Elucidating the Blood Group Regulome

Wu, Ping Chun

2024

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

Total number of authors:
1

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Elucidating the Blood Group Regulome

PING CHUN (GLORIA) WU

DEPARTMENT OF LABORATORY MEDICINE | FACULTY OF MEDICINE | LUND UNIVERSITY

Elucidating the Blood Group Regulome

Elucidating the Blood Group Regulome

Ping Chun (Gloria) Wu

DOCTORAL DISSERTATION

Doctoral dissertation for the degree of Doctor of Philosophy (PhD) at the
Faculty of Medicine at Lund University to be publicly defended
on 19th of April at 13.00 in Segerfalksalen, BMC A10, Lund, Sweden

Faculty opponent
Associate Professor Peter Bugert, Ph.D.

Medical Faculty Mannheim
Heidelberg University, Mannheim, Germany

Organization: LUND UNIVERSITY			
Document name:	**Doctoral Dissertation**	**Date of issue:2024-03-12**	
Author(s):	**Ping Chun (Gloria) Wu**	**Sponsoring organization:**	

Title and subtitle: Elucidating the Blood Group Regulome

Abstract:

Matching for blood group antigens in immunized patients is essential for the safety of blood transfusion. However, variable antigen expression can make serological blood typing difficult or result in a falsely negative crossmatch. Although the genetic background for most blood group antigens is now established, the genetics and mechanisms underlying the varying antigen levels are not well understood. The overall aim of this thesis is to provide a big data analysis approach to identify blood group gene regulatory regions systematically and to pave the way for real-world applications, thus improving precision in genotypic blood group predictions.

By setting up a dedicated bioinformatics pipeline, we were able to identify 193 candidate regulatory regions containing GATA1 binding sites across 33 blood group genes. As a proof of principle, we further defined two GATA1 binding sites in intron 4 of the *CR1* gene (Knops, ISBT 022) as a functional enhancer to boost gene transcription. Through cohort studies using multiple methods including genotyping, a gene expression assay and protein expression analysis via flow cytometry and western blot, we found that the minor allele of rs11117991C resulted in loss of GATA1 binding and is the genetic basis underlying the very low expression of CR1 on red blood cells, known as the Helgeson phenotype. However, rs11117991C has a very low allele frequency in the African population. Thus, a list of other potential regulatory transcription factor (TF) binding sites and SNVs altering motifs in *CR1* were presented to investigate a possible mechanism behind the Helgeson phenotype in African populations.

The analysis was also expanded to other erythroid TFs: KLF1, RUNX1, NFE2 and also histone modifications indicating enhancers, promoters and open chromatin regions. The analysis identified a total of 814 candidate regulatory sites within 47 blood group genomic regions and opens up the possibility to study the interaction between TFs at sites showing co-occupancy. Co-occupancy of all four TFs was identified in four blood group genes only. Various levels of transcript activity at these sites were observed as follows: *CR1* and *ABCC4* (PEL, ISBT 040) enhanced, *ABCB6* (LAN, ISBT 033) did not alter transcript levels, and *EMP3* (MAM, ISBT 041) either up- or down-regulated levels depending on the orientation of the element. The *KEL* (Kell, ISBT 006) promoter was further characterized and both GATA1 and KLF1 were found to bind to this region, and the disruption of the corresponding motifs downregulated transcription.

Lastly, the GATA1 regulatory sites in *RHD* (Rh, ISBT 004) promoter intron 1 and intron 2 were examined in weak D samples with normal *RHD* exon sequences. We identified a novel variant that disrupted the GATA1 site in the proximal promoter in one such sample. In addition, two samples were found to be chimeras carrying both RhD+ and RhD– cell populations, thus masquerading as weakly expressing RhD.

In conclusion, this study provides both concrete findings, tools and insight for future studies regarding the regulatory landscape governing blood group antigen expression by developing and applying a systematic big data analysis approach with the overarching goal to improve blood group typing.

Key words: blood group, gene regulation, transcription factors, erythrocyte, transfusion medicine

Classification system and/or index terms (if any)	Supplementary bibliographical information
Language English	**ISSN and key title**: 1652-8220
ISBN: 978-91-8021-542-8	
Recipient's notes	**Number of pages**:102
Price	Security classification

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature	Date 2024-03-11

Elucidating the Blood Group Regulome

Ping Chun (Gloria) Wu

LUND UNIVERSITY

Coverphoto create by AI-generated art *NightCafe Studio* with the text prompt “DNA double helix fade out into the background of dark blue starry night” in Candy style.

Faculty of Medicine

Department of Laboratory Medicine

ISBN 978-91-8021-542-8

ISSN 1652-8220

Lund University, Faculty of Medicine Doctoral Dissertation Series 2024:49

Printed in Sweden by Media-Tryck, Lund University

Lund 2024

NORDIC SWAN ECOLABEL
Printed matter
3041 0903

Media-Tryck is a Nordic Swan Ecolabel
certified provider of printed material.
Read more about our environmental
work at www.mediatryck.lu.se

MADE IN SWEDEN

To the One and the ones with abounding grace

A land of waterbrooks, of springs and of fountains,
flowing forth in valleys and in mountains;

A land of wheat and barley and vines and fig trees and
pomegranates; a land of olive trees with oil and of honey.

Deuteronomy 8:7-8 (recovery version)

Table of Contents

Abstract

Matching for blood group antigens in immunized patients is essential for the safety of blood transfusion. However, variable antigen expression can make serological blood typing difficult or result in a falsely negative crossmatch. Although the genetic background for most blood group antigens is now established, the genetics and mechanisms underlying the varying antigen levels are not well understood. The overall aim of this thesis is to provide a big data analysis approach to identify blood group gene regulatory regions systematically and to pave the way for real-world applications, thus improving precision in genotypic blood group predictions.

