud M**, Storry JR, Olsson ML. Erythrogene: a database for in-depth analysis of the extensive variation of 36 blood group systems in the 1000 Genomes Project. *Blood Advances*. 2016;1(3):240–249.

IV. **Jöud M**, Möller M, Olsson ML. Identification of human glycosyltransferase genes expressed in erythroid cells predicts potential carbohydrate blood group loci. *Scientific Reports*. 2018;8:6040.

The following papers, written during the course of the studies, are not included in this thesis.

V. Swaminathan B*, Thorleifsson G*, **Jöud M***, Ali M*, Johnsson E, Ajore R, Sulem P, Halvarsson BM, Eyjolfsson G, Haraldsdottir V, Hultman C, Ingelsson E, Kristinsson SY, Kähler AK, Lenhoff S, Masson G, Mellqvist UH, Månsson R, Nelander S, Olafsson I, Sigurðardottir O, Steingrimsdóttir H, Vangsted A, Vogel U, Waage A, Nahi H, Gudbjartsson DF, Rafnar T, Turesson I, Gullberg U, Stefánsson K, Hansson M, Thorsteinsdóttir U, Nilsson B. Variants in *ELL2* influencing immunoglobulin levels associate with multiple myeloma. *Nature Communications*. 2015;6:7213

VI. Mitchell JS, Li N, Weinhold N, Försti A, Ali M, van Duin M, Thorleifsson G, Johnson DC, Chen B, Halvarsson BM, Gudbjartsson DF, Kuiper R, Stephens OW, Bertsch U, Broderick P, Campo C, Einsele H, Gregory WA, Gullberg U, Henrion M, Hillengass J, Hoffmann P, Jackson GH, Johnsson E, **Jöud M**, Kristinsson SY, Lenhoff S, Lenive O, Mellqvist UH, Migliorini G, Nahi H, Nelander S, Nickel J, Nöthen MM, Rafnar T, Ross FM, da Silva Filho MI, Swaminathan B, Thomsen H, Turesson I, Vangsted A, Vogel U, Waage A, Walker BA, Wihlborg AK, Broyl A, Davies FE, Thorsteinsdottir U, Langer C, Hansson M, Kaiser M, Sonneveld P, Stefansson K, Morgan GJ, Goldschmidt H, Hemminki K, Nilsson B, Houlston RS. Genome-wide association study identifies multiple susceptibility loci for multiple myeloma. *Nature Communications*. 2016;7:12050.

VII. Ajore R, Raiser D, McConkey M, **Jöud M**, Boidol B, Mar B, Saksena G, Weinstock DM, Armstrong S, Ellis SR, Ebert BL, Nilsson B. Deletion of ribosomal protein genes is a common vulnerability in human cancer, particularly in concert with *TP53* mutations. *EMBO Molecular Medicine*. 2017;9(4):498–597.

Introduction

Access to compatible blood for transfusion is a prerequisite for modern health care. Today, blood transfusions are routinely administered in response to severe or symptomatic anemia, for example following major surgery or bone marrow-suppressing cancer treatments. In 2016, 411,097 red blood cell (RBC) units were transfused in Sweden to 83,923 patients, or 41.3 RBC units per 1,000 inhabitants[1]. In nation-wide data from the United States, it was found that in 2004, out of almost 39 million hospital discharges, 2.3 million (5.8%) were associated with blood transfusions[2]. Globally, as many as 112.5 million blood donations were made in 2013, with the highest donation rates in high-income countries[3]. However, blood transfusions are not always uncomplicated, although the safety has improved considerably in the last century. To reduce transfusion rates, there are active efforts in the context of patient blood management (PBM) to optimize patients and procedures prior to transfusion[4].

This thesis studies one of the main obstacles of safe blood transfusions, namely the presence of blood group antigens on the RBC surface[5]. Antibodies to blood group antigens can cause severe hemolytic reactions in the patient upon transfusion with incompatible blood[6]. If these antibodies are directed to antigens expressed by most blood donors, the patients have worse availability of compatible blood and are at risk of limited treatment options or significant treatment delay[7]. The four papers included in the thesis studies the genetic variation underlying the expression of blood group antigens. The results presented can help improve the availability of compatible blood and the matching between patients and blood donors.

Blood transfusions – a historical background

Some of the earliest experiments with blood transfusions were performed in England in the 1660s when blood was transfused between dogs[8]. In contemporary France, the first transfusion to a human (with blood from dog) was recorded[6]. After a number of deaths following transfusion of blood from animal to human,

blood transfusions were banned. Experiments with blood transfusions did not restart until 1818 when dr James Blundell, a London obstetrician and physiologist, observed a woman passing away from post-partum hemorrhage. Blundell realized that a transfusion of blood could well have saved her life:

> *Reflecting afterwards on this melancholy scene, for there were circumstances which gave it a peculiar interest, I could not forbear considering, that the patient might very probably have been saved by transfusion [...]*[9]

From his own experiments with transfusion between dogs, and from human to dog[9], Blundell was convinced that transfusions was only possible between individuals of the same species. The same year (1818), he was able to perform the first recorded blood transfusion between humans[10], and only a few years later the first successful transfusion[11]. In 1829, apart from describing his transfusion apparatus (the Gravitator), Blundell laid the principles of blood transfusions, generally valid as of today:

> *States of the body really requiring the infusion of blood into the veins are probably rare; yet we sometimes meet with cases in which the patient must die unless such operation can be performed; and still more frequently with cases which seem to require a supply of blood, in order to prevent the ill health which usually arises from large losses of the vital fluid, even when they do not prove fatal*[12].

At the time of James Blundell's writing, there was no knowledge on the concept of blood groups other than the observations that blood transfusion sometimes was harmful or even fatal. German physiologist Leonard Landois showed in 1875 that a mixture of RBCs and serum from different species sometimes caused hemolysis[6,13]. Following his work, Austrian Karl Landsteiner could show that this hemolytic reaction could be seen also after transfusion of blood between humans[14,15] and that humans could be divided into one of three groups with differing compatibility amongst them. This was the discovery of the human A, B, and O* blood groups of the ABO blood group system (Table 1). The AB blood group phenotype, where both of A and B are expressed, was recognized by von Decastello and Stürli only a few years later[16]. Knowledge of ABO paved the way for safer transfusions and Karl Landsteiner was awarded the Nobel Prize in Physiology or Medicine 1930 "for his discovery of human blood groups". The rules for

* The O blood group was initially called C but was renamed O for German *ohne*, 'without', 'lacking'. It is a common mistake to write the O with a 0 (zero) character in writing (at least in Swedish) since the system is referred to as "AB0" in spoken language.

transfusion of RBCs and plasma according to Landsteiner's findings are outlined in Figure 1.

Phenotype	RBC antigens	Antibodies in serum	Genotype
O	None	Anti-A and -B	*O*/*O*
A	A	Anti-B	*A*/*A* or *A*/*O*
B	B	Anti-A	*B*/*B* or *B*/*O*
AB	A and B	None	*A*/*B*

Table 1 Simplified overview of antigens, antibodies and genotypes in the ABO blood group system.

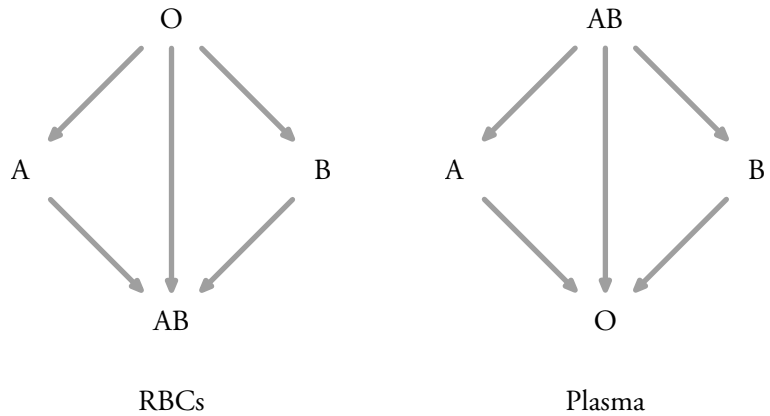


Figure 1 Compatibility of transfusions of RBCs and plasma in the ABO blood group system. According to Landsteiner's rule, an individual produces naturally-occurring antibodies against the antigens lacking on his or her own cells from infancy, explaining the compatibility patterns. Not taking other blood groups systems into account, group O RBCs, lacking A and B, can be transfused to any individual regardless of the presence of antibodies to A or B. Similarly, group AB plasma, without antibodies to A and B, can be transfused to any individual regardless of the presence of A or B on their RBCs.

Blood groups

What Landsteiner had discovered was that the compatibility of blood between individuals is limited by *blood groups*. Blood groups are inherited polymorphisms in proteins and glycans (carbohydrate structures) on the RBC surface that are antigenic (Terminology, p. 20). Depending on their genetic background, blood group antigens are organized into systems[17]. At the time of this writing, 36 blood group systems and 354 antigens* are recognized by the International Society of

* The Fy6 antigen in the Duffy system would not have been approved as a blood group antigen by the current definition since no human anti-Fy6 has been described[18], leaving 353 proper antigens.

Blood Transfusion (ISBT) (Table 2). Blood group antigens of some systems are not confined to RBCs and can also be found on other cell types, sometimes even in soluble form in secretions. These systems are referred to as histo-blood group systems. In some cases blood group antigens function as receptors for pathogens, and absence of an antigen can provide at least partial resistance to infection[19]. Histo-blood group systems can be of importance not only in blood transfusion, but also in solid organ transplantation (Adverse effects of blood group immunization, p. 21).

Table 2a Blood group systems recognized by the ISBT[17].

System name	Number	Symbol	Gene name(s)	Type	Number of antigens
ABO	001	ABO	*ABO*[20,21]	Carbohydrate	4
MNS	002	MNS	*GYPA*, *GYPB*[22,23], (*GYPE*)[a]	Protein	49
P1PK	003	P1PK	*A4GALT*[24–27]	Carbohydrate	3
Rh	004	RH	*RHD*, *RHCE*[28–32]	Protein	55
Lutheran	005	LU	*BCAM*[33]	Protein	24
Kell	006	KEL	*KEL*[34]	Protein	36
Lewis	007	LE	*FUT3*[35]	Carbohydrate	6
Duffy	008	FY	*ACKR1*[36]	Protein	5
Kidd	009	JK	*SLC14A1*[37]	Protein	3
Diego	010	DI	*SLC4A1*[38]	Protein	22
Yt	011	YT	*ACHE*[39,40]	Protein	3
Xg	012	XG	*XG*[41], *CD99*[42]	Protein	2
Scianna	013	SC	*ERMAP*[43]	Protein	7
Dombrock	014	DO	*ART4*[44]	Protein	10
Colton	015	CO	*AQP1*[45]	Protein	4
Landsteiner-Wiener	016	LW	*ICAM4*[46]	Protein	3
Chido/Rodgers	017	CH/RG	*C4A*, *C4B*[47,48]	Protein	9
H	018	H	*FUT1*[49]	Carbohydrate	1
Kx	019	XK	*XK*[50]	Protein	1
Gerbich	020	GE	*GYPC*[51]	Protein	11
Cromer	021	CROM	*CD55*[52]	Protein	19
Knops	022	KN	*CR1*[53–55]	Protein	9
Indian	023	IN	*CD44*[56]	Protein	5

Table 2b Blood group systems recognized by the ISBT[17].

System name	Number	Symbol	Gene name(s)	Type	Number of antigens
Ok	024	OK	*BSG*[57,58]	Protein	3
Raph	025	RAPH	*CD151*[59]	Protein	1
John Milton Hagen	026	JMH	*SEMA7A*[60,61]	Protein	6
I	027	I	*GCNT2*[62]	Carbohydrate	1
Globoside	028	GLOB	*B3GALNT1*[63]	Carbohydrate	2
Gill	029	GIL	*AQP3*[64]	Protein	1
Rh-associated glycoprotein	030	RHAG	*RHAG*[65]	Protein	4
FORS	031	FORS	*GBGT1*[66]	Carbohydrate	1
JR	032	JR	*ABCG2*[67,68]	Protein	1
LAN	033	LAN	*ABCB6*[69]	Protein	1
Vel	034	VEL	*SMIM1*[Paper I,70,71]	Protein	1
CD59	035	CD59	*CD59*[72]	Protein	1
Augustine	036	AUG	*SLC29A1*[73]	Protein	2

[a] The protein product of *GYPE* in the MNS system is not detectable at the RBC surface, but hybrids with *GYPA*[74] and *GYPB*[75] exist.

A central knowledge in blood group serology is that individuals can raise antibodies against the antigens not expressed by themselves, alloantigens*. This occurs in response to allogeneic blood transfusion or other contact with blood, but also after transplantation and in pregnancy[6]. For some blood groups, naturally-occurring antibodies can be seen despite no previous immunizing event. Possibly, the naturally-occurring antibodies are raised against mimicking antigens expressed in the surrounding environment[76]. For the ABO blood group system, the presence of these antibodies is an expected finding for serological determination of ABO phenotypes other than AB. According to Landsteiner's (and others') findings, an individual raises antibodies to the ABO antigens not expressed on his or her own cells from infancy (Table 1, Figure 1). Exceptions to this rule are rarely seen, but can occur in some congenital immunodeficiencies and following hematopoietic stem cell transplantation (HSCT) with an ABO-mismatched donor. In the latter group, the recipient's original ABO phenotype is still present on all cells but those of hematopoietic origin and antibodies to the recipient's own antigens are not expected to be raised.

* From Greek *állos*, 'other', 'different'

Terminology

ISBT* is the international organization that develops and maintains guidelines for blood group antigen and allele nomenclature. While there are many polymorphic proteins and glycans bound to the RBC surface, they are not acknowledged as blood group antigens unless known to fulfill the criteria outlined by the ISBT (Table 3). These criteria are particularly relevant for Paper I, where the Vel blood group system is established, for Paper III, where potentially novel blood group antigens are recognized, and for Paper IV, where the existence of candidate blood group genes is predicted.

Table 3 Summarized ISBT terminology for blood groups[17].

Term	Definition	Number of antigens
Blood group antigen	An inherited RBC antigen defined by a human alloantibody.	354
System	A collection of one or more antigens controlled at a single gene locus, or by two or more very closely linked homologous genes with little or no observable recombination between them.	316 (in 36 systems; Table 2)
Collection (200 series)	Serologically, biochemically, or genetically related antigens, which do not (yet) fit the criteria required for system status.	15 (in 6 collections)
700 series	Antigens of low frequency (less than 1% of individuals positive) that cannot be included in a system or collection.	17
901 series	Antigens of high frequency (more than 90% of individuals positive) that cannot be included in a system or collection.	6

Every blood group antigen is included in one of the systems, collections or low- or high-frequency† series. Most novel antigens are found to belong in known systems and are of low or high frequency[77]. All antigens are given a name and are numbered sequentially, prefixed with the number of the system or series where it is placed.

* Specifically, the Red Cell Immunogenetics and Blood Group Terminology working party of the ISBT.

† The term 'frequency' is used throughout this thesis, despite that the ISBT uses 'incidence' in the context of low- and high-frequency antigens[17]. This is to avoid confusion with the epidemiological meaning of incidence, the probability of occurrence of a given condition within a specified period of time.

Antigens in the blood group collections and low- and high-frequency series are often referred to as *orphan* antigens since they do not yet have a genetic home. Thus, no genetic screening for the expression of these antigens is possible (Genetic basis of blood groups, p. 27). Another category are the *emerging* antigens, not previously acknowledged by the ISBT or even known to be expressed on the RBC surface. An emerging antigen recently described[66] is the FORS blood group antigen, located on the Forssman glycolipid*. The Forssman glycolipid was known to be present in some mammals, but not on the RBCs in humans or primates, and was unexpectedly found to be present on RBCs of some individuals. All or most others, not expressing FORS, have naturally-occurring antibodies to this antigen. These findings and the discovery of the genetic basis of Forssman expression on human RBCs formed the basis of the FORS blood group system (Table 2).

Throughout this thesis, the symbols + and – are used to denote the presence or absence, respectively, of a particular antigen on the surface of RBCs, e.g., Vel+ and Vel– to denote Vel blood group status.

Adverse effects of blood group immunization

Upon transfusion, preformed blood group antibodies may bind the transfused RBCs expressing the corresponding antigen. Depending on the specificity, isotype and titer of the antibody, this can lead to adverse reactions with a decreased RBC survival, or, in some cases, to no reaction at all[6]. The hemolytic transfusion reaction (HTR) is the most severe consequence of preformed antibodies in blood transfusion, leading to the hemolysis of transfused cells, either intravascular by activation of the complement system, or extravascular by the reticuloendothelial system[6]. Sometimes harmful, and at worst even fatal, HTR are prevented by the detection of blood group antibodies in patient serum prior to transfusion and subsequent selection of antigen-negative RBC units. Due to this fundamental principle in clinical transfusion medicine, transfusion can nowadays generally be considered safe from an immunologic perspective. Consequently, only three acute and eight delayed HTR were reported in Sweden 2016[5]. While many blood group systems and antigens exist (Table 2), blood units to patients are not routinely matched for clinically relevant blood groups beyond ABO and the D antigen of the Rh system†. This can result in formation of antibodies to mismatched blood group antigens which will be detected before the next transfusion

* Named after its discoverer, professor John Forssman (1868-1947), coincidentally a professor at Lund University and director of the Lund hospital 1927–1930.