By setting up a dedicated bioinformatics pipeline, we were able to identify 193 candidate regulatory regions containing GATA1 binding sites across 33 blood group genes. As a proof of principle, we further defined two GATA1 binding sites in intron 4 of the *CR1* gene (Knops, ISBT 022) as a functional enhancer to boost gene transcription. Through cohort studies using multiple methods including genotyping, a gene expression assay and protein expression analysis via flow cytometry and western blot, we found that the minor allele of rs11117991C resulted in loss of GATA1 binding and is the genetic basis underlying the very low expression of CR1 on red blood cells, known as the Helgeson phenotype. However, rs11117991C has a very low allele frequency in the African population. Thus, a list of other potential regulatory transcription factor (TF) binding sites and SNVs altering motifs in *CR1* were presented to investigate a possible mechanism behind the Helgeson phenotype in African populations.

The analysis was also expanded to other erythroid TFs: KLF1, RUNX1, NFE2 and also histone modifications indicating enhancers, promoters and open chromatin regions. The analysis identified a total of 814 candidate regulatory sites within 47 blood group genomic regions and opens up the possibility to study the interaction between TFs at sites showing co-occupancy. Co-occupancy of all four TFs was identified in four blood group genes only. Various levels of transcript activity at these sites were observed as follows: *CR1* and *ABCC4* (PEL, ISBT 040) enhanced, *ABCB6* (LAN, ISBT 033) did not alter transcript levels, and *EMP3* (MAM, ISBT 041) either up- or down-regulated levels depending on the orientation of the element. The *KEL* (Kell, ISBT 006) promoter was further characterized and both GATA1 and KLF1 were found to bind to this region, and the disruption of the corresponding motifs downregulated transcription.

Lastly, the GATA1 regulatory sites in *RHD* (Rh, ISBT 004) promoter intron 1 and intron 2 were examined in weak D samples with normal *RHD* exon sequences. We identified a novel variant that disrupted the GATA1 site in the proximal promoter in one such sample. In addition, two samples were found to be chimeras carrying both RhD+ and RhD– cell populations, thus masquerading as weakly expressing RhD.

In conclusion, this study provides both concrete findings, tools and insight for future studies regarding the regulatory landscape governing blood group antigen expression by developing and applying a systematic big data analysis approach with the overarching goal to improve blood group typing.

Key words: blood group, gene regulation, transcription factors, erythrocyte, transfusion medicine

Populärvetenskaplig sammanfattning på svenska

Precis som människor ser olika ut på utsidan är de också olika på insidan, inte minst på ytan av våra röda blodkroppar. Det finns proteiner och lipider dekorerade med kolhydrater (ja, faktiskt socker), och även bara proteiner utan kolhydrater, på ytan av de röda blodkropparna. Dessa molekyler skiljer lite från person till person (såvida du inte har en enäggstvilling att jämföra med!). För personer som behöver få en blodtransfusion är det säkrast att ge röda blodkroppar från blodgivare som liknar patientens. Människor kan också ha olika mängder av proteiner och sockerarter på sina röda blodkroppar. Vi vet mycket om varför människor har olika typer av proteiner och sockerarter men betydligt mindre om varför mängderna skiljer sig åt. Inom området transfusionsmedicin kallar vi de olika typerna för blodgrupper, och de kodas av våra blodgruppsgener. I denna avhandling var syftet att studera varför människor har olika antal av dessa blodgruppsproteiner och sockerarter på sina röda blodkroppar.

Det finns en grupp molekyler, så kallade transkriptionsfaktorer, som påverkar produktionen av blodgruppsproteiner på ungefär samma sätt som spenat fungerar för seriefiguren Karl-Alfred. Innan han äter upp sin burk med spenat är han svag och orkar inte så mycket. Efter att han ätit spenaten växer hans muskler och blir större så att han får mycket mer kraft. På liknande sätt är genernas aktivitet svag innan transkriptionsfaktorerna binder till speciella platser i våra blodgruppsgener. När dessa faktorer binder till sin favoritplats i genen "slår de på" genen och ökar produktionen från genen så att den gör mycket mer protein. Den speciella grupp av transkriptionsfaktorer jag studerade här är särskilt bra på att stimulera proteinproduktionen i just röda blodkroppar, inklusive våra blodgruppsproteiner så om faktorerna binder ökar mängden blodgrupp medan väldigt lite produceras om de inte kan binda till genen. De speciella områden av generna som transkriptionsfaktorer behöver binda till kallas regulatoriska regioner och finns av två huvudvarianter: promotorer (som startar upp genens aktivitet) och enhancers (som förstärker genens aktivitet).

Under mina studier har jag försökt kartlägga var de regulatoriska regionerna finns i alla våra blodgruppsgener och var transkriptionsfaktorerna binder. Först tog jag hjälp av datorer för att vägleda mig i arbetet (s.k. bioinformatik). Sedan bekräftade jag med experiment i laboratoriet om datorförutsägelserna var korrekta. Jag har också undersökt vad som händer om reglerande regioner ändras. Genom att i detalj studera var transkriptionsfaktorer binder i dessa regulatoriska regioner löste vi ett halvt sekel gammalt mysterium, orsaken till den s.k. Helgeson-fenotypen. I denna fenotyp ses ovanligt låga mängder av ett specifikt protein som kallas komplementreceptor 1 (CR1) på röda blodkroppar. Vi fann att bindning av en specifik transkriptionsfaktor vid namn GATA1 till *CR1*-genen resulterade i en hög mängd CR1-protein. Men när den regulatoriska regionen är ändrad (som är fallet hos vissa människor) känner inte GATA1 igen sin landningsplats och kan därför inte längre binda till den. Detta orsakar Helgeson-fenotypen då CR1-nivåerna är nästan oupptäckbara i blodbankslaboratoriet.

Jag studerade också fler transkriptionsfaktorer som är involverade i produktionen av mogna röda blodkroppar. Genom att studera var i blodgruppsgenernas reglerande regioner som flera transkriptionsfaktorer binder tillsammans lärde vi oss att gener reagerar olika trots att samma grupp av transkriptionsfaktorer samverkar. Vissa gener reagerade med högre aktivitet, andra visade ingen förändring och ytterligare andra visade ökning eller minskning av aktiviteten beroende på vilken riktning av genen som transkriptionsfaktorerna band. Tack vare detta delarbete kunde vi också förstå mycket mer om promotorn för en kliniskt viktig blodgruppsgen som heter *KEL*.

Forskningen som presenteras i den här boken hjälpte också till att identifiera en ny förändring i DNA-sekvensen hos en person som minskade inbindningen av transkriptionsfaktorn GATA1, vilket resulterade i ett mycket svagt uttryck av en av de viktigaste blodgrupperna, RhD (den som folk säger är positiv om du har den och negativ om du saknar den).

Ju mer vi lär oss om var och hur transkriptionsfaktorerna binder, och om det sker förändringar i de reglerande regionerna av blodgruppsgenerna, desto bättre blir vi på blodgruppstypning med hög precision så att vi ger så väl matchat blod från donatorer som möjligt till patienterna för att skydda dem.

Låt oss klarlägga blodgruppernas reglering!

Popular Summary in English

People look different in their external appearances and their internal make-up is also different, including that of our red blood cells (RBCs). There are proteins and lipids coated with carbohydrates (yes, sugars), and also just proteins without coating, on the surface of the RBCs. These molecules are slightly different from person to person (unless you have an identical twin to compare with!). For people who need to receive a blood transfusion, it is safest to match the same-looking RBCs between the blood donor and the patient. People can also have different amounts of proteins and sugars on their RBCs. We know much about why people have different types of proteins and sugars but less about why the amounts differ. Here, we wanted to study why people have different numbers of these proteins and sugars on their RBCs. In the field of transfusion medicine, we call these differences blood group antigens, and they are encoded by our blood group genes.

There are some molecules, called transcription factors, that act on the production of the blood group proteins as spinach works for Popeye the sailorman. Before Popeye eats his can of spinach, he is weak and has not much strength. After he eats the spinach, his muscles grow bigger and he gains much more power. In a similar way, before the transcription factors bind to a special place in our blood group genes, the activity of the gene is really weak. However, when these factors bind to their favorite place in the gene, they "turn on" the gene, and boost the production from the gene to make much more protein. The special group of transcription factors I studied here are particularly good at stimulating protein production in RBCs, including our blood group proteins, so if the factors bind the amount of blood group increases while very little is produced if they cannot bind to the gene. The special areas of the genes where transcription factors need to bind to are called regulatory regions and come in two main variants: promoters and enhancers.

During my studies, I have set out to map where the regulatory regions are located in all our blood group genes and where the transcription factors bind. First, I used computer power to guide me (so-called bioinformatics). Then, I confirmed with experiments in the lab if the computer predictions were correct. I have also investigated what happens if the regulatory region is altered. By studying in detail where transcription factors bind in these regulatory regions, we solved a half century old mystery, the Helgeson phenotype. Helgeson phenotype RBCs have an unusually low amount of a specific protein called Complement receptor 1 (CR1). We found that binding of a specific transcription factor named GATA1 to the *CR1* gene resulted in a high amount of CR1. But when the regulatory region is changed, GATA1 does not recognize its landing place and can therefore no longer bind to it. This causes the Helgeson phenotype, in which CR1 is almost undetectable to blood bankers.

I also studied more transcription factors that are involved in the production of mature RBCs. By studying different blood group gene regions bound by the transcription factors, we learned that not all genes react the same to the same group of transcription factors, some genes react with higher gene activity, others showed no change and yet others showed up- or down-regulation that depended on what direction of the gene the transcription factors bound. I was also able to understand much more about the promoter of a clinically important blood group gene called *KEL*.

The research presented in this book also helped to identify a novel change in the DNA sequence of one person that made it less recognizable for transcription factor GATA1 to bind, which resulted in a very weak expression of one of the most important blood group antigens, RhD (the one which people say is positive if you have it and negative if you lack it).

The more we learn about where and how the transcription factors bind, and if there are changes in the regulatory regions of the blood group genes, the better we become at blood group typing with high precision so that we provide as similar RBCs from donors as possible to patients and keep them safe.

Let’s elucidate the blood group regulome!

中文科普摘要

人們的外表不同，就連紅血球也長得不一樣。而在需要輸血的時候，最安全的做法是能為患者匹配外觀相同的紅血球。紅血球表面有不同數量的蛋白質跟脂質，而且有的有碳水化合物包覆住（是的，也就是糖份），有的則沒有。我們對人們的紅血球上為什麼有不同類型的蛋白質和糖了解很多，但對它們為什麼存在不同的數量則知之甚少。這個研究即是針對後者進行更深入的探討。

在輸血醫學領域中，我們將這些長得不一樣蛋白質及醣類稱為血型抗原。血型抗原是由血型基因所控制。血型基因會產生血型蛋白質，至於數量的關鍵，則在於「轉錄因子」。

有一些稱為轉錄因子對於血型蛋白質的產生，就像大力水手卜派的菠菜一樣。大力水手吃菠菜罐頭之前很虛弱、沒有太多力氣。但吃了菠菜以後他的肌肉就變大了，力量也增強了。同理，血型基因在透過特殊的位置與轉錄因子結合之前，它的活性很弱；但結合之後，這個血型基因的產線就會被「打開」，血型蛋白質的產量就會大幅上升。