† There are local exceptions. For example, children and women of childbearing age are given K– (KEL:-1) blood units in Region Skåne and elsewhere, to prevent immunization[78].

or even give rise to delayed hemolysis of the transfused RBCs. Better matching is desirable but often not feasible[79].

The presence of antibodies to blood group antigens can also complicate pregnancies, when the mother is immunized to blood group antigens expressed by the fetus's RBCs (inherited from the father). Antibodies can be raised during pregnancy when the mother's immune system recognizes blood group antigens expressed on fetal RBCs[80]. The immunization occurs when fetal cells pass into the mother's circulation by fetomaternal hemorrhage. Unlike the larger IgM isotype antibodies, IgG antibodies produced can pass the placental barrier and cause hemolytic disease of the fetus and newborn (HDFN)[81]. IgG antibodies are in fact actively transported across the placental barrier[82]. Upon the detection of antibodies, fetuses are monitored for signs of anemia and newborns for jaundice. When signs of significant anemia are seen in the fetus, intrauterine transfusions are occasionally necessary[80]. In Sweden, intrauterine transfusions were required by 14–21 patients yearly between 2013–2015[83]. All pregnant women in Sweden are screened at least once during pregnancy for presence of blood group antibodies. The D antigen of the Rh blood group system is a major cause of HDFN in D– mothers. However, immunization to D can be prevented by the administration of anti-D immunoglobulin (Rh-prophylaxis) to D– mothers mainly postnatally[84], but also antenatally[85] or following potentially immunizing procedure. Since the D– phenotype is due to homozygosity for deleted *RHD*[86] (in the Western population at least[18]), the presence of *RHD* from cell-free fetal DNA (cffDNA)[87] in maternal plasma can be screened for[88,89]. In Region Skåne and elsewhere[89–91], D– mothers are screened for the presence of fetal *RHD* in maternal plasma to avoid administration of immune prophylaxis unless it is necessary[89]. For antibodies other than anti-D, the major risk factor for immunization to blood group antigens is previous blood transfusion, rather than parity[92].

Blood groups are important also in solid organ transplantation and HSCT. Donor-recipient mismatches can cause rejection of the transplanted organs but transplantation is possible with careful planning[93–96]. Besides the risk of graft rejection, an lesser known adverse event in mismatched transplantation is the passenger lymphocyte syndrome (PLS)[97]. In PLS, donor lymphocytes remaining in the graft causes hemolysis in the recipient due to the continued production of recipient-specific blood group antibodies[98]. Additionally, in HSCT, recipient- or donor-derived blood group antibodies can cause hemolysis of RBCs[96]. While antibodies to donor blood group antigens are normally cleared within days or weeks after transplantation[99], the persistence of antibodies can cause pure red cell aplasia (PRCA). PRCA is not common, but can require transfusion support for

months or years after transplantation[96]. The most important blood group system in transplantation is ABO, where the same compatibility rules as for RBCs apply (Figure 1). Rejection of transplanted kidneys have been ascribed to antibodies in the Kidd and other systems[100,101] indicating that these are also of importance. PLS has also been reported for systems other than ABO, including Rh, Duffy and Kidd[98].

Protein blood groups

Protein-based blood group systems are the most abundant and encompass the highest number of antigens (Table 2, p. 18). The proteins harboring blood group antigens are anchored in the RBC membrane, most of them endogenously expressed in erythroid precursors. The complement component C4 (available in two isoforms, C4A and C4B) is, however, synthesized primarily in the liver and only adsorbed and covalently bound onto RBCs from plasma. C4A and C4B carries the Rodgers and Chido antigens, respectively, of the Chido/Rodgers blood group system.

Many blood group-related proteins have important functions (Table 4), and some null phenotypes have clinical consequences.

Table 4 Function of proteins encoded by blood group genes[76].

Category	Blood group systems
Membrane transporters	Augustine, Colton, Diego, Gill, JR, Kidd, Kx, LAN, Rh, Rh-associated glycoprotein
Receptors and adhesion molecules	Duffy, Indian, JMH, LW, Lutheran, Ok, Raph, Scianna, Xg[a]
Complement regulatory glycoproteins	CD59, Chido/Rodgers, Cromer, Knops
Enzymes	Kell, Dombrock, Yt
Structural components	Diego, Gerbich, Lutheran, Kx, Rh-associated glycoprotein
Components of the glycocalyx/unknown	MNS, Vel

[a] Based on its homology to CD99.

The most common null phenotype seen, D–, caused by homozygosity for deletion of *RHD*, is without consequence for the carrier, most likely because of similar functionality of the homologous RhCE protein. The very rare Rh_{null} phenotype, caused by a simultaneously inactivated *RHCE* or inactivating variation in *RHAG*, has membrane abnormalities in the RBC and a usually mild and compensated hemolytic anemia[76]. Individuals with a null phenotype in the Kidd blood group

system have no clinical symptoms, but a reduced urine concentrating ability[102]. This seemingly surprising finding is explained by the fact that the protein carrying the Kidd antigens, the UT-B urea transporter, is expressed and functions also in the kidney[18]. Other negative clinical phenotypes are associated with the Diego[103], Kx[104] and Raph[105] blood group systems. There are also positive consequences of null phenotypes. Low or no expression of antigens in the MNS, Duffy, Diego, Gerbich, Knops, and Ok blood group systems all provide some degree of protection from malaria infection[18,19,106].

The Vel blood group system

Vel is a a high-frequency antigen according to the ISBT definition, with most people expressing the antigen. This was recognized already in the first publication of Vel in 1952 where a patient sample was reactive with all except four out of 10,000 samples tested[107]. Since its first description, anti-Vel has been known as a clinically significant antibody, capable of causing severe hemolysis upon transfusion of an immunized individual with Vel+ RBCs[108,109]. Combined surveys from Western countries found that only 95 out of 251,170 samples (or about 1 in 2,644) were Vel–[108]. Intriguingly, 52 of the 95 Vel– were found in Umeå, in a sample of 91,605 blood donors[110]. Hence, the frequency of Vel– was much higher in Umeå (1 in 1762)* as compared to other Western countries (1 in 3,711). Studies in other populations have indicated an increased frequency of Vel– in Thais (four in 328 samples)[111] and Chilcotin Indians in Canada (two in 160 samples)[112]. However, the former frequency was recently disputed in a study of 223 Thai blood donors[113] where no Vel– individuals were found. Instead, almost all of the Thai donors had a relatively weak expression of Vel, suggesting that the Thai Vel– samples found in the 1960s were in fact Vel+ or Vel+^{weak}. The variable expression is a known feature of Vel[108], which has lead to mistyping of Vel+^{weak} donors as being Vel–†.

Prior to the publication of Paper I and papers by other groups[70,71], the genetics underlying Vel expression were not known. The inheritance pattern of the Vel– trait, however, suggested an autosomal recessive inheritance[108,110], which was confirmed by the findings in Paper I (p. 43). Associations with other blood group phenotypes have been reported. Studies in Umeå demonstrated a higher frequency of the P_2 blood group phenotype in Vel– blood donors[110]. Likely a

* Umeå has since been a major global supplier of Vel– blood.

† This was actually the case in the study by Cvejic et al.[70], where one of five presumed Vel– was in fact Vel+^{weak}. Despite this, the authors managed to identify the 17-base-pair deletion in *SMIM1* (Paper I, p. 43), by yet undisclosed methods.

false positive finding, this was later disproved[109]. Anti-Vel has been reported not to react with cells lacking antigens in the Gerbich system[108]. Recent flow cytometric studies of the expression of Vel and glycophorin C (the carrier of Gerbich antigens) showed, however, that there was no difference in glycophorin C expression between Vel– and Vel+ cells[114]. Lastly, it was shown that antibodies to the ABTI high-frequency antigen reacted weakly or, in one case, not at all with Vel– RBCs[108]. These results placed Vel and ABTI in the same blood group collection due to the presumed serological relationship. When the genetic background of Vel was defined in Paper I, the genetic connection to ABTI could not be confirmed[115]. Vel was promoted to its own blood group system while ABTI remains an orphan antigen.

Many examples of anti-Vel are documented[108]. Anti-Vel is commonly a mixture of IgM and IgG (IgG1 and IgG3[108]), reacts weaker with cord blood cells than adult cells, and has a wide thermal range[116]. These are all features of antibodies to carbohydrate antigens, which could suggest that the Vel antigen is also carried on a glycan. Unlike other carbohydrate blood group antigens, however, no examples of naturally-occurring anti-Vel are seen but only ever produced in response to contact with the antigen upon transfusion or pregnancy[108]. While strongly hemolytic, Vel is not usually causing HDFN, possibly due to the weak expression on neonatal cells and the predominance of IgM-anti-Vel[76,109] although examples of even severe HDFN have been reported[108,117]. Interestingly, one abstract reported that 5 out of 6 individuals with anti-Vel were carriers of the *HLA-DRB1*11* group of alleles[118], indicating that only a subset of Vel– individuals form anti-Vel.

Carbohydrate blood groups

Carbohydrate blood group antigens (Table 2) reside on glycans present on glycoproteins and glycosphingolipids in the RBCs membrane. These glycans are part of the glycocalyx surrounding all cells, contributing to recognition, communication and intercellular adhesion[119]. The blood group antigens are some of the terminal structures of the glycans, available for immune recognition. The function of the blood group antigens is unknown but the *ABO* locus at 9q34.2 is a surprisingly common finding in genome-wide association studies (GWAS). Currently, there are results at the *ABO* locus for 92 traits (The NHGRI-EBI Catalog of published genome-wide association studies; cited 2018-04-24). Of importance, the blood group O phenotype provides protection from severe infection with *Plasmodium falciparum* malaria by means of reduced rosetting[120], at least partially explaining

the high prevalence of blood group O in *P. falciparum*-endemic areas[121]. Rosetting is a microscopic phenomenon seen when uninfected RBCs are bound to a central infected cell, forming a flower shape, and is associated with severe disease[122]. A and B blood group antigens on the RBC surface has been shown to function as co-receptors in the formation of rosettes[123]. However, the reason for this interaction was unknown until recently when it was shown that the RIFINs family of *P. falciparum* proteins bind preferentially to the A antigen[124]. RIFINs (repetitive interspersed families of polypeptides) is a protein family of which some are expressed on the surface of infected RBCs[122].

Most carbohydrate antigens are synthesized endogenously in RBCs, but they can also be adsorbed onto the membrane from glycosphingolipids in plasma[76]. This is seen for the Lewis blood group antigens, not synthesized in erythroid cells. The A and B antigens in the ABO system are found as soluble antigens in plasma of blood group A, B and AB individuals if they are *secretors*. The secretor status is defined by a functional *FUT2* gene, necessary for the synthesis of secreted, ABH-active glycans[76]. Soluble A antigens in the plasma from the transfusion recipient can be adsorbed onto the transfused RBCs[125]. Similarly, the RBCs produced following a HSCT procedure, can adsorb antigens onto the newly-produced RBCs[126]. The same appears to be true for B antigens[127].

Glycosyltransferases and glycosylation

Glycans are synthesized by the glycosyltransferase (GT) family of enzymes (enzyme commission [EC] 2.4]). GTs catalyze the transfer of sugar moieties from activated donor molecules to acceptor substrates, forming the glycosidic bond[128]. The acceptor substrate can be of various types including other sugars, lipids, proteins, nucleic acids or other small molecules. GTs are classified into families based on amino acid similarity, available in the Carbohydrate-Active enZYmes Database (CAZy)[129]. Currently, 105 families are classified in CAZy, but not all of these are found in humans[130]. Despite the large number of families, merely three protein topologies or folds have been described for GTs: GT-A, GT-B, and GT-C, where GT-A and GT-B are the predominant folds[131]. The domain configuration and structure is similar within each fold. There is an almost linear correlation between the number of GT genes and the total number of genes in sequenced genomes, with GTs accounting for 1–2% of all genes[128].

The action of GTs creates a wealth of glycans present on glycoproteins, glycosphingolipids and proteoglycans. The large variety is due to the many combinations of sugar moieties and possible linkages. There are two types of glycosidic bonds, α and β, and a number of possible carbon links, referred to as 1–2, 1–4

etc. Furthermore, hydroxyl groups of sugar moieties can be subject to sulphation or phosphorylation[119]. Glycans are synthesized by the sequential action of several GTs and a dysfunction of a proximally acting GT abolishes the synthesis of the remaining glycan.

There are two main types of bindings of glycans to proteins, N- and O- linked. The N-linked is the most common, with glycans binding to asparagine residues in the Asn-X-Ser/Thr sequence*. For the O-linked binding, glycans bind to serine or threonine[119] or more rarely to tyrosine[132]. It has been estimated that more than 50% of proteins are glycosylated as a post-translational modification while in the Golgi apparatus. The presence of glycans is often crucial for the correct protein function[133]. Erythopoietin[134] and immunoglobulins[135] are two examples of proteins depending on their glycans for correct function. The removal of glycans on immunoglobulins is currently explored as a therapeutic possibility in antibody-mediated diseases, by the action of the EndoS enzyme of streptococcal origin[136–138]. The dysfunction of many GTs, in addition to other enzymes active in glycan biosynthesis and metabolism, cause syndromes in a heterogeneous group of rare disorders referred to as congenital disorders of glycosylation (CDG)[139]. The CDG group of disorders have a variety of phenotypes, reflecting the ubiquitous distribution and function of glycans. However, the dysfunction of GTs can be completely benign, as shown by an absence of disease phenotypes in individuals with inactivating variation in blood-group-related GT genes. On the other hand, there is a notable exception in the P1PK/GLOB blood group systems. Individuals with the p, P_1^k and P_2^k phenotypes, all lacking the P antigen (globoside) due to a deficiency of *B3GALNT1*, are healthy but have a high rate of spontaneous abortions[140]. This is due to the presence of naturally-occurring IgM and IgG3 antibodies to the P antigen, highly expressed in the placenta[141].

Genetic basis of blood groups

Throughout this thesis, the term single nucleotide polymorphism (SNP) is used to denote single nucleotide variants regardless of population frequency, for practical reasons. In other contexts, the term SNP is sometimes reserved for variants with a population minor allele frequency (MAF) greater than 1%, and single nucleotide variant (SNV) used otherwise. The term indel is used to denote insertion and/or deletion.

* X denotes any amino acid but proline.

The genetic variation underlying the polymorphic blood groups are no different from those causing other traits, and all types of variants are seen including SNPs, indels and larger structural variations[17,18]. Structural variations are prevalent as deletions, copy number variations and in the formation of hybrid genes, within, for example, the Rh, Chido/Rodgers and MNS blood group systems, respectively. For protein-based blood group systems, the absence or presence of an antigen is mostly determined by a SNP, often resulting in an *antithetical* pair of antigens. For example, the SNP rs1058396 has two alleles, G and A, where the former gives rise to the Jk^a and the latter to the Jk^b antigen in the Kidd blood group system. A heterozygous individual will express both antigens, and a homozygote either of them. The antithetical relation can also be found between a low- and a high-frequency antigen. For some protein antigens, an antithetical antigen have not been described. Homozygosity for null alleles in protein-based blood group genes can be problematic from a clinical perspective since all antigens on the protein are lacking. As polyclonal antibodies can be raised to all possible epitopes on the protein, transfusion of blood from a blood donor with the same (null) phenotype is necessary. This highlights the importance of having methods and reagents to screen for rare blood donors.

Carbohydrate blood group antigens are synthesized by GTs and deleterious variants in the underlying genes result in a null phenotype and absence of the antigen. Genetic variation in GTs can also cause an altered specificity of the enzyme, thereby changing its donor or acceptor substrate preference. This is seen in the ABO blood group system, where only a few SNPs determine the donor substrate specificity and the resulting A or B phenotype[20]. Alternatively, a variant can change the activity of the enzyme while maintaining its specificity. This can for instance be observed in inherited subgroups of ABO and is often manifested phenotypically as an altered, often weakened, expression pattern[142].

Genetic variation in a regulatory region has also been recognized as an important modifier of blood group antigen expression. The expression of blood group genes is largely determined by the binding of transcription factors to regulatory regions[143,144]. A variant disrupting binding sites of transcription factors can cause an abrogated expression. This was first observed for a null allele of *ACKR1* in the Duffy blood group system, where a SNP in the promoter, rs2814778, disrupts a GATA1 binding site and causes the Fy(a–b–) phenotype[145]. Interestingly, since GATA1 is an almost erythroid-specific transcription factor, the Duffy glycoprotein can be found on other cell types. The Fy(a–b–) phenotype is highly prevalent in parts of West Africa since it provides resistance to *P. vivax* and *P. knowlesi* malaria[19] since these parasites use the Duffy glycoprotein as an invasion

receptor[146,147]. By similar mechanisms, disrupted transcription binding sites have since also explained the B_m[148], A_m[149] and P_2[26,27] phenotypes.