這份研究發現一組特殊的轉錄因子特別擅長刺激紅血球中的蛋白質產生，包括血型蛋白質。如果這些因子結合，特定血型蛋白質的數量就會增加，反之數量就會很少。轉錄因子需要與基因結合的特殊區域稱為調控區，主要含有兩種區域：啟動子(promoter)和增強子(enhancer)。研究過程首先利用電腦來預測特定調控區在血型基因中的位置以及轉錄因子的結合位置，這部分屬於生物資訊學的範疇，再透過實驗證實電腦的預測是否正確。除此之外，還進一步瞭解改變調控區域會發生什麼事。

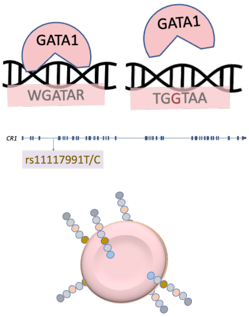
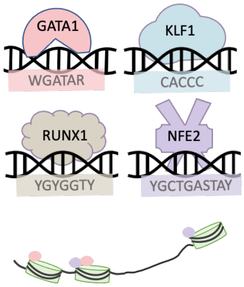
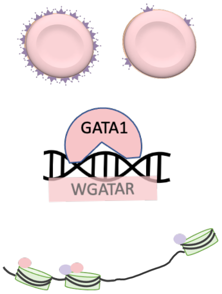
透過詳細研究轉錄因子在這些調控區中的結合位置，我們解決了半個世紀以來的謎團，即 Helgeson 表型。Helgeson 表型紅血球中一種稱為 CR1 的蛋白質含量異常地低。而我們發現，一種名為 GATA1 的轉錄因子與 *CR1* 基因結合能使 CR1 的數量大增。當調控區域發生變化、GATA1 無法辨識這個調控區而沒有與其結合，就會導致 Helgeson 表型的產生，使血庫人員幾乎檢測不到紅血球表面的 CR1。除此之外，研究參與成熟紅血球生成的特定轉錄因子及其結合的不同血型基因區域，我們了解到：並非所有血型基因對同一組轉錄因子的反應都相同，有些基因反應出較高的基因活性，有些則沒有變化，而有些則能透過調節轉錄因子與基因結合的方向、而有升高

或降低的表現。藉此，我們對臨床上重要的血型基因 *KEL* 的啟動子有了更深的認識。本研究還發現了一個新的 DNA 序列變化。這種變化使得轉錄因子 GATA1 的結合變得難以識別，從而導致最重要的血型抗原之一 RhD 的表達非常地弱。

愈瞭解轉錄因子結合的位置、方式以及血型基因的調控區的變化，愈能做到更好的高精度血型分型，從而盡可能匹配出愈相似的紅血球給患者以保障他們的安全。

讓我們來一起研究血型的調控吧！

Thesis at a glance

Paper	I	II	III
Graphical abstract		 GATA1 WGATAR KLF1 CACCC RUNX1 YGYGGTY NFE2 YGCTGASTAY	
Aim	To establish the GATA1 blood group regulome and investigate the mechanism behind the low CR1 expressing Helgeson phenotype.	To map the blood group regulome with key erythroid TFs, histone modifications and open chromatin regions.	To investigate GATA1 binding in open regulatory regions of *RHD* in weak RhD samples with normal *RHD* exon sequences.
Key results	◊ Identified 193 GATA1 binding sites in 33 blood group genes. ◊ Identified *CR1* intron 4 as a GATA1-driven enhancer. ◊ Showed that SNV rs11117991T>C disrupts GATA1 binding. ◊ LD found with rs11117991 & previous Helgeson markers for Caucasians but not Africans.	◊ Identified 814 potential TF regulatory sites in 47 blood group genes. ◊ Co-occupancy of 4 TFs at *CR1*, *ABCB6*, *ABCC4* & *EMP3* showed varied effects on transcription. ◊ *KEL* promoter region regulated by GATA1 and KLF1 motifs. ◊ Decreased binding of TFs to *KEL* promoter when motifs are disrupted by natural variants.	◊ GATA1 showed binding to the *RHD* promoter, and regions in intron 1 and intron 2. ◊ *RHD* intron 2 displays enhancer ability in the R^2 haplotype with SNV rs675072G>A. ◊ Novel variant identified in the promoter region (c.1–110A>C) of a Del sample; disrupts GATA1 binding. ◊ RhD–/RhD+ chimeras found in two "weak D" samples.
Abbreviations: CR1, complement receptor 1; LD, linkage disequilibrium, SNV, single nucleotide variant; TF, transcription factor.			

List of Papers

Paper I

Ping Chun Wu, Yan Quan Lee, Mattias Möller, Jill R. Storry and Martin L. Olsson. Elucidation of the low-expressing erythroid CR1 phenotype by bioinformatic mining of the GATA1-driven blood-group regulome.

Nature Communications, 2023; 14(1): 5001
(12 electronic pages + supplements, doi: 10.1038/s41467-023-40708-w).

Paper II

Ping Chun Wu, Eunike C McGowan, Yan Quan Lee, Sudip Ghosh, Jenny Hansson and Martin L Olsson. Epigenetic dissection of human blood group genes reveals unknown regulatory elements and detailed characteristics of *KEL* and other loci.

Manuscript, submitted to *Transfusion*.

Paper III

Eunike C McGowan, **Ping Chun Wu**, Åsa Hellberg, Genghis H Lopez, Catherine A Hyland and Martin L Olsson. A bioinformatically initiated approach to evaluate GATA1 regulatory regions in samples with weak D, Del or D– phenotypes despite normal *RHD* exons.

Manuscript, submitted to *Transfusion Medicine and Hemotherapy.*

Author’s contribution to the papers

Paper I

P.C.W., M.M., and M.L.O. conceived and designed the bioinformatic part of the study, whilst P.C.W., Y.Q.L., J.R.S., and M.L.O. designed the *in vitro* experiments of the study. P.C.W. and M.M. performed the bioinformatic analysis, P.C.W. and Y.Q.L. performed experiments. All authors interpreted bioinformatic and *in vitro* experimental data. Y.Q.L., J.R.S., and M.L.O. supervised the study. P.C.W. wrote the manuscript and all authors read and revised the manuscript.