Applications of blood group genotyping

Knowledge of the genetic basis of blood groups makes genotyping to predict the blood group status in an individual possible. There are four principally different situations where genotyping is necessary or preferred to serological typing[76]:

1. there are no RBCs available for serological typing
2. there is no serological reagent available
3. genotyping will provide more or better information than serological typing
4. genotyping is more efficient and/or more cost effective than serological typing

The clinical importance of studying the genetic background of blood groups is demonstrated firstly by the possibility of finding compatible blood[79,150]. While it is challenging to find compatible blood to patients immunized to a high-frequency* antigen[7], knowledge of the underlying genetic background enables large-scale genetic screening for compatible blood donors. Without this knowledge, screening must be undertaken using serological methods, provided that reagents are available. The Vel blood group antigen was difficult to type for until the publication of Paper I, since antisera to Vel were scarce and only collected from immunized individuals – no commercial serological reagent was readily available.

In multi-transfused patients it is close to impossible to determine blood group phenotypes with serological methods since donors' RBCs are present. Genotyping is, however, possible in this situation (RBCs are enucleated)[151]. To the benefit of chronically transfused patients, an improved matching between recipient and donor will also extend the survival of the transfused cells and decrease the risk of alloimmunization[152,153].

Genotyping is necessary for the detection of *RHD* from cffDNA in maternal plasma (Adverse effects of blood group immunization, p. 21). Protocols for prediction of other antigens by testing cffDNA have been developed, including the K, C, c, and E antigens[154,155].

Another important application of genetic blood group determination is quality assurance of reagent RBCs used in antibody detection assays in the clinical laboratory[156]. Genetic determination of blood group gene alleles can, e.g., assure the homozygosity of screening cells for clinically important blood group antigens

* It is not difficult to find a blood donor lacking a low-frequency antigen since most blood donors will not express the antigen on their RBCs.

to maximize the ability to detect antibodies. In Sweden, genotyping of screening cells is recommended for blood group antigens in the Rh, Kell, Duffy, Kidd and MNS systems[157].

Methods used for blood group genotyping

Methods for blood group genotyping and phenotype prediction were developed soon after the cloning of blood group genes. Already in the paper defining the *A* and *B* alleles, the authors found that the differing polymorphisms could be cleaved with restriction enzymes and used for diagnostic purposes[20]. ABO was also the most studied blood group system initially, given its clinical importance[158,159]. Soon thereafter, it was recognized that the *ABO* locus was more diverse than previously expected[160], necessitating an expansion of the original protocols. Based on the polymerase chain reaction (PCR) and variants such as restriction fragment length polymorphism analysis of PCR-amplified fragments (PCR-RFLP), PCR with allele-specific primers (PCR-ASP) and PCR with sequence-specific primers (PCR-SSP), some of these early protocols are still in use for genotyping in ABO and other blood group systems[161,162]. Quantitative PCR (qPCR) is also used, with applications in fetal *RHD* screening in maternal plasma[163] among others. Sanger sequencing is an option and serves as the gold standard for genotyping.

While specific, simple PCR-based methods have comparatively low throughput. They are also limited in that they can only find what they were developed to detect. Expecting an increased usage of blood group genotyping, even for routine purposes[79], efforts were made to develop methods with higher throughput and that also included several blood group systems in the same analysis[164–167]. The EU-funded BloodGen project, a collaboration between the transfusion medicine community and the industry, had the goal of developing an array solution for blood group genotyping purposes[150,168]. The project succeeded and resulted in the commercialization of a blood group genotyping array[169]. BloodGen was not, however, the only effort in this area and methods based on multiplex PCR and other array formats were developed and made available[150,170,171]. In recent years, truly high-throughput methods based on qPCR[172], Luminex[173,174] and matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF)[113,175,176] have been introduced, with a capacity of genotyping hundreds, if not thousands, of samples a day.

The next natural step in blood group genotyping is the application of next-generation sequencing (NGS)* methods[153,177–180]. NGS enables an unbiased sequencing of all relevant genes in parallel, for many samples at a time. Pilot studies exploring the use of NGS in blood group genotyping have been published[181–185]. However, many of these efforts have focused on single systems (mainly Rh) or phenotypes defined by SNPs, excluding indels and structural variation. The exception is a study by Lane et al.[183], that comprehensively predicted blood group and platelet phenotypes in a single individual. The authors, however, had to develop custom software workflows and faced the absence of well-defined databases for annotations of blood group genotyping data†. Exome sequencing approaches have been tested for unresolved serological typings[187] but exome sequencing is likely insufficient as a clinical method since crucial polymorphisms are located in introns[17]. The excellent dynamic range of NGS also has applications in antenatal predictions of fetal genotypes[155].

* The term next-generation sequencing is used throughout this thesis, although it is somewhat misleading. Many of the methods covered by this term are not the next, but rather the current generation of sequencing methods. The term massively parallel sequencing (MPS) has sometimes been used as an alternative with the benefit of being descriptive.

† A previous database, BGMUT[186], was incomplete and has been discontinued.

Aims

The overall aim of this thesis was to study the genetic variation at known, novel and candidate blood group loci using bioinformatic methods. The specific aims were as follows:

Paper I	To elucidate the genetic background of the Vel– blood group phenotype and thereby establish a new blood group system
Paper II	To investigate the effects of genetic variants in *SMIM1* on the expression of the Vel blood group antigen
Paper III	To determine and tabulate the alleles of known blood group-related genes in 1000G and to study the potential impact of previously unrecognized genetic variants located in these genes
Paper IV	To identify human glycosyltransferase genes expressed in erythroid tissue and assess their potential as candidate blood group loci

Methods

Data sources

Donor and patient samples

Anonymized buffy coat waste materials from routine blood donations and peripheral blood samples from donors and immunized patients were obtained from several blood centers in Sweden, Europe and the United States. No individuals were approached solely for the purposes of these studies.

Patient and donor samples were used in Papers I and II.

The 1000 Genomes project

The 1000 Genomes project (1000G) was an international study where whole-genome sequencing (WGS) was performed in 2,504 individuals from different populations. The goal of 1000G was to find most genetic variants with frequencies of at least 1% in five major population groups[188–190] (Table 5). The preceding HapMap project had identified a then impressive 3.1 million common variants, and created linkage disequilibrium maps on a genome-wide scale[191]. Still, there was an insufficient number of low-frequency variants known to study complex genetic disease and genotype-phenotype relations. At the same time, massively parallel sequencing technologies had become increasingly available. This enabled WGS studies to be performed on an (at the time) unparalleled scale.

The project was carried out using a number of sequencing technologies at several centers and a combination of exome sequencing and low-coverage (~1–4×) WGS was used. In total, 84.7 million single nucleotide variants, 3.6 million smaller indels and 60,000 structural variants were discovered[190] in the 26 populations sequenced. While recent studies of single populations, with sometimes even greater sample sizes (such as the UK10K[192], Genome of the Netherlands[193], Swedish SweFreq[194] and the upcoming 100,000 Genomes Project[195] studies), capture even more rare variation, the inclusion of different populations has proven valuable. All data from 1000G are publicly available without registration, unlike

Table 5 Description of populations included in the final phase 3 of the 1000 Genomes project.

Abbreviation	Description	Super population	Sample size
AFR	African	–	669
ACB	African Caribbean in Barbados	AFR	96
ASW	African Ancestry in Southwest USA	AFR	66
ESN	Esan in Nigeria	AFR	99
GWD	Gambian in Western Division, the Gambia	AFR	113
LWK	Luhya in Webuye, Kenya	AFR	101
MSL	Mende in Sierra Leone	AFR	85
YRI	Yoruba in Ibadan, Nigeria	AFR	109
AMR	American	–	352
CLM	Colombian in Medellín, Colombia	AMR	94
MXL	Mexican Ancestry in Los Angeles, California	AMR	67
PEL	Peruvian in Lima, Peru	AMR	86
PUR	Puerto Rican in Puerto Rico	AMR	105
EAS	East Asian	–	515
CDX	Chinese Dai in Xishuangbanna, China	EAS	99
CHB	Han Chinese in Beijing, China	EAS	103
CHS	Southern Han Chinese, China	EAS	108
JPT	Japanese in Tokyo, Japan	EAS	104
KHV	Kinh in Ho Chi Minh City, Vietnam	EAS	101
EUR	European	–	505
CEU	Utah residents with Northern and Western European ancestry	EUR	99
FIN	Finnish in Finland	EUR	99
GBR	British in England and Scotland	EUR	92
IBS	Iberian populations in Spain	EUR	107
TSI	Toscani in Italy	EUR	108
SAS	South Asian	–	494
BEB	Bengali in Bangladesh	SAS	86
GIH	Gujarati Indian in Houston, Texas	SAS	106
ITU	Indian Telugu in the United Kingdom	SAS	103
PJL	Punjabi in Lahore, Pakistan	SAS	96
STU	Sri Lankan Tamil in the United Kingdom	SAS	103

other studies where commonly only summarized variant frequencies are accessible without registration and application procedures.

Another aspect of 1000G – commonly forgotten – is the bioinformatic spin-off effect. Many of today's most used bioinformatic tools and file formats, such as the BWA sequence aligner[196], SAMtools sequence processing toolkit[197] and VCF (variant call format) file format[198], were developed for the purposes of 1000G and are still standard tools in genomic processing pipelines and studies. The programs benefit from being free and open-sourced, which enables collaborative development and in the end faster, better tested and more useful tools.

Data from 1000G were used in Paper I as a reference for allele frequency comparisons, in Paper II for determining haplotypes in the European and African populations, in Paper III as the basis for the *Erythrogene* database of blood group gene alleles, and in Paper IV as a source of genetic variation in GT genes.

The Encyclopedia of DNA Elements

The goal of the ongoing Encyclopedia of DNA Elements (ENCODE) project is to build a comprehensive list of functional elements in the human genome. This includes elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active. ENCODE started after the completion of the human genomes sequencing projects[199,200] to explain the function of most genomic elements, not only the coding regions. Consequently, ENCODE examined the genome from different perspectives[201]:

- transcribed and protein-coding regions
- RNA expression
- protein bound-regions, including ChIP-Seq (chromatin immunoprecipitation followed by massively parallel sequencing)
- DNase I hypersensitive sites and footprints
- regions of histone modification
- DNA methylation
- chromosome-interacting regions.

In summary, 80.4% of the human genome was covered with at least one of the elements detected in ENCODE, a massive increase in the understanding of the human genome[201]. In comparison, only 1.22% of the genome was found to be coding exons.

Several laboratories participated in ENCODE, all sharing the same cell lines, protocols and data processing pipelines. The experiments were performed in 147 different cell lines, divided by the number of experiments performed into three

groups (tiers)[201]. The group with the most experiments (tier 1) consisted of the lymphoblastoid cell line GM12878, the H1 embryonic stem cell (H1 hESC) and the erythroleukemic cell line K562[201]. The inclusion of the K562 cell line was of importance due to its erythroleukemic phenotype, generally extrapolated to erythroid cells. To the benefit of RBC-related research, the ChIP-Seq experiments included the GATA1 transcription factor[202], a major regulator of gene expression in erythroid cells[143,144].

Data from ENCODE were used in Papers I, II and IV, in particular the GATA1 ChIP-Seq data generated in K562 cells.

Other databases

A number of other databases were used, including, but not limited to, UCSC Genome Browser[203], Ensembl[204], ExAC[205] (most of which was later merged into the larger gnomAD), KEGG GLYCAN[206], UniProt[207] and CAZy[129]. Many of these databases cross-reference each other and the choice of source was often driven by the ease of their programming interfaces.

Statistical analysis

Genetic association analysis

The purpose of genetic association analysis is to establish an association between a marker and a trait[208]. In the era of genomics, this is typically employed on a whole-genome scale in GWAS with unrelated individuals, but local association studies are possible. In a GWAS, subjects are genotyped for genetic markers across the whole genome with SNP arrays or sequencing. For SNP arrays, the most common method used, the markers are selected to tag haplotype blocks with many SNPs in high linkage disequilibrium (see below). The subjects can be cases and controls (for binary traits) or consist of a cohort where quantitative variables, such as biomarkers, have been measured. The allele frequencies are compared between groups using statistical tests, typically contingency table tests (Pearson χ^2, Fisher's exact or Cochran-Armitage trend tests) or logistic regression, that also allow for the inclusion of covariates[208]. Linear regression is used for continuous variables. Due to the very high number of tests being performed – typically in the millions – the risk of false positive findings (type I errors) is very high, and the genome-wide significance level is often set to $p < 5 \times 10^{-8}$. This value is derived from Bonferroni correction for the 1 million tests employed in early GWAS. Other approaches to balance power and the risk of type I errors exist[209].

Genetic association analysis is sensitive to population stratification[208,210]. A small group of individuals with deviating allele frequencies can cause false findings. The population stratification problem can be resolved by modeling the ethnicity of study participants by principal component analysis of their genotypes, and subsequently using the strongest principal components as covariates in the association analyses[211]. The principal components can also be used to exclude individuals that do not cluster with other participants, suggestive of a different ethnicity. Since exclusion will decrease the sample size and statistical power, it is generally not preferred but, nevertheless, sometimes necessary. Studies on appropriate statistical methods in GWAS is an area of intensive research and other approaches to overcome stratification are described, such as linear mixed modeling[212,213]. With a large number of participants, that could be well over 100,000[214], genotyped for millions of markers, the analysis is also becoming more computationally challenging with a need for more efficient methods.

GWAS can be augmented with family-based designs[215]. While GWAS typically include only unrelated individuals (close relatives are actively excluded[216]), the classical way of resolving the genetic basis of traits, linkage analysis, is based on the family structure[217]. The family-based linkage analysis is uniformly more powerful than unrelated GWAS given the same sample size[218] but is considered more laborious to perform in addition to other drawbacks[215]. To take advantage of both study designs, family-based association study designs and statistics have been developed[208,215]. Adding knowledge on family structure to association studies increase their power while still keeping them feasible.

Genetic association analysis was used in Paper I, incorporating elements of family-based designs.

Linkage disequilibrium measurements

Linkage disequilibrium (LD), the statistical association in a population of the alleles at two loci, is an important concept in statistical genetics[208,219]. Several measures are defined and are useful for different purposes.

We denote the alleles of two loci A/a and B/b and assign them probabilities p_A, p_a, p_B and p_b. This gives us four possible haplotypes, with probabilities p_{AB}, p_{aB}, p_{Ab} and p_{ab}. At linkage equilibrium (LE), the probability of a haplotype AB (p_{AB}) will equal the product of the included alleles (p_{AB} is equal to $p_A p_B$) (Table 6)[208]. Deviation from this equality indicates LD, and can be measured by the LD coefficient D, usually defined as $D = p_{AB} - p_A p_B$. D falls within the range -1 to 1, with $D = 0$ at LE.

	B	b	Total
A	$p_{AB} = p_A p_B$	$p_{Ab} = p_A p_b$	p_A
a	$p_{aB} = p_a p_B$	$p_{ab} = p_a p_b$	p_a
Total	p_B	p_b	

Table 6 Random population allele frequencies for two loci under linkage equilibrium.

D is, however, sensitive to low minor allele frequencies and could incorrectly suggest LE when rare alleles are present. Other, alternative and more robust, measures have been derived, and are commonly used[208]. Two such measures, D' and the squared correlation coefficient r^2, are commonly used. For values of $D > 0$, D' is defined as[208,220]

$$D' = \frac{D}{-\min(p_A p_B,\ p_a p_b)}$$

and, for $D < 0$,

$$D' = \frac{D}{\min(p_A p_b,\ p_a p_B)}.$$

It follows that $0 \leq D' \leq 1$, with $D' = 0$ at LE and $D' = 1$ when any haplotype count is 0. While D' is an improvement over D for haplotypes including rare alleles, r^2 is a measure that can be used to predict the other locus[208]. This is necessary for genotype imputation, where dense genetic datasets (such as 1000G) are used to predict missing genotypes in lower resolution data[221]. The lower resolution data can be generated in, e.g., SNP arrays. Genotype imputation is employed in GWAS to increase the study resolution without having to sequence all samples. One way of calculating r^2, given D and the allele frequencies at two loci A and B, is

$$r^2 = \frac{D^2}{p_A p_B p_a p_b}.$$

r^2 will be 1 only when there is a single haplotype, and the other locus can be perfectly predicted.

LD measurements were performed in Papers I and II.

Other statistical analyses

R (https://www.r-project.org/) and Python with the Pandas (https://pandas.pydata.org/) and SciPy (https://www.scipy.org/) libraries were the main programming languages used for statistical analysis. When necessary, tests were implemented using published algorithms and formulas, or adapted from open-source implementations. Non-parametric tests were used unless assumptions were met for

using parametric tests. Two-sided tests (where applicable) were used throughout. A p-value less than 0.05 was considered statistically significant, unless otherwise noted.

Statistical analyses other than genetic association analysis and LD measurements were used in Papers I, II and IV.

Software implementation

A large variety of software was developed for most aspects of the work in this thesis. When possible, existing programs and libraries were used and combined with scripts. The coding generally followed good coding practices[222], with increasing adherence over time. Precision and correctness was always prioritized and never sacrificed for performance or elegance.

A number of programming languages were used, including Python, gnuplot, bash with GNU coreutils, awk, sed, C++, R and Nim. The choice was made depending on the availability of tooling and libraries, and the need for performance and low memory usage. All programs and libraries used in this thesis were free* and open-source.

Software was developed for Papers I–IV.

* *Free software is a matter of liberty, not price. To understand the concept, you should think of free as in free speech, not as in free beer.* – Richard Stallman

Results and discussion

Elucidation of the genetic background of Vel (Paper I)

Results

Samples were collected from 20 Vel– and 8 Vel+ blood donors and genotyped for 2.44 million SNPs using Illumina HumanOmni 2.5M SNP arrays. In the 28 genotyped individuals, there were 5 Vel– and 4 Vel+ related individuals, representing two separate families (Figure 2).

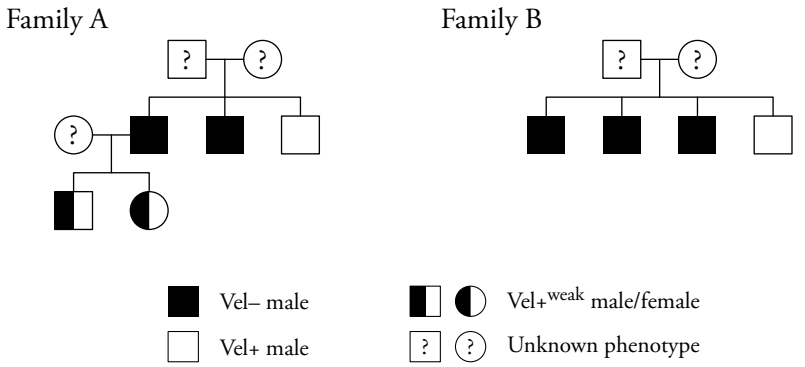


Figure 2 Pedigrees showing the relatedness between the genotyped individuals in Family A and B, respectively.

Since the Vel– trait follows an autosomal recessive inheritance pattern[110], the relatedness was used to design a filtering algorithm for selection of SNPs following the expected pattern. The algorithm selected all SNPs that were identical between the Vel– siblings and non-identical in the Vel+ family members within each of the two families. Applying this algorithm on the genotype data, 8,780 out of the original 2.44 million SNPs remained.

After discarding the related samples from further analysis – to minimize risk of bias – the genotype frequencies of the 15 remaining Vel– individuals were used

for association testing*. As controls, the 379 samples in the 1000G (phase 1) EUR super population were used. The genotype frequencies for all SNPs in the filtered Vel– samples were compared to the frequency in the 1000G samples using Fisher's exact test (due to the small number of Vel– samples).

The association analysis identified 25 SNPs in a 97-kb haplotype block on chromosome 1p36, the limits of which coincided with recombination hotspots as calculated from 1000G data (Figure 3). All 20 genotyped Vel– individuals were homozygous and identical within the identified block. The block contained five genes: *CCDC27*, *SMIM1*, *LRRC47*, *CEP104* and *DFFB*. Of these genes, the previously uncharacterized and uncited *SMIM1* (Small membrane integral protein 1) was found to be expressed in erythroid cell line data and predicted to be a transmembrane protein in databases, thus sharing the properties of many previously known blood group genes.

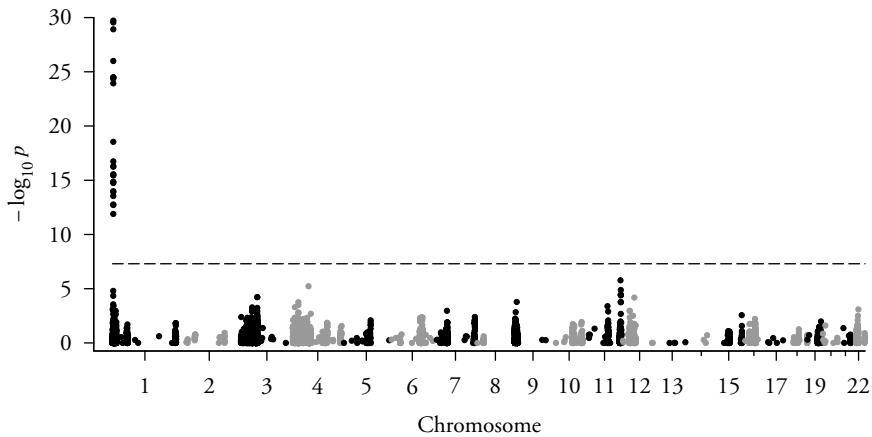


Figure 3 Manhattan plot showing $-\log_{10} p$-values for Fisher's exact test comparing the allele frequencies of the Vel– individuals to 379 individuals in the EUR population of phase 1 of 1000G. In all, 8,780 SNPs remained after applying the filtering algorithm and tested for association. The strong association of the 97-kb haplotype block at 1p32 is seen at the far left. The dashed line indicates the genome-wide significance threshold of $p < 5 \times 10^{-8}$.

Sequencing of *SMIM1* showed all Vel– individuals to be homozygous for a 17-base-pair (bp) deletion in exon 3 (c.64_80del, p.Ser22fs), causing a frameshift and an abolished expression. The findings were validated in 15 additional Vel– samples. Screening of 520 Swedish blood donors found 30 heterozygous carriers, but no homozygous individuals. The heterozygous carriers had lower expression

* The data from the 4 Vel+ non-related individuals were used as an internal control only.

of Vel than individuals homozygous for the wild-type allele. Rabbit antibodies were raised against an extracellular peptide sequence of the SMIM1 protein, and the expression of SMIM1 was found to correlate with the genotype in Western blotting experiments. The Vel– samples had no expression of SMIM1. *SMIM1* was overexpressed in K562 cells and resulted in higher expression of the Vel antigen.

SMIM1 was found to be a small gene (3,187 bp) with four exons, and predicted to encode a type 1 transmembrane protein of 78 amino acids. An extracellular domain of approximately 50 amino acids, with potential O-glycosylation sites, was predicted. Searches in protein databases found no SMIM1 homologs in man, but in 45 other species, from primates to sea squirt. A potential GXXXG dimerization motif[223] was found in the predicted transmembrane domain. *SMIM1* was found to be expressed mainly in erythroid tissue, and had an evolutionarily conserved GATA1 binding motif.

Discussion

The genetic background of the highly elusive Vel blood group antigen was found to be a 17-bp deletion (assigned dbSNP id rs566629828) in the previously completely uncharacterized gene, *SMIM1** All Vel– samples tested were homozygous for the deletion whereas Vel+ individuals carried at least one wildtype allele. The fact that all Vel– had the same deletion suggests a founder effect. Our results were independently confirmed by two other groups[70,71].

The genetic background underlying the Vel– phenotype was unraveled by combining association analysis with family information, drastically decreasing the search space and increasing the study power. This particular approach was not tested previously in studies of genes underlying blood groups but proved successful. In part, this was due to all Vel– individuals having the same deletion. Homozygosity mapping (which could be regarded a form of non-parametric linkage analysis[215]) in related samples was, however, used to map *ABCG2* to the Jr(a–) phenotype[67]. This approach would also, in retrospect, have been successful in our study since all Vel– samples were sharing the same haplotype.

Establishing the genetic background of Vel has made genetic screening for Vel– donors possible. Previously, there were no antisera readily available and large-scale screening campaigns were not feasible. In a project at the Department of Clinical Immunology and Transfusion Medicine in Lund, several new Vel–

* While the manuscript was prepared, the gene was still called *LOC388588*.

donors were identified in a short period of time, proving how the results were directly translational into clinical practice.

While the genetic background of the Vel– phenotype has been found, there are still questions that need to be addressed. The Vel antigen is known for its variability in expression amongst individuals and this is not fully explained in this study. However, this was addressed in Paper II. The SMIM1 protein is almost completely uncharacterized and most of what we know is based on predictions in databases. The prediction of a type I transmembrane protein has been questioned[224] but later results are contradictory[225]. Intriguingly, while the SMIM1 protein is conserved in a range of species, without any homologs providing redundancy, Vel– individuals are healthy enough to become blood donors. Published GWAS have found associations for the 17-bp deletion with RBC distribution width[226], and for a regulatory SNP in *SMIM1* intron 1, rs1175550 (studied in Paper II), with blood copper levels[227] and mean corpuscular hemoglobin concentration (MCHC)[228]. The function of SMIM1, a small transmembrane protein without annotated domains, remains to be determined.

On the basis of Paper I, Vel was promoted into a novel blood group system, the Vel blood group system, ISBT number 034 (Table 2, p. 18), with a single antigen, Vel.

Effects of genetic variation in *SMIM1* (Paper II)

Results

A regulatory region in intron 1 of *SMIM1* was brought to attention from a GWAS study of erythrocyte traits[228], where a SNP in *SMIM1*, rs1175550 was associated with MCHC. The region was found to contain an enrichment of ChIP-Seq peaks and markers of open chromatin in ENCODE data, and was sequenced in 150 Vel+ Swedish blood donors. All donors were verified not to be carriers of the *SMIM1* 17-bp deletion (Paper I). The sequenced region was found to contain eight polymorphic genetic variants (Table 7). For a subset of donors, *SMIM1* mRNA expression in blood was measured by qPCR, Vel expression levels on RBCs by flow cytometry and SMIM1 protein levels by Western blotting of RBC membranes with anti-SMIM1. *SMIM1* mRNA and Vel expression levels were tested for association with genotypes in linear regression models.

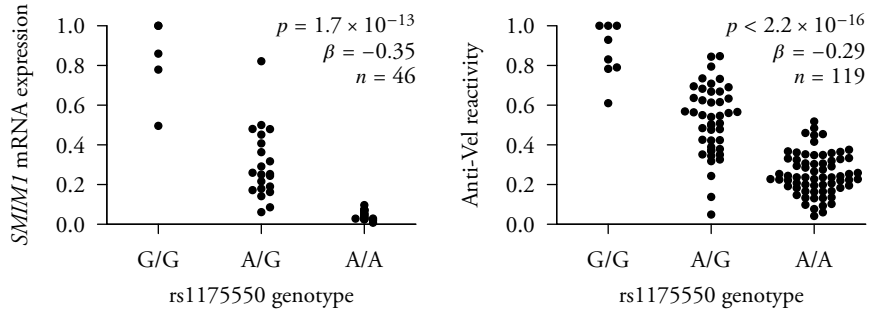


Figure 4 *SMIM1* mRNA expression and anti-Vel reactivity in Swedish blood donors, categorized by rs1175550 genotype.

A strong association of rs1175550 with *SMIM1* ($\beta = -0.35$, $p = 1.7 \times 10^{-13}$) and Vel expression ($\beta = -0.29$, $p < 2.2 \times 10^{-16}$), was found in the models, with higher expression for the minor G allele (Figure 4). The effect was also seen in Western blotting of RBCs membranes with anti-SMIM1. An equally strong association as for rs1175550 was found for a 3-bp indel, rs143702418 (rs70940313 in reverse complement), but not for the six remaining SNPs in the sequenced region. However, the independent effect of rs143702418 could not be established due to strong LD with rs1175550 in the Swedish donors. The strong LD was confirmed in the 1000G EUR super population ($r^2 = 0.88$) (Table 7).

	EUR		AFR	
SNP	r^2	D'	r^2	D'
rs1175550	reference	reference	reference	reference
rs1184341	0.98	0.99	0.38	0.89
rs2797432	0.22	1.00	0.08	0.98
rs143702418	0.88	0.95	0.22	0.98
rs1181893	0.00	1.00	0.00	1.00
rs6673829	0.13	–1.00	0.07	–1.00
rs9424296	0.02	–0.93	0.01	–1.00
rs1175549	0.88	0.94	0.96	0.98

Table 7 Identified SNPs in intron 1 of *SMIM1* and their linkage disequilibrium with rs1175550 in the EUR and AFR super populations in 1000G.

LD between rs1175550 and rs143702418 was measured for other populations in 1000G, and was found to be lower in the AFR super population (Table 7), that includes African Americans. This was mostly due to the higher prevalence of the rs1175550G-rs143702418C haplotype, uncommon in the 1000G EUR super population.

The identified region in *SMIM1* intron 1 was sequenced in blood samples from 202 African American blood donors. The rs1175550 and rs143702418 genotypes were tested for association with *SMIM1* mRNA and Vel antigen expression in linear regression models. The effects of rs1175550 were confirmed (Vel antigen expression $p < 2.2 \times 10^{-16}$), but rs143702418 showed less significant effects on *SMIM1* and Vel antigen expression in the African American ($p = 0.009$ and $p = 6.0 \times 10^{-5}$, respectively) than in the Swedish donors. Consistent with the Swedish data, no association was found for the six remaining SNPs in the sequenced region.

To test the hypothesis that rs143702418 had an independent effect on *SMIM1* mRNA and Vel antigen expression, a multiple linear regression model including both of rs1175550 and rs143702418 was fitted. The model revealed a small but significant effect of rs143702418 on Vel antigen ($\beta = -0.05$, $p = 0.006$) but not *SMIM1* mRNA expression levels ($p = 0.29$) in the African American samples (Figure 5). The expression was lower with the minor CGCA allele of rs143702418. The independent effects of rs1175550 and rs143702418 were supported by luciferase experiments.

To explain the effects of the two variants, ENCODE data for the identified region were again examined for the presence of ChIP-Seq peaks. The erythroid transcription factors GATA1, KLF1, TAL1 and ZBTB7A were identified as potential modulators of *SMIM1* expression, and examined in electrophoretic mobility shift assays (EMSA). The transcription factor TAL1 was suggested to bind to the G allele of rs1175550, although the exact mechanisms were unclear. No antibody supershifts were seen for rs143702418 in EMSA.

Discussion

The strong, previously reported[70,114], association of rs1175550 with *SMIM1* and Vel expression, was confirmed in our data. By careful examination of local LD patterns in different populations, we could utilize a different population to untangle the contributions of individual variants. This established rs143702418 as an independent expression quantitative trait locus (eQTL) in *SMIM1* and found six additional SNPs not to be significant.

The minor CGCA allele of rs143702418 had lower expression in the African American donors, opposite of what was expected from the Swedish data. This was due to the strong LD between rs1175550 and rs143702418 in the Swedish donors, with the G and CGCA alleles in a common haplotype. The strong effects of rs1175550 obscured the weaker effect of rs143702418 and highlights

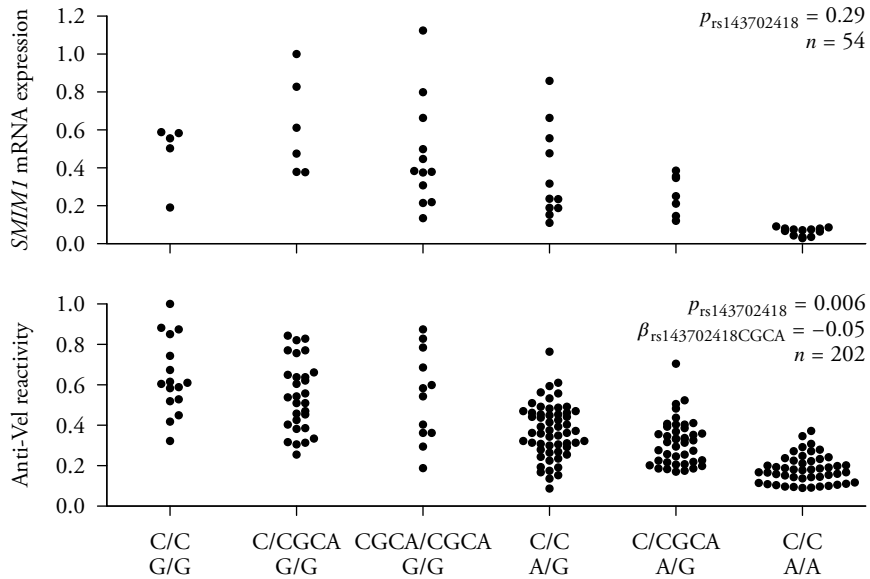


Figure 5 *SMIM1* mRNA expression and anti-Vel reactivity in African American blood donors, categorized by rs143702418-rs1175550 genotype.

the importance of carefully examining LD patterns and performing validating experiments before concluding on causality of genetic variants.

While this work focused on expression regulators in samples not carrying the 17-bp Vel-determining deletion identified in Paper I, previous work has shown that the deletion is the single strongest determinant of Vel antigen expression[114]. In addition to the 17-bp deletion and the effects of rs1175550 and rs143702418 reported here, two additional missense SNPs, c.152T>A and c.152T>G (yet unnamed and not included in dbSNP[229] version 150), have been shown to cause the Vel+^{weak} phenotype[70]. These variants lead to an amino acid shift close to the transmembrane region. They seem to be very rare, and neither of them are found in almost 150,000 alleles in gnomAD[205]. More eQTLs affecting Vel antigen strength are expected to be found in future studies.

The transcription factor TAL1, a regulator of gene expression in RBCs[143], was suggested to cause the eQTL effects seen for rs1175550. This was in concordance

with findings of another study that also found an increase of expression upon GATA1 binding[230]. Further studies are needed to fully explain the findings.

Genetic variation at blood group loci in the 1000 Genomes project (Paper III)

Results

Genotype data for 41 blood group genes (Table 2, p. 18) and the two erythroid transcription factors *GATA1* and *KLF1* were extracted for all individuals in phase 3 of 1000G and imported to a database developed by fellow PhD student Mattias Möller. Coordinates were remapped to Locus Reference Genomic (LRG) records[231]. Variants were classified according to terms defined by Sequence Ontology (SO)[232].