Paper II

P.C.W. and M.L.O. conceived the study. P.C.W., E.C.M. and S.G. performed experiments. Y.Q.L., J.H. and M.L.O. supervised the study. All authors analyzed and interpreted results. P.C.W. drafted the manuscript and all authors contributed to, revised and approved the final version.

Paper III

E.C.M., P.C.W. and M.L.O. conceived the study. E.C.M. wrote the manuscript, designed primers, performed PCR assays, EMSA and luciferase assay and data analysis, including the sequencing data from PCRs performed by P.C.W. P.C.W. helped write the manuscript, performed the analysis that provided the GATA1 candidates, designed primers, performed PCR assays and constructed vectors for the luciferase assay. Å.H. oversaw the NRLGBT samples and analysis of the D+/D– chimeras. G.H.L. and C.A.H. provided and managed the data from the Lifeblood samples. E.C.M., P.C.W. and M.L.O. interpreted and reviewed the data. M.L.O. supervised the study. All authors contributed to the preparation and review of this manuscript and approved the final version.

Abbreviations

1000G	1000 Genomes Project
AIHA	autoimmune hemolytic anemia
AML	acute myeloid leukemia
ASP	allele-specific primers
basoE	basophilic erythroblast
BFU-E	burst-forming unit-erythroid
bp	base pair
BPG	2,3-biphosphoglycerate
CDA	congenital dyserythropoietic anemia
cDNA	complementary DNA
cffDNA	cell-free fetal DNA
CFU-E	colony-forming unit-erythroid
CFU-GEMM	colony-forming unit-granulocyte, erythrocyte, monocyte and megakaryocyte
CMP	common myeloid progenitor
CR1	complement receptor 1
CTCG	CCCTC-binding factor
DAT	direct antiglobulin test
DHS	DNase hypersensitivity sites
DNA	deoxyribonucleic acid
EBI	erythroblastic islands
EPO	erythropoietin
GPA	glycophorin A
HDFN	hemolytic disease of the fetus and newborn
HSC	hematopoietic stem cell
HTR	hemolytic transfusion reaction
IAT	indirect antiglobulin test
IgG	immunoglobulin G

IgM	immunoglobulin M
LCR	locus control region
LD	linkage disequilibrium
Maf	muscular aponeurotic fibrosarcoma
miRNA	microRNA
MNV	multiple nucleotide variants
ncRNA	non-coding RNA
NGS	next-generation sequencing
orthoE	orthochromatophilic erythroblast
PBM	patient blood management
PCR	polymerase chain reaction
polyE	polychromatophilic erythroblast
qPCR	quantitative polymerase chain reaction
RBC	red blood cell
RCIBGT	Red Cell Immunogenetics and Blood Group Terminology
RFLP	restriction fragment length polymorphism
RhAG	Rh-associated glycoprotein
RNA	ribonucleic acid
SNV	single nucleotide variant
SSP	sequence-specific primers
SV	structural variation
TAD	topologically associated domain
TF	transcription factor
TTI	transfusion-transmitted infection
WBC	white blood cell
WES	whole exome sequencing
WGS	whole-genome sequencing
WHO	World Health Organization

Introduction

"What we know is a drop, what we don't know is an ocean."
Isaac Newton (1642-1727)

Transfusion medicine

Transfusion medicine is a branch of medicine where blood or blood components are transfused as a therapeutic treatment. This field started with the discovery of blood circulation by physician William Harvey, described in his publication concerning motion of the heart and blood in 1628[1]. Almost 40 years later, there was an attempt of xenotransfusion of blood from a lamb to a 15-year-old patient in 1667 by Jean-Baptiste Denis[2]. The earliest published record of human-to-human blood transfusion was performed in 1818 by obstetrician James Blundell trying to save a woman suffering from postpartum hemorrhage[3]. Blundell later performed a successful transfusion to treat a postpartum hemorrhage patient in 1825[4]. Later, Adolf Creite in 1869, and Leonard Landois in 1875, showed cell clumping (hemagglutination) and hemolysis when red blood cells (RBCs) from one species were mixed with serum taken from another species[5]. This agglutination was later observed by Karl Landsteiner in human blood taken from different individuals mixed together[6], where he classified three blood groups A, B and C (C was renamed O for *ohne* in German meaning "without" or "zero") according to their agglutination patterns. Landsteiner discovered that transfusing blood within the same blood group does not lead to RBC destruction. This finding laid the foundation of modern blood transfusion and the first successful blood transfusion according to blood group was performed in 1907 by Reuben Ottenberg. Ottenberg later published on preliminary tests in 1913, which laid the ground work for modern blood banking and led to the use of group O individuals as universal donors for blood transfusion[7].

Today, transfusion medicine involves all aspects from patient blood management, blood donation management, immunohematology laboratory testing, transfusion-transmitted pathogen testing, to monitoring adverse reactions. It remains as an irreplaceable treatment for patients with transfusion-dependent anemia caused by hematological malignancies or hemoglobinopathies, bleeding patients in surgery or trauma, and patients with coagulopathy[8]. In fact, current practice in well-established blood banks is not to transfuse whole blood other than in massive bleeding or special circumstances such as in war or mass-injury situations, but relies on transfusion of the appropriate blood components as indicated by the patient's clinical condition[9]. However, there are recent revivals of the use of whole blood, especially the use of low-titer group O whole blood for trauma patients[10,11]. A single whole blood collection, using differential centrifugation and with freeze thaw procedures, can be separated into several blood components. The most common components with their specific implication in the clinic are packed RBCs, platelets, white blood cell (WBC) concentrates, and plasma. Blood components can also be collected through apheresis collection, where the desired blood component from a single donor is enriched[12].