Among the 2,504 individuals in 1000G, 52,305 variants at 52,955 sites were found. Of all variants, 50,076 were SNPs and 2,168 were indels. There was a large variation in the number of variants per kb gene, with *BSG*, *CD151* and *GYPB* having the highest frequency of variants per kb/gene, and *C4A* and *C4B* the least. Comparing individual variant frequencies between populations, the variant rs2814778 in *ACKR1* was the most unevenly distributed. This variant disrupts a GATA1 binding site and causes the Fy(a–b–) phenotype, common in African populations and providing a relative resistance to *P. vivax* and *P. knowlesi* malaria[145].

Allele lists were compiled from official ISBT lists[17], The Blood Group Antigen Gene Mutation Database (BGMUT)[186], and tables in Reid et al.[18]. The 1000G genotype data were matched to the allele lists and annotated with the names of known alleles. In total, 2,462 unique alleles were found for the 43 included genes, 958 (38.9%) of which could be matched to a known allele, named by ISBT. In the coding regions, 1,241 non-synonymous variants were found, only 241 of which had known association with blood group variation. Thus, 1,000 variants were not previously recognized.

For the 31 protein blood group genes, the positions corresponding to extracellular portions of proteins were listed and compared to the list of variants not previously recognized. A total of 357 of the 1,000 variants matched an extracellular amino acid. Since these variants were predicted to cause an amino acid exchange, they may represent novel antigens or modified phenotypes (Figure 6).

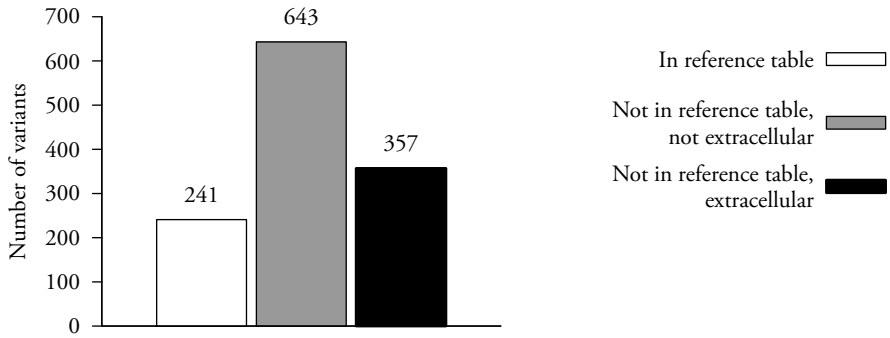


Figure 6 Non-synonymous coding region variants in protein blood group genes in the 1000G data.

The results were made available through a public online database, *Erythrogene* (https://www.erythrogene.com/), enabling scientists and clinical reference laboratory staff to search for variants and alleles in blood group genes.

Discussion

The genetic variation in all known blood group-related genes was analyzed in the full 1000G data and compiled into a publicly available database. Previously, only proof-of-concept studies of these methods for blood group genotyping purposes were published[182,183]. This study analyzes the genetic variation and maps alleles to lists of known alleles, on an unprecedented scale, and for all blood group systems.

While NGS is an established method for tissue typing purposes, the classical HLA genes (the most polymorphic human genes) have a lower frequency of indels, structural variants and hybrids than blood group genes. Although most blood group antigens arise from SNPs, that are easily detected given sufficient read depth, it is much more challenging to comprehensively annotate alleles that include other types of variation. This is underscored by the fact that some prominent variants were missing from the results, including the deletion of *RHD* carried by about 40% of individuals in Western populations[18], and the 17-bp deletion defining the Vel– phenotype (Paper I). The proper software support to help resolving this is missing. The availability of searchable allele databases is the first step towards a comprehensive toolkit for blood group genotyping with NGS methods.

Of 1,241 non-synonymous variants found, 1,000 were not included in current lists of variants in blood group genes. These variants, in particular the subset of 357 variants predicted to alter the extracellular amino acid sequence, can be

of importance in the investigation of serological discrepancies in the clinical laboratory. However, the 643 variants altering the amino acid sequence in the transmembrane and intracellular regions are also of relevance since they can cause a modified expression of the protein[233,234].

The Erythrogene database is now used on a routine basis in the Nordic Reference Laboratory for Blood Group Typing and has also been used for follow-up studies including Paper IV and a study of deletions at the *ABO* locus[235].

Prediction of candidate carbohydrate blood group loci (Paper IV)

Results

The UniProt and Ensembl databases were searched for human genes annotated to have GT activity (Glycosyltransferases and glycosylation, p. 26), and with predicted or proven expression on protein level. The list of genes was cross-checked against the CAZy database without any additional findings. In total, 244 GT genes that matched the criteria were found in the human genome, distributed over 44 families.

SNPs and indels at GT loci were collected from 1000G data. In total, there were 550,275 variants at the 244 loci, out of which 543,040 were SNPs and 7,235 were indels (length ≤ 48 bp). All variants were classified with the Variant Effect Predictor[236] tool and almost all (98.7%) were found to be located outside exons and splice site regions, or were synonymous. There were 329 variants in 149 GT genes classified to have a high impact, generally predicted to result in a null allele.

To evaluate the expression of GT genes in erythroid tissue, RNA sequencing (RNA-Seq) data generated in erythroid cells were used. The list of GT genes was limited to those with an expression similar to, or higher than, the known blood group-related GT genes (Table 2, p. 18) except *FUT2* and *FUT3*, known not to be expressed endogenously in erythroid cells. This identified 155 (64%) GT genes to be expressed at the defined level (data missing for 8 genes). Furthermore, an enrichment of ChIP-Seq peaks for the erythroid transcription factor GATA1 was found among the 155 identified genes, as compared to genes with lower expression ($p < 1.7 \times 10^{-6}$, χ^2-test).

From the list of GT genes expressed in erythroid tissue, non-candidate genes were removed in a stepwise procedure. Specifically, any gene that had a known disease phenotype, that was reported to be essential for glycosylation, or was a known blood group gene was removed, leaving a list of 30 candidate blood group genes (Table 8). These candidates were predicted to result in a benign phenotype in homozygotes or compound heterozygotes for null alleles, similar to known

carbohydrate blood group genes. Out of the 30 candidates, 16 had high impact variants in 1000G, predicted to cause null alleles, and all 30 had these types of variants in the larger gnomAD[205] dataset.

Gene	Family	1000G high impact variants
B3GNT2	GT31	0
B3GNT9	GT31	0
B3GNTL1	GT2	5
B4GALT2	GT7	0
B4GALT3	GT7	0
B4GALT4	GT7	3
B4GALT6	GT7	0
DPY19L1	unknown	1
DPY19L3	GT98	0
DPY19L4	unknown	4
FUT4	GT10	0
FUT7	GT10	0
FUT10	GT10	3
FUT11	GT10	2
GCNT1	GT14	1
GLT8D1	GT8	1
GTDC1	GT4	1
GXYLT1	GT8	0
KDELC1	GT90	0
ST3GAL1	GT29	0
ST3GAL2	GT29	1
ST3GAL4	GT29	1
ST3GAL6	GT29	1
ST6GAL1	GT29	0
ST6GAL2	GT29	0
ST6GALNAC1	GT29	1
ST6GALNAC4	GT29	1
ST6GALNAC6	GT29	2
ST8SIA4	GT29	0
ST8SIA6	GT29	3

Table 8 Candidate GT genes expressed in RBCs with a benign predicted impact. Indicated is their family and the number of high impact variants present in 1000G.

Discussion

In all expressed GT genes identified in the human genome, 30 were found to be blood group gene candidates. These genes shared the properties of carbohydrate-based blood group systems presently recognized by the ISBT and were selected in a stepwise model.

Among the existing carbohydrate blood group genes, most are expressed in erythroid tissue. RNA-Seq data from erythroid cells were used to select the GT genes expressed at a level higher than or similar to that of the known blood group GT gene with the lowest expression. However, as shown by the Lewis blood group antigens, produced in epithelial cells and adsorbed from plasma onto the RBC surface, the possibility of false negatives cannot be excluded.

By definition, any blood group antigen requires a human alloantibody to have been formed by at least one individual (Terminology, p. 20). The added layer of complexity in carbohydrate-based blood groups, where the antigen is not located on the protein itself, but on the glycan product, makes predictions on immunogenicity challenging. In addition, the identified variants predicted to result in a null allele had low allele frequencies, and it can be expected that any novel antigens are either very high- or very low-frequency. Beside null alleles, SNPs that cause qualitative alterations in the donor or acceptor specificity of GTs are a possible cause of loss of enzyme function.

The list of candidates includes four members of the GT29 family with α-2,3-sialyltransferase activity (*ST3GAL1*, *ST3GAL2*, *ST3GAL4* and *ST3GAL6*), a property required for the synthesis of the LKE orphan blood group antigen[18]. These four candidates may well be a starting point for studies of LKE, and the other candidate genes presented here could serve as a help in investigating the genetic background of other orphan and emerging blood groups.

General discussion

This thesis comprises four papers investigating the genetic variation that underlies the expression of human blood group genes and antigens. Bioinformatic methods were used to find variation in known blood group systems, to define variation underlying a novel system, and to predict the presence of, and potential for, candidate systems.

The thesis highlights the possibilities generated by the incorporation of bioinformatic competence in laboratory transfusion medicine research, creating synergistic effects. The next step in blood group genotyping, utilization of NGS technologies for blood group typing purposes, is coming closer and closer. As these studies show, bioinformatic competence is needed until robust and user-friendly tools for the correct interpretation of this data are available. The database assembled in Paper III has laid the foundations for the development of tools to simplify the analysis of NGS data in the transfusion medicine field.

The studies included display the use of genomic databases from two perspectives. The first perspective is to broaden the views of a finding, for example by studying the allele frequency of a variant found by sequencing of a patient or donor sample. This is indeed very helpful, since the available information quickly deepens the understanding of the finding. The other perspective has the reverse approach, by predicting the existence of variation and its consequences from the database first, and only then confirming its existence in a patient or donor sample. The benefit is that much of the work can be done without the need for samples until the confirmation stage, and much of the work can be at least semi-automated. This latter approach was tested in all four papers, with great success in Papers I–II, and with the results of Paper IV pending confirmation in future studies. The predictions in Paper III have been examined in a recently published follow-up study, where the presence of a previously unknown deletion in *ABO* was confirmed while another was dismissed[235].

The reverse perspective outlined above has its limitations. There is great potential for utilizing data from exome or whole-genome sequencing for a multitude of purposes, blood group genotyping being one of them. One of the challenges

faced with this perspective is the availability of phenotypes. 1000G and other public datasets are excellent resources but lack information on, among other limitations, blood group phenotypes. A cohort of whole-genome sequenced individuals, serologically typed for antigens of interest, would be an excellent data source for the characterization of the expression of blood group antigens. Sequencing of established blood donor cohorts would be an appropriate beginning. These studies could be performed in the Nordic countries with an abundance of cross-linked registers[237,238].

A proposed strategy for resolving the genetic background of current orphan blood group antigens is to utilize the available genomic data. Potential genes, and even variants in these genes, can be filtered out from the available data. These data include gene expression in erythroid tissue and variant frequencies in the different population in 1000G or other datasets. The approach is particularly feasible when there are candidate genes present. A brief list of candidate genes and variants can be collected and samples later tested for the presence of variation explaining the phenotype.

I expect that future studies will resolve the genetic background of all current orphan blood group antigens, provided that patient samples are available. Based on current knowledge of variation in genes expressed in erythroid tissue, there are probably many, many more blood group antigens emerging, most of which will probably be very low- (or very high)-frequency. The usage of NGS technologies for blood group genotyping is a natural first step for these future studies.

Conclusions

The main conclusions of this thesis are:

Paper I — Homozygosity for a 17-bp deletion in exon 3 of the previously uncharacterized gene *SMIM1* defines the Vel– blood group phenotype.

Paper II — Two SNPs in a regulatory region of *SMIM1* intron 1, rs1175550 and rs143702418, have independent effects on the expression of the Vel blood group antigen.

Paper III — A database of known and unknown alleles in known blood group-related genes in 1000G was established. Among all identified genetic variants, 357 were non-synonymous and predicted to occur on the extracellular portion of blood group-carrying proteins and could represent novel blood group antigens or modified phenotypes.

Paper IV — Among all human genes, 244 expressed genes with GT activity were identified. Out of these, 30 were predicted to have the properties of known genes defining carbohydrate-based blood group systems and could represent candidate blood group loci.

Populärvetenskaplig sammanfattning

Bakgrund

Transfusion av de syrebärande röda blodkropparna är en viktig del av modern sjukvård och är nödvändigt vid t.ex. större kirurgi och cancerbehandling. I dag är blodtransfusioner säkra och en mycket liten andel av de som transfunderas (får blodtransfusion) drabbas av biverkningar. De första blodtransfusioner som gjordes mellan människor var långt ifrån säkra och ledde ibland, men inte alltid, till allvarliga reaktioner hos mottagaren av blodet. Österrikaren Karl Landsteiner intresserade sig i slutet av 1800-talet för varför transfusionen endast ibland gav upphov till reaktioner. Han fann i sina experiment att om blodplasma från en individ blandades med röda blodkroppar från en annan så klumpades de röda blodkropparna ihop eller förstördes. Detta skedde dock bara ibland och endast i vissa kombinationer av plasma och röda blodkroppar. Landsteiner drog slutsatsen att det måste finnas olika faktorer i blodet som förklaring till resultaten. De faktorer Landsteiner (och senare andra) upptäckt var blodgrupperna A, B, AB och O som idag ingår i ett gemensamt system med namnet ABO (Figur 1, s. 17). År 1930 belönades Karl Landsteiner med Nobelpriset i fysiologi eller medicin för sina upptäckter inom området. Med denna nya kunskap kunde de tidigare allvarliga konsekvenserna till största del undvikas.

Blodgruppsantigener är strukturer på de röda blodkropparnas cellyta som har en förmåga att stimulera immunförsvaret till att börja bilda antikroppar. Blodgruppsantigenerna A och B (som Landsteiner upptäckte) är de två mest välkända men idag känner vi också till hundratals andra (Tabell 2, s. 18). Vid transfusion av blod kan de blodgruppsantigener som saknas på de egna röda blodkropparna uppfattas som främmande med följd att blodgruppsantikroppar bildas. Antikropparna kan i sin tur aktivera andra delar av immunförsvaret och förstöra de blodkroppar som tillförts vid transfusionen. Därmed uteblir den önskade effekten av transfusionen och i vissa fall uppstår även läckage av skadliga ämnen. Det är därför viktigt att påvisa blodgruppsantikroppar innan transfusioner ges. Blod från blodgivare med samma blodgrupp som patienten kan då väljas till patienten.

Arbete med påvisning av blodgruppsantikroppar och urval av blodkomponenter utförs på sjukhusens blodcentraler.

Det är genetisk variation (olikheter i vår arvsmassa) som bestämmer vilka blodgruppsantigener som finns på våra röda blodkroppar. Med kännedom om vilken genetisk variant som ger upphov till en blodgrupp kan individer med en viss blodgrupp hittas på genetisk väg, utan att de röda blodkropparna undersökts. Detta är fördelaktigt i många situationer, t.ex. för att snabbare kunna hitta en blodgivare med en sällsynt blodgrupp eller för att bestämma blodgruppen hos en patient som redan har hunnit transfunderas och därför har en blandning av eget och blodgivares blod.

Avhandlingens delarbeten

Avhandlingen innehåller fyra delarbeten som fokuserar på den genetiska bakgrunden till förekomst av blodgruppsantigener på ytan av röda blodkroppar. I avhandlingen användes bioinformatiska metoder som lånar kunskap och metodik från bl.a. biologi, statistik och datavetenskap.

I det första delarbetet studerades varför vissa individer saknar ett blodgruppsantigen med namnet Vel. Vel förekommer hos de allra flesta och till en patient som bildat antikroppar mot Vel är det därför svårt att hitta passande blod att transfundera eftersom bara ca en på tusen blodgivare i Sverige saknar Vel. Den genetiska bakgrunden till varför Vel saknas har varit okänd sedan Vel upptäcktes 1952, vilket försvårat sökandet efter passande blodgivare.

I prover från blodgivare med och utan Vel på de röda blodkropparna analyserades en stor mängd genetiska varianter spridda över hela arvsmassan. Vissa av blodgivarna var besläktade, och tack vare detta kunde merparten av alla undersökta varianter sållas bort. Förekomsten av de kvarvarande varianterna jämfördes sedan statistiskt mellan blodgivarna och individer från en stor studie av genetisk variation, 1000 Genomes project. En tidigare okänd gen, Small membrane integral protein 1 (*SMIM1*), identifierades som det troligaste upphovet till Vel. Riktade analyser bekräftade att genetisk variation i *SMIM1* förklarar varför vissa helt saknar blodgruppsantigenet Vel. Fyndet har lett till att de tidigare svårfunna blodgivare som saknar Vel nu lättare kan hittas.