Blood transfusion is often regarded as a lifesaving procedure; however, it is not without risk. There is the innate risk of transfusion-transmitted infection (TTI) by blood-borne pathogens. Depending on where you are in the world, there are different levels of risk for malaria, HBV, HCV, HIV, syphilis and other agents, but blood transfusion does not appear to pose a threat for SARS-CoV-2 transmission[13,14]. Besides risk for infections, other transfusion adverse reactions including non-hemolytic febrile reactions, allergic reactions, hemolysis, transfusion-associated graft versus host disease (TA-GvHD), transfusion-related acute lung injury (TRALI) and transfusion-associated circulatory overload (TACO), which may cause fatality in severe cases[15,16].

To ensure the safety of blood transfusion, hemovigilance, a set of surveillance processes covering the entire transfusion procedure starting from the collection of blood and its components to the follow-up of recipients, is implemented by many nations[17]. This requires the collaboration of government agencies, international organizations, blood establishments, health care professionals and experts in the field to establish and implement systems and processes concerning transfusion[18]. Sweden in recent years also utilizes the Scandinavian Donations and Transfusions database (SCANDAT), a database including electronically recorded blood donations, transfusions, blood donors, transfused patients, and persons with a blood typing result to facilitate the health of both donors and recipients[19-21].

In the latest report from World Health Organization (WHO), 40% of the 118.5 million blood donations collected were from high-income countries, and there is a drastic decrease of the donation rate from high-income countries (31.5 donations per 1000 people, 3.15%) to low-income countries (0.5%)[22] with the consequence that the scarcity of blood may impose a threat on individual and national health in the less developed countries[23]. Although Taiwan, the country and place I called home and served before this study, has more than twice the donation rate (7.5%) than the average of other high-income countries, the donation rate has declined due to the aging population and will face the problem of blood shortage if no preventive measurements are taken[24]. Patient blood management (PBM), an approach to improve patient outcomes by managing and preserving a patient's own blood, while promoting patient safety, is one of the approaches to mitigate blood shortages. Other innovative ideas to help with the availability of safe blood include enzyme conversion of group A, B or AB to type as group O (ECO)[25,26], growing blood cells from patients' own hematopoietic stem cells, immortalized cell lines or even fibroblasts[27-29], or simply get to the root cause for transfusion by curing hemoglobinopathies and other erythropoietic diseases by gene therapy[30,31].

Blood groups

Blood group antigens and systems

The first blood group system, ABO, was discovered by Karl Landsteiner in 1900[6]. Landsteiner continued to make discoveries of additional blood group antigens, such as the M, N and P antigens[32]. He received the Nobel Prize in Physiology or Medicine in 1930 for his discovery of human blood groups[33]. The requirements to define a blood group antigen are: a) to be present on the RBC surface; b) to be defined by a human alloantibody, that is, at least one person in the world lacks it and makes antibodies against it; c) it should be inherited. Furthermore, for one or more such defined antigens to form a system, the additional criteria should be met: d) the genetic basis should be defined, the gene encoding the antigen should have been identified and sequenced; e) the gene must be different from, and not a closely-linked homologue of, all other genes encoding antigens of existing blood group systems. To date, there are 390 blood group antigens recognized by the International Society of Blood Transfusion (ISBT) Red Cell Immunogenetics and Blood Group Terminology (RCIBGT) working party, and 362 of out the known 390 blood group antigens have been classified into 45 blood group systems (Table 1). The remaining antigens outside the systems are grouped into either a *Collection*- if two or more antigens are related serologically, biochemically or genetically but do not yet fulfil the system criteria; or *700 series*- if neither system or collection criteria are met and the incidence of the antigen is <1%; or *901 series*- if the incidence is >90%.

Blood group antigens are essentially carried on the surface molecules of the RBC membrane, such as proteins, glycoproteins, or glycolipids, and their specificity is determined either by the oligosaccharide or amino acid sequence. Here, we can also divide the blood group systems into two categories: 1) the carbohydrate-based blood group systems, where the gene encodes a glycosyltransferase that will add the terminal sugar to the precursor to synthesize a certain blood group antigen, such as ABO, P1PK, Lewis, H etc.; 2) the protein-based blood group systems, where the gene encodes the protein that carries the antigen itself. MNS, Rh, Kell, Duffy are well-known examples of such systems.

Table 1. Blood group systems recognized by the ISBT (as of Dec 2023).

No.	System	Symbol	Gene name	Antigen(s)	Chr. location	CD number
001	ABO	ABO	*ABO*	4	9q34.2	
002	MNS	MNS	*GYPA,GYPB,(GYPE)*	50	4q31.21	CD235a CD235b
003	P1PK	P1PK	*A4GALT*	3	22q13.2	CD77
004	Rh	RH	*RHD, RHCE*	56	1p36.11	CD240
005	Lutheran	LU	*BCAM*	28	19q13.2	CD239
006	Kell	KEL	*KEL*	38	7q33	CD238
007	Lewis	LE	*FUT3*	6	19p13.3	
008	Duffy	FY	*ACKR1*	5	1q21-q22	CD234
009	Kidd	JK	*SLC14A1*	3	18q11-q12	
010	Diego	DI	*SLC4A1*	23	17q21.31	CD233
011	Yt	YT	*ACHE*	6	7q22	
012	Xg	XG	*XG,CD99*	2	Xp22.32	CD99†
013	Scianna	SC	*ERMAP*	9	1p34.2	
014	Dombrock	DO	*ART4*	10	12p13-p12	CD297
015	Colton	CO	*AQP1*	4	7p14	
016	Landsteiner-Wiener	LW	*ICAM4*	4	19p13.2	CD242
017	Chido/Rodgers	CH/RG	*C4A,C4B*	9	6p21.3	
018	H	H	*FUT1, FUT2*	1	19q13.33	CD173
019	Kx	XK	*XK*	1	Xp21.1	
020	Gerbich	GE	*GYPC*	13	2q14-q21	CD236
021	Cromer	CROM	*CD55*	20	1q32	CD55
022	Knops	KN	*CR1*	13	1q32.2	CD35
023	Indian	IN	*CD44*	6	11p13	CD44
024	Ok	OK	*BSG*	3	19p13.3	CD147
025	Raph	RAPH	*CD151*	1	11p15.5	CD151
026	JohnMiltonHagen	JMH	*SEMA7A*	8	15q22.3-q23	CD108
027	I	I	*GCNT2*	1	6p24.2	
028	Globoside	GLOB	*B3GALNT1*	3	3q25	
029	Gill	GIL	*AQP3*	1	9p13	
030	Rh-associated glycoprotein	RHAG	*RHAG*	6	6p12.3	CD241
031	FORS	FORS	*GBGT1*	1	9q34.13-q34.3	
032	JR	JR	*ABCG2*	1	4q22.1	CD338
033	LAN	LAN	*ABCB6*	1	2q36	
034	Vel	VEL	*SMIM1*	1	1p36.32	
035	CD59	CD59	*CD59*	1	11p13	CD59
036	Augustine	AUG	*SLC29A1*	4	6p21.1	
037	Kanno	KANNO	*PRNP*	1	20p13	
038	SID	SID	*B4GALNT2*	1	17q21.32	
039	CTL2*	CTL2	*SLC44A2*	4	19p13.2	
040	PEL	PEL	*ABCC4*	1	13q32.1	
041	MAM	MAM	*EMP3*	1	19q13.33	
042	EMM	EMM	*PIGG*	1	4p16.3	
043	ABCC1	ABCC1	*ABCC1*	1	16p13.11	
044	Er	ER	*PIEZO1*	5	16q24.3	
045	CD36[#]	CD36	*CD36*	1	7q21.11	CD36

*Cs^a and Cs^b were added to the ISBT 039 CTL2 system at the ISBT RCIBGT meeting in Dec 2023[34].

[#] The CD36 system was added as system ISBT 045 at the ISBT RCIBGT meeting in June 2023[35].

Blood group genetics

Determining the genetics governing blood group antigen expression can be tricky. For instance, the ABO blood group antigens were known since 1900, but it was not until almost a century later at 1990 that the gene encoding transferases synthesizing the A/B antigens, *ABO*, was characterized by Yamamoto[36].

Variations in the blood group genes are often the explanations for the diversity of antigens observed. The molecular mechanism for genetic variations altering antigens can be as simple as a single nucleotide variation (SNV), e.g. *KEL* c.578C>T, which changes k (KEL2) to K (KEL1) antigen, to a large scale of structural variations (SVs) such as deletion of the entire gene, e.g. deletion of *RHD*, which abolishes expression of the entire RhD protein and all its antigens[37] (Table 2).

Most of the blood group systems are governed by a single gene, with few systems having multiple genes. Most of them are homologous genes located closely on the same chromosomes, e.g. *RHD* and *RHCE*, *GYPA* and *GYPB*. The interactions of these homologous genes can abolish or give rise to an antigen. Moreover, the carbohydrate-based antigens, even though each system is governed by its own gene, lack of the glycotransferase for the precursor can also affect the downstream synthesis of other antigens, e.g. variants in *FUT1* make the enzyme lose its ability to make H antigen, and therefore subsequent A or B antigens are also abolished. Hence, the testing for blood group antigens are not always so straight forward.

Table 2. Molecular mechanisms affecting blood group antigen expression.

Molecular mechanisms	Type of change	Example(s) of gene variation(s)	Resulting phenotype
SNV	Antithetical antigen	*DI*01 or DI*A, SLC4A1* c.2561C>T	Di(b+) → Di(a+)
	Novel antigen	*GYPB*24 or GYPB*Mit, GYPB* c.161G>A	Mit+
	Weakening antigen expression	*FUT2*01W.02.01*, *FUT2* c.418A>T *RHD*01EL.01 RHD*DEL1*, *RHD* c.1227G>A	Le(a+b+)[#] Del
	Splice site variant	*JK*02N.01*, *SLC14A1* c.342-1G>A *ABO*B3.03*, *ABO* c.155+5G>A	Jk(a–b–) B_3
	Premature stop codon	*ABCG2*01N.01*, *ABCG2* c.376C>T	Jr(a–)
	Loss of TF binding	*ABO* c.28+5890 T>A (one *B* allele, disruption of GATA1 binding site GATA to GAAA)	B_{weak}
MNV[*]	Antithetical antigen	*GYPA*02 or GYPA*N*, *GYPA* c.59C>T, c.71G>A, c.72T>G	M+ → N+
Small indel	Framshift	*FUT1*01N.13*, *FUT1* c.881_882delTT	H–[†] (found in Para-Bombay)
SV	Gene deletion	*RHD*01N.01*, *RHD* deletion	D–
	Exon deletion	*GE*01N.01*, del exons 3 & 4	Ge:–2,–3,–4, or Leach type (PL)
	Exon duplication	*GE*01.06.01*, duplicated exon 3	GE:6 or Ls(a+)
	Gene rearrangement	*GYP*501 or GYP*Mur*, GYP(B1-136-Bψ137199-A200-229-B230366) *RHD*01N.06*, *RHD* hybrid CE exons 4-7 with 733C>G 1006G>T	MNS:–3,4,7,10, 20,34,35 or S–s+, Mi(a+), Mur+, Hil+, MINY+, MUT+ D–, partial C+ r'^S haplotype
	Large indel	*ABO*c.28+5445_11350del	A_{weak}

Hand-picked variants from previous personal blood group genotyping experiences, with exceptions for examples relating to exon deletion and exon duplication in the Gerbich system.