I det andra delarbetet studerades varför mängden Vel på ytan av de röda blodkropparna varierar mellan olika individer. Vid mycket låga mängder kan slutsatsen felaktigt dras att Vel saknas på cellytan. Den i delarbete 1 identifierade genen *SMIM1* studerades hos blodgivare och två olika genetiska varianter konstaterades kunna påverka mängden Vel. Hur mycket de två varianterna var för sig bidrog

kunde dock inte klarläggas hos svenska blodgivare p.g.a. stark statistisk koppling varianterna emellan. Analys av data från studien 1000 Genomes project talade för att den statistiska kopplingen inte var lika stark hos individer med afrikanskt ursprung. Prover samlades in från afroamerikanska blodgivare i New York och de två genetiska varianterna analyserades hos dessa. Resultaten bekräftade att de två identifierade varianterna hade var för sig oberoende effekter på mängden Vel på cellytan. Fyndet kan leda till att metoderna för att bestämma att en individ saknar Vel blir säkrare.

I det tredje delarbetet inventerades samtliga förekommande varianter av blodgruppsgener i de 2504 individer som deltagit i studien 1000 Genomes project. Kunskap om samtliga variationer i blodgruppsgener är, utöver de vetenskapliga aspekterna, viktig för att utforma robusta genetiska metoder för bestämning av blodgrupper. Data insamlades till en upprättad databas och jämfördes med listor över tidigare kända varianter. I databasen hittades en stor mängd tidigare okända varianter. Av dessa okända varianter var 357 av den typ som skulle kunna orsaka förekomst av nya, hittills okända blodgruppsantigener. Databasen är publicerad på Internet (https://www.erythrogene.com/) och är fritt åtkomlig att användas av blodgruppsforskare och referenslaboratorier.

I det fjärde och sista delarbetet studerades den familj av enzymer som bygger sockermolekyler på cellytan. Socker i form av sammankopplade kedjor finns på ytan av samtliga levande celler. Dessa sockerkedjor byggs av enzymer i familjen glykosyltransferaser. Genetisk variation i enzymernas gener orsakar ibland att vissa enzymer inte är aktiva, med följd att sockerkedjorna kan se olika ut från individ till individ. Denna olikhet kan orsaka bildning av antikroppar mot de sockerkedjor man själv inte har och utgöra hinder vid transfusion. Blodgruppsantigenerna A och B, som Landsteiner upptäckte, utgörs t.ex. av sockerkedjor.

Målet med delarbete fyra var att hitta alla gener som ger upphov till enzymfamiljen glykosyltransferaser och undersöka den genetiska variationen i dessa. Vid sökningar i protein- och gendatabaser återfanns totalt 244 gener, vilket är ca 1% av samtliga mänskliga gener. En stor bredd i antalet och typen av genetiska varianter fanns i de funna generna. Slutligen identifierades en mindre grupp av 30 gener som bedömdes kunna utgöra basen för hittills okända blodgruppssystem. Dessa 30 gener har egenskaper som liknade idag kända blodgruppsgener. Fyndet kan vara en resurs vid utredning av antikroppar mot tidigare ej beskrivna blodgruppsantigener.

Acknowledgements

There are a great number of people to whom I would like to express my gratitude, including, but not limited to:

- Martin L Olsson, my main supervisor, for introducing me to science and for all your support over the years. You bridged my troubled waters while navigating your own. Thank you.
- Ann-Sofie Liedberg, my co-supervisor, for your reasoning manners and pragmatic thinking. Your support has been very helpful, and became critical in the last few years. Thank you.
- My de facto co-supervisor, colleague and friend, Jill R Storry, for almost anything that I can think of in relation to this thesis.
- My co-authors and fellow PhD-students, Mikael Kronborg Christophersen and Mattias Möller, for the great collaborations and discussions that lead to the papers included in this thesis. Our different perspectives have moved all papers forward and have made them a lot better.
- All members of the research group, past and present, for the friendly, social, inviting and supporting group environment. Many of you have supported and helped me more than you know.
- All colleagues and co-workers at the department of Clinical Immunology and Transfusion Medicine for the enjoyable working environment. I would especially like to thank Josefina Dykes for all the support given during the course of the studies.
- I acknowledge Björn Nilsson for introducing me to bioinformatics.

Most of all, I thank my family in Stockholm and Skåne for supporting me up to this point in life, and beyond. My parents have always believed in me and promoted the spirit that nothing is impossible.

Anna, Måns och Elsa, jag älskar er.

References

1. Abedi MR, Andersson C, Medioni C, Bohl L, Forsberg PO, Norda R. Blodverksamheten i Sverige 2016. [Internet]. Svensk förening för transfusionsmedicin; 2017 [cited 2018 Apr 2]. Available from: http://www.kitm.se/sv/wp-content/uploads/2018/01/Blodverksamheten-i-Sverige-2016.pdf.
2. Morton J, Anastassopoulos KP, Patel ST, Lerner JH, Ryan KJ, Goss TF, et al. Frequency and outcomes of blood products transfusion across procedures and clinical conditions warranting inpatient care: an analysis of the 2004 healthcare cost and utilization project nationwide inpatient sample database. *Am J Med Qual.* 2010;25:289–296.
3. World Health Organization. Global status report on blood safety and availability 2016.Geneva: World Health Organization; 2017.
4. Goodnough LT, Shander A. Patient blood management. *Anesthesiology.* 2012;116:1367–1376.
5. Palfi M, Säfwenberg J. Hemovigilans i Sverige 2014–2016. [Internet]. Svensk förening för transfusionsmedicin; 2017 [cited 2018 Apr 17]. Available from: http://www.kitm.se/sv/wp-content/uploads/2017/08/Hemovigilans-i-Sverige-2014-2016.pdf.
6. Klein HG, Anstee DJ. Mollison's Blood Transfusion in Clinical Medicine.12th ed. Chichester: John Wiley and Sons, Inc.; 2014.
7. Seltsam A, Wagner FF, Salama A, Flegel WA. Antibodies to high-frequency antigens may decrease the quality of transfusion support: an observational study. *Transfusion.* 2003;43:1563–1566.
8. Lower R. The method observed in transfusing the bloud out of one animal into another. *Phil Trans.* 1666;1:353–358.
9. Blundell J. Experiments on the transfusion of blood by the syringe. *Med Chir Trans.* 1818;9:56–92.
10. Blundell J. Some account of a Case of Obstinate Vomiting, in which an attempt was made to prolong Life by the Injection of Blood into the Veins. *Med Chir Trans.* 1819;10:296–311.

11. Waller C. Case of Uterine Hemorrhage, in which the Operation of Transfusion was successfully performed. *Med Phys J.* 1825;54:273–277.
12. Blundell J. Observations of transfusion of blood. *Lancet.* 1829;12:321–324.
13. Landois L. Die Transfusion des Blutes; Versuch einer physiologischen Begrundung nach eigenen Experimental-Untersuchungen mit Berucksichtigung der Geschichte, der Indicationen, der operativen Technik und der Statistik.Leipzig: F. C. W. Vogel; 1875.
14. Landsteiner K. Zur Kenntnis der antifermentativen, lytischen und agglutinierenden Wirkungen des Blutserums und der Lymphe. *Zbl Bakt.* 1900;27:357–362.
15. Landsteiner K. Über Agglutinationserscheinungen normalen menschlichen Blutes. *Klin Wschr.* 1901;14:1132.
16. von Decastello A, Stürli A. Über die Isoagglutinine im Serum gesunder und kranker Menschen. *München Med Wochenschr.* 1902;49:1090–1095.
17. Red Cell Immunogenetics and Blood Group Terminology Working Party. Red Cell Immunogenetics and Blood Group Terminology. [Internet]. International Society of Blood Transfusion; 2018 [cited 2018 Apr 2]. Available from: http://www.isbtweb.org/working-parties/red-cell-immunogenetics-and-blood-group-terminology/.
18. Reid ME, Lomas-Francis C, Olsson ML. The Blood Group Antigen FactsBook.3rd ed. Amsterdam: Elsevier/Academic Press; 2012.
19. Anstee DJ. The relationship between blood groups and disease. *Blood.* 2010;115:4635–4643.
20. Yamamoto F, Marken J, Tsuji T, White T, Clausen H, Hakomori S. Cloning and characterization of DNA complementary to human UDP-GalNAc:Fucα1→2Gal α1→3GalNAc transferase (histo-blood group A transferase) mRNA. *J Biol Chem.* 1990;265:1146–1151.
21. Yamamoto F, Clausen H, White T, Marken J, Hakomori S. Molecular genetic basis of the histo-blood group ABO system. *Nature.* 1990;345:229–233.
22. Siebert PD, Fukuda M. Isolation and characterization of human glycophorin A cDNA clones by a synthetic oligonucleotide approach: nucleotide sequence and mRNA structure. *Proc Natl Acad Sci U S A.* 1986;83:1665–1669.

23. Siebert PD, Fukuda M. Molecular cloning of a human glycophorin B cDNA: nucleotide sequence and genomic relationship to glycophorin A. *Proc Natl Acad Sci U S A.* 1987;84:6735–6739.
24. Steffensen R, Carlier K, Wiels J, Levery SB, Stroud M, Cedergren B, et al. Cloning and expression of the histo-blood group P^k UDP-galactose: Galβ1–4Glcβ1-Cer α1,4-galactosyltransferase. Molecular genetic basis of the p phenotype. *J Biol Chem.* 2000;275:16723–16729.
25. Thuresson B, Westman JS, Olsson ML. Identification of a novel *A4GALT* exon reveals the genetic basis of the P_1/P_2 histo-blood groups. *Blood.* 2011;117:678–687.
26. Westman JS, Stenfelt L, Vidovic K, Möller M, Hellberg Å, Kjellström S, et al. Allele-selective RUNX1 binding regulates P1 blood group status by transcriptional control of *A4GALT*. *Blood.* 2018;131:1611–1616.
27. Yeh CC, Chang CJ, Twu YC, Hung ST, Tsai YJ, Liao JC, et al. The differential expression of the blood group *P^1-A4GALT* and *P^2-A4GALT* alleles is stimulated by the transcription factor early growth response 1. *Transfusion.* 2018;58:1054–1064.
28. Avent ND, Ridgwell K, Tanner MJ, Anstee DJ. cDNA cloning of a 30 kDa erythrocyte membrane protein associated with Rh (Rhesus)-blood-group-antigen expression. *Biochem J.* 1990;271:821–825.
29. Chérif-Zahar B, Bloy C, Le Van Kim C, Blanchard D, Bailly P, Hermand P, et al. Molecular cloning and protein structure of a human blood group Rh polypeptide. *Proc Natl Acad Sci U S A.* 1990;87:6243–6247.
30. Le van Kim C, Mouro I, Chérif-Zahar B, Raynal V, Cherrier C, Cartron JP, et al. Molecular cloning and primary structure of the human blood group RhD polypeptide. *Proc Natl Acad Sci U S A.* 1992;89:10925–10929.
31. Arce MA, Thompson ES, Wagner S, Coyne KE, Ferdman BA, Lublin DM. Molecular cloning of RhD cDNA derived from a gene present in RhD-positive, but not RhD-negative individuals. *Blood.* 1993;82:651–655.
32. Kajii E, Umenishi F, Iwamoto S, Ikemoto S. Isolation of a new cDNA clone encoding an Rh polypeptide associated with the Rh blood group system. *Hum Genet.* 1993;91:157–162.

33. Parsons SF, Mallinson G, Holmes CH, Houlihan JM, Simpson KL, Mawby WJ, et al. The Lutheran blood group glycoprotein, another member of the immunoglobulin superfamily, is widely expressed in human tissues and is developmentally regulated in human liver. *Proc Natl Acad Sci U S A.* 1995;92:5496–5500.
34. Lee S, Zambas ED, Marsh WL, Redman CM. Molecular cloning and primary structure of Kell blood group protein. *Proc Natl Acad Sci U S A.* 1991;88:6353–6357.
35. Kukowska-Latallo JF, Larsen RD, Nair RP, Lowe JB. A cloned human cDNA determines expression of a mouse stage-specific embryonic antigen and the Lewis blood group α(1,3/1,4)fucosyltransferase. *Genes Dev.* 1990;4:1288–1303.
36. Chaudhuri A, Polyakova J, Zbrzezna V, Williams K, Gulati S, Pogo AO. Cloning of glycoprotein D cDNA, which encodes the major subunit of the Duffy blood group system and the receptor for the *Plasmodium vivax* malaria parasite. *Proc Natl Acad Sci U S A.* 1993;90:10793–10797.
37. Olivès B, Mattei MG, Huet M, Neau P, Martial S, Cartron JP, et al. Kidd blood group and urea transport function of human erythrocytes are carried by the same protein. *J Biol Chem.* 1995;270:15607–15610.
38. Bruce LJ, Anstee DJ, Spring FA, Tanner MJ. Band 3 Memphis variant II. Altered stilbene disulfonate binding and the Diego (Di^a) blood group antigen are associated with the human erythrocyte band 3 mutation Pro^{854}→Leu*. *J Biol Chem.* 1994;269:16155–16158.
39. Bartels CF, Zelinski T, Lockridge O. Mutation at codon 322 in the human acetylcholinesterase (ACHE) gene accounts for YT blood group polymorphism. *Am J Hum Genet.* 1993;52:928–936.
40. Rao N, Whitsett CF, Oxendine SM, Telen MJ. Human erythrocyte acetylcholinesterase bears the Yt^a blood group antigen and is reduced or absent in the Yt(a–b–) phenotype. *Blood.* 1993;81:815–819.
41. Ellis NA, Tippett P, Petty A, Reid M, Weller PA, Ye TZ, et al. *PBDX* is the *XG* blood group gene. *Nat Genet.* 1994;8:285–290.
42. Darling SM, Banting GS, Pym B, Wolfe J, Goodfellow PN. Cloning an expressed gene shared by the human sex chromosomes. *Proc Natl Acad Sci U S A.* 1986;83:135–139.
43. Wagner FF, Poole J, Flegel WA. Scianna antigens including Rd are expressed by ERMAP. *Blood.* 2003;101:752–757.

44. Gubin AN, Njoroge JM, Wojda U, Pack SD, Rios M, Reid ME, et al. Identification of the Dombrock blood group glycoprotein as a polymorphic member of the ADP-ribosyltransferase gene family. *Blood.* 2000; 96:2621–2627.
45. Smith BL, Preston GM, Spring FA, Anstee DJ, Agre P. Human red cell aquaporin CHIP. I. Molecular characterization of ABH and Colton blood group antigens. *J Clin Invest.* 1994;94:1043–1049.
46. Bailly P, Hermand P, Callebaut I, Sonneborn HH, Khamlichi S, Mornon JP, et al. The LW blood group glycoprotein is homologous to intercellular adhesion molecules. *Proc Natl Acad Sci U S A.* 1994;91:5306–5310.
47. Yu CY, Belt KT, Giles CM, Campbell RD, Porter RR. Structural basis of the polymorphism of human complement components C4A and C4B: gene size, reactivity and antigenicity. *EMBO J.* 1986;5:2873–2881.
48. Yu CY. The complete exon-intron structure of a human complement component *C4A* gene. DNA sequences, polymorphism, and linkage to the 21-hydroxylase gene. *J Immunol.* 1991;146:1057–1066.
49. Kelly RJ, Ernst LK, Larsen RD, Bryant JG, Robinson JS, Lowe JB. Molecular basis for H blood group deficiency in Bombay (Oh) and para-Bombay individuals. *Proc Natl Acad Sci U S A.* 1994;91:5843–5847.
50. Ho M, Chelly J, Carter N, Danek A, Crocker P, Monaco AP. Isolation of the gene for McLeod syndrome that encodes a novel membrane transport protein. *Cell.* 1994;77:869–880.
51. Colin Y, Rahuel C, London J, Roméo PH, d'Auriol L, Galibert F, et al. Isolation of cDNA clones and complete amino acid sequence of human erythrocyte glycophorin C. *J Biol Chem.* 1986;261:229–233.
52. Telen MJ, Hall SE, Green AM, Moulds JJ, Rosse WF. Identification of human erythrocyte blood group antigens on decay-accelerating factor (DAF) and an erythrocyte phenotype negative for DAF. *J Exp Med.* 1988; 167:1993–1998.
53. Wong WW, Cahill JM, Rosen MD, Kennedy CA, Bonaccio ET, Morris MJ, et al. Structure of the human *CR1* gene. Molecular basis of the structural and quantitative polymorphisms and identification of a new CR1-like allele. *J Exp Med.* 1989;169:847–863.
54. Moulds JM, Nickells MW, Moulds JJ, Brown MC, Atkinson JP. The C3b/C4b receptor is recognized by the Knops, McCoy, Swain-langley, and York blood group antisera. *J Exp Med.* 1991;173:1159–1163.

55. Rao N, Ferguson DJ, Lee SF, Telen MJ. Identification of human erythrocyte blood group antigens on the C3b/C4b receptor. *J Immunol.* 1991; 146:3502–3507.
56. Spring FA, Dalchau R, Daniels GL, Mallinson G, Judson PA, Parsons SF, et al. The In^a and In^b blood group antigens are located on a glycoprotein of 80,000 MW (the CDw44 glycoprotein) whose expression is influenced by the *In(Lu)* gene. *Immunology.* 1988;64:37–43.
57. Spring FA, Holmes CH, Simpson KL, Mawby WJ, Mattes MJ, Okubo Y, et al. The Ok^a blood group antigen is a marker for the M6 leukocyte activation antigen, the human homolog of OX-47 antigen, basigin and neurothelin, an immunoglobulin superfamily molecule that is widely expressed in human cells and tissues. *Eur J Immunol.* 1997;27:891–897.
58. Guo H, Majmudar G, Jensen TC, Biswas C, Toole BP, Gordon MK. Characterization of the gene for human EMMPRIN, a tumor cell surface inducer of matrix metalloproteinases. *Gene.* 1998;220:99–108.
59. Karamatic Crew V, Burton N, Kagan A, Green CA, Levene C, Flinter F, et al. CD151, the first member of the tetraspanin (TM4) superfamily detected on erythrocytes, is essential for the correct assembly of human basement membranes in kidney and skin. *Blood.* 2004;104:2217–2223.
60. Mudad R, Rao N, Angelisova P, Horejsi V, Telen MJ. Evidence that CDw108 membrane protein bears the JMH blood group antigen. *Transfusion.* 1995;35:566–570.
61. Yamada A, Kubo K, Takeshita T, Harashima N, Kawano K, Mine T, et al. Molecular cloning of a glycosylphosphatidylinositol-anchored molecule CDw108. *J Immunol.* 1999;162:4094–4100.
62. Yu LC, Twu YC, Chang CY, Lin M. Molecular basis of the adult i phenotype and the gene responsible for the expression of the human blood group I antigen. *Blood.* 2001;98:3840–3845.
63. Hellberg Å, Poole J, Olsson ML. Molecular basis of the globoside-deficient P^k blood group phenotype. Identification of four inactivating mutations in the UDP-*N*-acetylgalactosamine: globotriaosylceramide 3-β-*N*-acetylgalactosaminyltransferase gene. *J Biol Chem.* 2002;277:29455–29459.
64. Roudier N, Ripoche P, Gane P, Le Pennec PY, Daniels G, Cartron JP, et al. AQP3 deficiency in humans and the molecular basis of a novel blood group system, GIL. *J Biol Chem.* 2002;277:45854–45859.

65. Tilley L, Green C, Poole J, Gaskell A, Ridgwell K, Burton NM, et al. A new blood group system, RHAG: three antigens resulting from amino acid substitutions in the Rh-associated glycoprotein. *Vox Sang.* 2010; 98:151–159.
66. Svensson L, Hult AK, Stamps R, Ångström J, Teneberg S, Storry JR, et al. Forssman expression on human erythrocytes: biochemical and genetic evidence of a new histo-blood group system. *Blood.* 2013;121:1459–1468.
67. Zelinski T, Coghlan G, Liu XQ, Reid ME. *ABCG2* null alleles define the Jr(a–) blood group phenotype. *Nat Genet.* 2012;44:131–132.
68. Saison C, Helias V, Ballif BA, Peyrard T, Puy H, Miyazaki T, et al. Null alleles of *ABCG2* encoding the breast cancer resistance protein define the new blood group system Junior. *Nat Genet.* 2012;44:174–177.
69. Helias V, Saison C, Ballif BA, Peyrard T, Takahashi J, Takahashi H, et al. ABCB6 is dispensable for erythropoiesis and specifies the new blood group system Langereis. *Nat Genet.* 2012;44:170–173.
70. Cvejic A, Haer-Wigman L, Stephens JC, Kostadima M, Smethurst PA, Frontini M, et al. *SMIM1* underlies the Vel blood group and influences red blood cell traits. *Nat Genet.* 2013;45:542–545.
71. Ballif BA, Helias V, Peyrard T, Menanteau C, Saison C, Lucien N, et al. Disruption of *SMIM1* causes the Vel– blood type. *EMBO Mol Med.* 2013;5:751–761.
72. Anliker M, von Zabern I, Höchsmann B, Kyrieleis H, Dohna-Schwake C, Flegel WA, et al. A new blood group antigen is defined by anti-CD59, detected in a CD59-deficient patient. *Transfusion.* 2014;54:1817–1822.
73. Daniels G, Ballif BA, Helias V, Saison C, Grimsley S, Mannessier L, et al. Lack of the nucleoside transporter ENT1 results in the Augustine-null blood type and ectopic mineralization. *Blood.* 2015;125:3651–3654.
74. Huang CH, Chen Y, Blumenfeld OO. A novel St^a glycophorin produced via gene conversion of pseudoexon III from glycophorin E to glycophorin A gene. *Hum Mutat.* 2000;15:533–540.
75. Willemetz A, Nataf J, Thonier V, Peyrard T, Arnaud L. Gene conversion events between *GYPB* and *GYPE* abolish expression of the S and s blood group antigens. *Vox Sang.* 2015;108:410–416.
76. Daniels G. Human blood groups.3rd ed. Chichester: Wiley-Blackwell; 2013.

77. Storry JR, Castilho L, Chen Q, Daniels G, Denomme G, Flegel WA, et al. International society of blood transfusion working party on red cell immunogenetics and terminology: report of the Seoul and London meetings. *ISBT Sci Ser.* 2016;11:118–122.
78. Solheim BG. Provision of K– (KEL1–) blood to women not more than 50 years of age. *Transfusion.* 2015;55:468–469.
79. Anstee DJ. Red cell genotyping and the future of pretransfusion testing. *Blood.* 2009;114:248–256.
80. Hendrickson JE, Delaney M. Hemolytic disease of the fetus and newborn: modern practice and future investigations. *Transfus Med Rev.* 2016;30:159–164.
81. de Haas M, Thurik FF, Koelewijn JM, van der Schoot CE. Haemolytic disease of the fetus and newborn. *Vox Sang.* 2015;109:99–113.
82. Story CM, Mikulska JE, Simister NE. A major histocompatibility complex class I-like Fc receptor cloned from human placenta: possible role in transfer of immunoglobulin G from mother to fetus. *J Exp Med.* 1994; 180:2377–2381.
83. Socialstyrelsen. Intrauterina behandlingar som rikssjukvård. Utvärdering och definitionsöversyn 2013–2015. [Internet]. Socialstyrelsen; 2017 [cited 2018 May 1]. Available from: http://www.socialstyrelsen.se/Lists/Artikelkatalog/Attachments/20463/2017-1-15.pdf.
84. Crowther C, Middleton P. Anti-D administration after childbirth for preventing Rhesus alloimmunisation. *Cochrane Database Syst Rev.* 2000; (2):CD000021.
85. Tiblad E, Taune Wikman A, Ajne G, Blanck A, Jansson Y, Karlsson A, et al. Targeted routine antenatal anti-D prophylaxis in the prevention of RhD immunisation–outcome of a new antenatal screening and prevention program. *PLoS One.* 2013;8:e70984.
86. Colin Y, Chérif-Zahar B, Le Van Kim C, Raynal V, Van Huffel V, Cartron JP. Genetic basis of the RhD-positive and RhD-negative blood group polymorphism as determined by Southern analysis. *Blood.* 1991; 78:2747–2752.
87. Lo YM, Corbetta N, Chamberlain PF, Rai V, Sargent IL, Redman CW, et al. Presence of fetal DNA in maternal plasma and serum. *Lancet.* 1997; 350:485–487.
88. Lo YM, Hjelm NM, Fidler C, Sargent IL, Murphy MF, Chamberlain PF, et al. Prenatal diagnosis of fetal RhD status by molecular analysis of maternal plasma. *N Engl J Med.* 1998;339:1734–1738.

89. Clausen FB, Christiansen M, Steffensen R, Jørgensen S, Nielsen C, Jakobsen MA, et al. Report of the first nationally implemented clinical routine screening for fetal *RHD* in D– pregnant women to ascertain the requirement for antenatal RhD prophylaxis. *Transfusion.* 2012;52:752–758.
90. Haimila K, Sulin K, Kuosmanen M, Sareneva I, Korhonen A, Natunen S, et al. Targeted antenatal anti-D prophylaxis program for RhD-negative pregnant women – outcome of the first two years of a national program in Finland. *Acta Obstet Gynecol Scand.* 2017;96:1228–1233.
91. de Haas M, van der Ploeg CPB, Scheffer PG, Verlinden DA, Hirschberg F, Abbink H, et al. A nationwide fetal *RHD* screening programme for targeted antenatal and postnatal antiD. *ISBT Sci Ser.* 2012;7:164–167.
92. Koelewijn JM, Vrijkotte TGM, de Haas M, van der Schoot CE, Bonsel GJ. Risk factors for the presence of non-rhesus D red blood cell antibodies in pregnancy. *BJOG.* 2009;116:655–664.
93. Morath C, Zeier M, Döhler B, Opelz G, Süsal C. ABO-incompatible kidney transplantation. *Front Immunol.* 2017;8:234.
94. Urschel S, West LJ. ABO-incompatible heart transplantation. *Curr Opin Pediatr.* 2016;28:613–619.
95. Worel N. ABO-mismatched allogeneic hematopoietic stem cell transplantation. *Transfus Med Hemother.* 2016;43:3–12.
96. Rowley SD, Donato ML, Bhattacharyya P. Red blood cell-incompatible allogeneic hematopoietic progenitor cell transplantation. *Bone Marrow Transplant.* 2011;46:1167–1185.
97. Yazer MH, Triulzi DJ. Immune hemolysis following ABO-mismatched stem cell or solid organ transplantation. *Curr Opin Hematol.* 2007; 14:664–670.
98. Nadarajah L, Ashman N, Thuraisingham R, Barber C, Allard S, Green L. Literature review of passenger lymphocyte syndrome following renal transplantation and two case reports. *Am J Transpl.* 2013;13:1594–1600.
99. Rowley SD, Liang PS, Ulz L. Transplantation of ABO-incompatible bone marrow and peripheral blood stem cell components. *Bone Marrow Transplant.* 2000;26:749–757.
100. Rourk A, Squires JE. Implications of the Kidd blood group system in renal transplantation. *Immunohematology.* 2012;28:90–94.

101. Lerut E, Van Damme B, Noizat-Pirenne F, Emonds MP, Rouger P, Vanrenterghem Y, et al. Duffy and Kidd blood group antigens: minor histocompatibility antigens involved in renal allograft rejection? *Transfusion.* 2007;47:28–40.
102. Sands JM, Gargus JJ, Fröhlich O, Gunn RB, Kokko JP. Urinary concentrating ability in patients with Jk(a–b–) blood type who lack carrier-mediated urea transport. *J Am Soc Nephrol.* 1992;2:1689–1696.
103. Bruce LJ, Tanner MJ. Erythroid band 3 variants and disease. *Baillieres Best Pract Res Clin Haematol.* 1999;12:637–654.
104. Danek A, Rubio JP, Rampoldi L, Ho M, Dobson-Stone C, Tison F, et al. McLeod neuroacanthocytosis: genotype and phenotype. *Ann Neurol.* 2001;50:755–764.
105. Kagan A, Feld S, Chemke J, Bar-Khayim Y. Occurrence of hereditary nephritis, pretibial epidermolysis bullosa and beta-thalassemia minor in two siblings with end-stage renal disease. *Nephron.* 1988;49:331–332.
106. Goheen MM, Campino S, Cerami C. The role of the red blood cell in host defence against falciparum malaria: an expanding repertoire of evolutionary alterations. *Br J Haematol.* 2017;179:543–556.
107. Sussman L, Miller E. Un nouveau facteur sanguin «Vel». *Rev Hemat.* 1952;7:368–371.
108. Issitt PD, Anstee DJ. Applied blood group serology.4th ed. Durham: Montgomery Scientific Publications; 1998.
109. Storry JR, Peyrard T. The Vel blood group system: a review. *Immunohematology.* 2017;33:56–59.
110. Cedergren B, Giles CM, Ikin EW. The Vel blood group in northern Sweden. *Vox Sang.* 1976;31:344–355.
111. Chandanayingyong D, Sasaki TT, Greenwalt TJ. Blood groups of the Thais. *Transfusion.* 1967;7:269–276.
112. Alfred BM, Stout TD, Lee M, Birkbeck J, Petrakis NL. Blood groups, phosphoglucomutase, and cerumen types of the Anaham (Chilcotin) Indians. *Am J Phys Anthropol.* 1970;32:329–337.
113. Jongruamklang P, Gassner C, Meyer S, Kummasook A, Darlison M, Boonlum C, et al. Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry analysis of 36 blood group alleles among 396 Thai samples reveals region-specific variants. *Transfusion.* 2018. doi:.
114. Haer-Wigman L, Stegmann TC, Solati S, Ait Soussan A, Beckers E, van der Harst P, et al. Impact of genetic variation in the *SMIM1* gene on Vel expression levels. *Transfusion.* 2015;55:1457–1466.

115. Storry J. Five new blood group systems – what next? *ISBT Sci Ser.* 2014; 9:136–140.
116. Storry JR. Investigation into the carrier molecule of the Vel blood group antigen. *Transfusion.* 2010;50:28A.
117. Moise KJ, Morales Y, Bertholf MF, Rossmann SN, Bai Y. Anti-Vel alloimmunization and severe hemolytic disease of the fetus and newborn. *Immunohematology.* 2017;33:151–154.
118. Haer-Wigman L, de Haas M, van der Schoot C. The immune response to the Vel antigen is HLA-class II *DRB1*11* restricted. *Vox Sang.* 2013; 105:29.
119. Varki A. Essentials of Glycobiology.2nd ed. Cold Spring Harbor: Cold Spring Harbor Laboratory Press; 2009.
120. Rowe JA, Handel IG, Thera MA, Deans AM, Lyke KE, Koné A, et al. Blood group O protects against severe *Plasmodium falciparum* malaria through the mechanism of reduced rosetting. *Proc Natl Acad Sci U S A.* 2007;104:17471–17476.
121. Cserti CM, Dzik WH. The ABO blood group system and *Plasmodium falciparum* malaria. *Blood.* 2007;110:2250–2258.
122. Wang CW, Hviid L. Rifins, rosetting, and red blood cells. *Trends Parasitol.* 2015;31:285–286.
123. Barragan A, Kremsner PG, Wahlgren M, Carlson J. Blood group A antigen is a coreceptor in *Plasmodium falciparum* rosetting. *Infect Immun.* 2000;68:2971–2975.
124. Goel S, Palmkvist M, Moll K, Joannin N, Lara P, Akhouri RR, et al. RIFINs are adhesins implicated in severe *Plasmodium falciparum* malaria. *Nat Med.* 2015;21:314–317.
125. Renton PH, Hancock JA. Uptake of A and B antigens by transfused group O erythrocytes. *Vox Sang.* 1962;7:33–38.
126. Hult AK, Dykes JH, Storry JR, Olsson ML. A and B antigen levels acquired by group O donor-derived erythrocytes following ABO-non-identical transfusion or minor ABO-incompatible haematopoietic stem cell transplantation. *Transfus Med.* 2017;27:181–191.
127. Hult AK, Crottet SL, Storry J, Hellberg Å, Hustinx H, Olsson ML. Investigation into an ABO discrepancy with an unexpected answer in the Lewis system. *Transfusion.* 2017;57:149A.
128. Lairson LL, Henrissat B, Davies GJ, Withers SG. Glycosyltransferases: structures, functions, and mechanisms. *Annu Rev Biochem.* 2008; 77:521–555.

129. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 2014;42:D490–D495.
130. Carbohydrate-Active enZYmes Database. GlycosylTransferase family classification. [Internet]. Université d'Aix-Marseille; 2018 [cited 2018 Apr 23]. Available from: http://www.cazy.org/GlycosylTransferases.html.
131. Liu J, Mushegian A. Three monophyletic superfamilies account for the majority of the known glycosyltransferases. *Protein Sci.* 2003;12:1418–1431.
132. Halim A, Brinkmalm G, Rüetschi U, Westman-Brinkmalm A, Portelius E, Zetterberg H, et al. Site-specific characterization of threonine, serine, and tyrosine glycosylations of amyloid precursor protein/amyloid beta-peptides in human cerebrospinal fluid. *Proc Natl Acad Sci U S A.* 2011; 108:11848–11853.
133. Apweiler R, Hermjakob H, Sharon N. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim Biophys Acta.* 1999;1473:4–8.
134. Dubé S, Fisher JW, Powell JS. Glycosylation at specific sites of erythropoietin is essential for biosynthesis, secretion, and biological function. *J Biol Chem.* 1988;263:17516–17521.
135. Jennewein MF, Alter G. The immunoregulatory roles of antibody glycosylation. *Trends in immunology.* 2017;38:358–372.
136. Collin M, Olsén A. EndoS, a novel secreted protein from Streptococcus pyogenes with endoglycosidase activity on human IgG. *EMBO J.* 2001; 20:3046–3055.
137. Allhorn M, Briceño JG, Baudino L, Lood C, Olsson ML, Izui S, et al. The IgG-specific endoglycosidase EndoS inhibits both cellular and complement-mediated autoimmune hemolysis. *Blood.* 2010;115:5080–5088.
138. Collin M, Björck L. Toward clinical use of the IgG specific enzymes IdeS and EndoS against antibody-mediated diseases. *Methods Mol Biol.* 2017; 1535:339–351.
139. Hennet T, Cabalzar J. Congenital disorders of glycosylation: a concise chart of glycocalyx dysfunction. *Trends Biochem Sci.* 2015;40:377–384.
140. Hellberg Å, Westman JS, Olsson ML. An update on the GLOB blood group system and collection. *Immunohematology.* 2013;29:19–24.

141. Lindström K, Von Dem Borne AE, Breimer ME, Cedergren B, Okubo Y, Rydberg L, et al. Glycosphingolipid expression in spontaneously aborted fetuses and placenta from blood group p women. Evidence for placenta being the primary target for anti-Tj^a-antibodies. *Glycoconjugate J.* 1992; 9:325–329.
142. Hult AK, Olsson ML. Many genetically defined ABO subgroups exhibit characteristic flow cytometric patterns. *Transfusion.* 2010;50:308–323.
143. Cantor AB, Orkin SH. Transcriptional regulation of erythropoiesis: an affair involving multiple partners. *Oncogene.* 2002;21:3368–3376.
144. Love PE, Warzecha C, Li L. Ldb1 complexes: the new master regulators of erythroid gene transcription. *Trends Genet.* 2014;30:1–9.
145. Tournamille C, Colin Y, Cartron JP, Le Van Kim C. Disruption of a GATA motif in the *Duffy* gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nat Genet.* 1995;10:224–228.
146. Wertheimer SP, Barnwell JW. *Plasmodium vivax* interaction with the human Duffy blood group glycoprotein: identification of a parasite receptor-like protein. *Exp Parasitol.* 1989;69:340–350.
147. Haynes JD, Dalton JP, Klotz FW, McGinniss MH, Hadley TJ, Hudson DE, et al. Receptor-like specificity of a *Plasmodium knowlesi* malarial protein that binds to Duffy antigen ligands on erythrocytes. *J Exp Med.* 1988;167:1873–1881.
148. Sano R, Nakajima T, Takahashi K, Kubo R, Kominato Y, Tsukada J, et al. Expression of ABO blood-group genes is dependent upon an erythroid cell-specific regulatory element that is deleted in persons with the B_m phenotype. *Blood.* 2012;119:5301–5310.
149. Takahashi Y, Isa K, Sano R, Nakajima T, Kubo R, Takahashi K, et al. Deletion of the RUNX1 binding site in the erythroid cell-specific regulatory element of the ABO gene in two individuals with the A_m phenotype. *Vox Sang.* 2014;106:167–175.
150. Quill E. Blood-matching goes genetic. *Science.* 2008;319:1478–1479.
151. Reid ME, Rios M, Powell VI, Charles-Pierre D, Malavade V. DNA from blood samples can be used to genotype patients who have recently received a transfusion. *Transfusion.* 2000;40:48–53.
152. Fasano RM, Chou ST. Red blood cell antigen genotyping for sickle cell disease, thalassemia, and other transfusion complications. *Transfus Med Rev.* 2016;30:197–201.

153. Belsito A, Magnussen K, Napoli C. Emerging strategies of blood group genotyping for patients with hemoglobinopathies. *Transfus Apher Sci.* 2017;56:206–213.
154. Finning K, Martin P, Summers J, Daniels G. Fetal genotyping for the K (Kell) and Rh C, c, and E blood groups on cell-free fetal DNA in maternal plasma. *Transfusion.* 2007;47:2126–2133.
155. Rieneck K, Clausen FB, Dziegiel MH. Noninvasive antenatal determination of fetal blood group using next-generation sequencing. *Cold Spring Harb Perspect Med.* 2015;6:a023093.
156. Storry JR, Olsson ML, Reid ME. Application of DNA analysis to the quality assurance of reagent red blood cells. *Transfusion.* 2007;47:73S–78S.
157. Storry J, Wikman A. Handbok för Blodcentraler. Kapitel 8: Erytrocytantikroppsidentifiering. [Internet]. Svensk Förening för Transfusionsmedicin; 2018 [cited 2018 May 2]. Available from: http://www.kitm.se/sv/kap-8-3-0/.
158. Johnson PH, Hopkinson DA. Detection of ABO blood group polymorphism by denaturing gradient gel electrophoresis. *Hum Mol Genet.* 1992;1:341–344.
159. Ugozzoli L, Wallace RB. Application of an allele-specific polymerase chain reaction to the direct determination of ABO blood group genotypes. *Genomics.* 1992;12:670–674.
160. Olsson ML, Chester MA. Polymorphism and recombination events at the *ABO* locus: a major challenge for genomic ABO blood grouping strategies. *Transfus Med.* 2001;11:295–313.
161. Olsson ML, Chester MA. A rapid and simple ABO genotype screening method using a novel B/O^2 versus A/O^2 discriminating nucleotide substitution at the *ABO* locus. *Vox Sang.* 1995;69:242–247.
162. Hosseini-Maaf B, Hellberg A, Chester MA, Olsson ML. An extensive polymerase chain reaction-allele-specific polymorphism strategy for clinical ABO blood group genotyping that avoids potential errors caused by null, subgroup, and hybrid alleles. *Transfusion.* 2007;47:2110–2125.
163. Clausen FB, Krog GR, Rieneck K, Nielsen LK, Lundquist R, Finning K, et al. Reliable test for prenatal prediction of fetal RhD type using maternal plasma from RhD negative women. *Prenat Diagn.* 2005;25:1040–1044.

164. Beiboer SHW, Wieringa-Jelsma T, Maaskant-Van Wijk PA, van der Schoot CE, van Zwieten R, Roos D, et al. Rapid genotyping of blood group antigens by multiplex polymerase chain reaction and DNA microarray hybridization. *Transfusion.* 2005;45:667–679.
165. Hashmi G, Shariff T, Seul M, Vissavajjhala P, Hue-Roye K, Charles-Pierre D, et al. A flexible array format for large-scale, rapid blood group DNA typing. *Transfusion.* 2005;45:680–688.
166. Denomme GA, Van Oene M. High-throughput multiplex single-nucleotide polymorphism analysis for red cell and platelet antigen genotypes. *Transfusion.* 2005;45:660–666.
167. Bugert P, McBride S, Smith G, Dugrillon A, Klüter H, Ouwehand WH, et al. Microarray-based genotyping for blood groups: comparison of gene array and 5′-nuclease assay techniques with human platelet antigen as a model. *Transfusion.* 2005;45:654–659.
168. Avent ND, Martinez A, Flegel WA, Olsson ML, Scott ML, Nogués N, et al. The BloodGen project: toward mass-scale comprehensive genotyping of blood donors in the European Union and beyond. *Transfusion.* 2007; 47:40S–46S.
169. Avent ND, Martinez A, Flegel WA, Olsson ML, Scott ML, Nogués N, et al. The Bloodgen Project of the European Union, 2003–2009. *Transfus Med Hemother.* 2009;36:162–167.
170. Avent ND. Large scale blood group genotyping. *Transfus Clin Biol.* 2007; 14:10–15.
171. Moulds JM. Future of molecular testing for red blood cell antigens. *Clin Lab Med.* 2010;30:419–429.
172. Denomme GA, Schanen MJ. Mass-scale donor red cell genotyping using real-time array technology. *Immunohematology.* 2015;31:69–74.
173. Goldman M, Núria N, Castilho LM. An overview of the Progenika ID CORE XT: an automated genotyping platform based on a fluidic microarray system. *Immunohematology.* 2015;31:62–68.
174. Finning K, Bhandari R, Sellers F, Revelli N, Villa MA, Muñiz-Díaz E, et al. Evaluation of red blood cell and platelet antigen genotyping platforms (ID CORE XT/ID HPA XT) in routine clinical practice. *Blood Transfus.* 2016;14:160–167.
175. Gassner C, Meyer S, Frey BM, Vollmert C. Matrix-assisted laser desorption/ionisation, time-of-flight mass spectrometry-based blood group genotyping–the alternative approach. *Transfus Med Rev.* 2013;27:2–9.

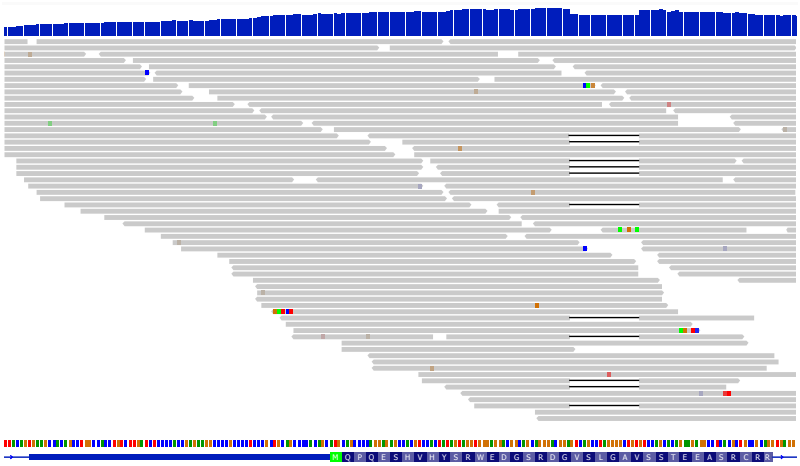
176. Meyer S, Vollmert C, Trost N, Brönnimann C, Gottschalk J, Buser A, et al. High-throughput Kell, Kidd, and Duffy matrix-assisted laser desorption/ionization, time-of-flight mass spectrometry-based blood group genotyping of 4000 donors shows close to full concordance with serotyping and detects new alleles. *Transfusion.* 2014;54:3198–3207.
177. Liu Z, Liu M, Mercado T, Illoh O, Davey R. Extended blood group molecular typing and next-generation sequencing. *Transfus Med Rev.* 2014; 28:177–186.
178. McBean RS, Hyland CA, Flower RL. Approaches to determination of a full profile of blood group genotypes: single nucleotide variant mapping and massively parallel sequencing. *Comput Struct Biotechnol J.* 2014; 11:147–151.
179. Tilley L, Grimsley S. Is Next Generation Sequencing the future of blood group testing? *Transfus Apher Sci.* 2014;50:183–188.
180. Fichou Y, Férec C. NGS and blood group systems: State of the art and perspectives. *Transfus Clin Biol.* 2017;24:240–244.
181. Fichou Y, Audrézet MP, Guéguen P, Le Maréchal C, Férec C. Next-generation sequencing is a credible strategy for blood group genotyping. *Br J Haematol.* 2014;167:554–562.
182. Fichou Y, Mariez M, Le Maréchal C, Férec C. The experience of extended blood group genotyping by next-generation sequencing (NGS): investigation of patients with sickle-cell disease. *Vox Sang.* 2016;111:418–424.
183. Lane WJ, Westhoff CM, Uy JM, Aguad M, Smeland-Wagman R, Kaufman RM, et al. Comprehensive red blood cell and platelet antigen prediction from whole genome sequencing: proof of principle. *Transfusion.* 2016;56:743–754.
184. Chou ST, Flanagan JM, Vege S, Luban NLC, Brown RC, Ware RE, et al. Whole-exome sequencing for *RH* genotyping and alloimmunization risk in children with sickle cell anemia. *Blood Adv.* 2017;1:1414–1422.
185. Jakobsen MA, Dellgren C, Sheppard C, Yazer M, Sprogøe U. The use of next-generation sequencing for the determination of rare blood group genotypes. *Transfus Med*. 2017. doi: 10.1111/tme.12496.
186. Patnaik SK, Helmberg W, Blumenfeld OO. BGMUT: NCBI dbRBC database of allelic variations of genes encoding antigens of blood group systems. *Nucleic Acids Res.* 2012;40:D1023–D1029.

187. Schoeman EM, Roulis EV, Liew YW, Martin JR, Powley T, Wilson B, et al. Targeted exome sequencing defines novel and rare variants in complex blood group serology cases for a red blood cell reference laboratory setting. *Transfusion.* 2018;58:284–293.
188. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature.* 2010;467:1061–1073.
189. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature.* 2012;491:56–65.
190. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature.* 2015;526:68–74.
191. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature.* 2007;449:851–861.
192. UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature.* 2015;526:82–90.
193. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet.* 2014;46:818–825.
194. Ameur A, Dahlberg J, Olason P, Vezzi F, Karlsson R, Martin M, et al. SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population. *Eur J Hum Genet.* 2017;25:1253–1260.
195. Turnbull C, Scott RH, Thomas E, Jones L, Murugaesu N, Pretty FB, et al. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ.* 2018;361:k1687.
196. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25:1754–1760.
197. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009; 25:2078–2079.
198. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics.* 2011;27:2156–2158.
199. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature.* 2001;409:860–921.
200. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science.* 2001;291:1304–1351.
201. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489:57–74.

202. Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, Kundaje A, et al. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.* 2012;13:R48.
203. Casper J, Zweig AS, Villarreal C, Tyner C, Speir ML, Rosenbloom KR, et al. The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res.* 2018;46:D762–D769.
204. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. *Nucleic Acids Res.* 2018;46:D754–D761.
205. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* 2016;536:285–291.
206. Hashimoto K, Goto S, Kawano S, Aoki-Kinoshita KF, Ueda N, Hamajima M, et al. KEGG as a glycome informatics resource. *Glycobiology.* 2006;16:63R–70R.
207. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 2018;46:2699.
208. Laird NM, Lange C. The Fundamentals of Modern Statistical Genetics.New York: Springer Science; 2011.
209. Fadista J, Manning AK, Florez JC, Groop L. The (in)famous GWAS *P*-value threshold revisited and updated for low-frequency variants. *Eur J Hum Genet.* 2016;24:1202–1205.
210. Bacanu SA, Devlin B, Roeder K. Association studies for quantitative traits in structured populations. *Genet Epidemiol.* 2002;22:78–93.
211. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006;38:904–909.
212. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet.* 2010;11:459–463.
213. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet.* 2012;44:821–824.
214. Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat Genet.* 2018;50:524–537.
215. Ott J, Kamatani Y, Lathrop M. Family-based designs for genome-wide association studies. *Nat Rev Genet.* 2011;12:465–474.

216. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. *Nat Protoc.* 2010;5:1564–1573.
217. Ott J, Wang J, Leal SM. Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet.* 2015;16:275–284.
218. Li M, Boehnke M, Abecasis GR. Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. *Am J Hum Genet.* 2006;78:778–792.
219. Slatkin M. Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nat Rev Genet.* 2008;9:477–485.
220. Lewontin RC. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics.* 1964;49:49–67.
221. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. *Annu Rev Genomics Hum Genet.* 2009;10:387–406.
222. Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy RT, et al. Best practices for scientific computing. *PLoS Biol.* 2014; 12:e1001745.
223. Russ WP, Engelman DM. The GxxxG motif: a framework for transmembrane helix-helix association. *J Mol Biol.* 2000;296:911–919.
224. Arnaud L, Kelley LP, Helias V, Cartron JP, Ballif BA. SMIM1 is a type II transmembrane phosphoprotein and displays the Vel blood group antigen at its carboxyl-terminus. *FEBS letters.* 2015;589:3624–3630.
225. Nylander A, Storry JR, Olsson ML. The ins and outs of SMIM1 and its relationship to the expression of Vel blood group antigen. *Vox Sang.* 2017;112:48–49.
226. Pilling LC, Atkins JL, Duff MO, Beaumont RN, Jones SE, Tyrrell J, et al. Red blood cell distribution width: Genetic evidence for aging pathways in 116,666 volunteers. *PLoS One.* 2017;12:e0185083.
227. Evans DM, Zhu G, Dy V, Heath AC, Madden PAF, Kemp JP, et al. Genome-wide association study identifies loci affecting blood copper, selenium and zinc. *Hum Mol Genet.* 2013;22:3998–4006.
228. van der Harst P, Zhang W, Mateo Leach I, Rendon A, Verweij N, Sehmi J, et al. Seventy-five genetic loci influencing the human red blood cell. *Nature.* 2012;492:369–375.
229. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001;29:308–311.

230. Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, et al. Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell.* 2016;165:1530–1545.
231. MacArthur JAL, Morales J, Tully RE, Astashyn A, Gil L, Bruford EA, et al. Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants. *Nucleic Acids Res.* 2014;42:D873–D878.
232. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, et al. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* 2005;6:R44.
233. Wagner FF, Gassner C, Müller TH, Schönitzer D, Schunter F, Flegel WA. Molecular basis of weak D phenotypes. *Blood.* 1999;93:385–393.
234. Olsson ML, Smythe JS, Hansson C, Poole J, Mallinson G, Jones J, et al. The Fy^x phenotype is associated with a missense mutation in the Fy^b allele predicting Arg89Cys in the Duffy glycoprotein. *Br J Haematol.* 1998; 103:1184–1191.
235. Möller M, Hellberg Å, Olsson ML. Thorough analysis of unorthodox *ABO* deletions called by the 1000 Genomes project. *Vox Sang.* 2018; 113:185–197.
236. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. *Genome Biol.* 2016;17:122.
237. Ludvigsson JF, Almqvist C, Bonamy AKE, Ljung R, Michaëlsson K, Neovius M, et al. Registers of the Swedish total population and their use in medical research. *Eur J Epidemiol.* 2016;31:125–136.
238. Edgren G, Rostgaard K, Vasan SK, Wikman A, Norda R, Pedersen OB, et al. The new Scandinavian Donations and Transfusions database (SCANDAT2): a blood safety resource with added versatility. *Transfusion.* 2015; 55:1600–1606.



Printed by Media-Tryck, Lund 2018 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

FACULTY OF MEDICINE

Division of Hematology and Transfusion Medicine
Department of Laboratory Medicine

Lund University, Faculty of Medicine
Doctoral Dissertation Series 2018:92
ISBN 978-91-7619-658-8
ISSN 1652-8220