[*] MNV: multiple nucleotide variants

[#] The variant in *FUT2* affects the secretor status (*Se*), thus the weakening allele of *FUT2* resulted in weak secretor (Se^w), incomplete conversion of type 1 precursor to H type 1 antigen, thus allowing synthesis of both Le^a and Le^b antigens by functional FUT3.

[†] Homozygosity for this *FUT1* null allele with at least one functional *FUT2*, resulted in the para-Bombay phenotype lacking serologically detectable H antigen on RBCs, but A or B antigens can be passively absorbed onto RBCs.

Phenotypes found more prevalently in the Taiwanese population are highlighted in red.

Functional relevance of blood group molecules and antigens

Together with the genetic variations that directly result in changes of (glyco)protein or glycosyltransferases, which affect the blood group antigens, the regulatory aspects can influence the antigen expression levels[38]. This may cause some donor blood to be incorrectly phenotyped and labelled, resulting in the increased associated risk for alloimmunization and/or pregnancy complications and hemolytic transfusions reactions (HTR). Therefore, matching for compatible blood is an unceasing quest for blood bankers.

Furthermore, most antigens have specific physiological functions. While this is not the main topic of this thesis, Table 3 summarizes known and hypothesized functional aspects of blood group molecules.

Table 3. Blood group systems categorized according to their known or putative functions.

Category	Blood group systems
Host defense, innate immunity	ABO, P1PK, LE, H, I, GLOB, FORS, SID
Adhesion, receptor molecules	LU, FY, XG, SC, LW, XK, IN, OK, JMH, CD36
Channels and transporters	RH, JK, DI, CO, RAPH, GIL, RHAG, JR, LAN, AUG, CTL2, PEL, MAM, ABCC1, ER
Enzymes	KEL, YT, DO, EMM
Complement regulation	CH/RG, CROM, KN, CD59
Glycoprotein structure or unknown functions	MNS, GE, VEL, KANNO

Not surprisingly, antigens can be associated with diseases or can alter disease susceptibility or outcome. A well-known example is the Duffy glycoprotein (atypical chemokine receptor 1, ACKR1) acting as the receptor for *Plasmodium vivax* or its absence being correlated with lower neutrophil levels in so-called benign ethnic neutropenia[39,40]. The selection pressure from malaria has given rise to variations of the glycophorins that carry antigens in the MNS blood group system since these molecules constitute receptors for *P. falciparum*. One distinct antigen from this variation, Dantu, reduces risk for severe malaria by 40%[41]. In addition, the ABO blood group distribution is also under the selection pressure of *P. falciparum*, type O individuals tend to have milder disease outcome than type A and are favored in malaria-endemic regions[42]. Another good example of this selection pressure is found in Southeast Asia: the GP.Mur hybrid glycophorin is found in 88% of the Taiwanese aboriginal tribe Ami, as it expedites CO_2 exchange[43] but is virtually absent in Caucasians.

The carrier molecule, the antigens and their association to diseases can facilitate the classification of uncharacterized antigens into blood group systems. One recent example of such is the incidental finding that the Cs^a and Cs^b antigens turned out to correspond to neutrophil antigens, HNA-3a and -3b, which are associated with venous thromboembolism and TRALI, a discovery that made the antigens new members of the CTL2 blood group system[34].

Clinical significance of blood group antibodies

The formation of blood group antibodies is a product of an immune response. These can be alloimmune, usually produced in response to incompatible transfusion, or if mother and fetus/newborn carry different blood types. However, there are also so-called naturally-occurring antibodies, as in ABO, that are formed without prior exposure to allogeneic blood types. The hypothesis for such antibody formation is exposure to environmental agents such as bacteria that are similar to RBC antigens[44]. The antibodies formed against blood group antigens are typically immunoglobulin G (IgG) and IgM. IgG is a monomer carrying two antigen binding sites, and some subclasses of IgG are capable of passing through the placental barrier (IgG1, IgG3, IgG4). IgM is a pentamer, thus carrying 10 antigen binding sites. Although IgM is incapable of passing through the placental barrier, it is capable of agglutinating RBCs even at room or cold temperatures. IgG, on the other hand, reacts best at 37°C and while binding to its antigen, it opsonizes the cells to facilitate phagocytosis for cell removal. However, *in vitro* testing for IgG antibodies, usually requires the addition of antiglobulin as in the indirect antiglobulin test (IAT) to facilitate binding and hemagglutination.

The antibodies are considered clinically significant if they cause the following pathological effects[45]:

1. Destruction of allogeneic RBCs. This may occur after incompatible transfusion and causes HTR.
2. Destruction of autologous RBCs. This is due to autoantibodies which can cause autoimmune hemolytic anemia. Autoantibodies are usually more broadly reactive than alloantibodies.
3. Destruction of fetal RBCs. As discussed above, some IgG can cross the placental barrier and can cause hemolytic disease of the fetus and newborn (HDFN).
4. Damage to transplanted tissues. This is significant if the transplanted tissue harbors many of the incompatible antigens to the host, such as in the case for kidney transplantations where ABO matching is preferred[46].

Current antigen testing methods

Testing for antigens is performed phenotypically by serological typing, or genetically by molecular testing. The former relies on antigen-antibody recognition and is suitable for most blood group antigens, especially the ones without a clear genetic characterization, while the latter focuses on the genetic variation usually at the deoxyribonucleic acid (DNA) level and is most useful when there are no suitable RBCs or phenotyping reagents available for testing.

Serological testing

The basis for serological testing is by antigen-antibody binding, where RBCs are incubated with plasma that may or may not contain blood group-specific antibodies. The end point is commonly agglutination and is classified at a scale from 0 to 4+, where 0 means no agglutination, hence a negative reaction, and 4+ denotes complete agglutination where RBCs are formed to one or a few large clump(s). Any level of agglutination is considered a positive reaction. In real-world blood bank practice, this process can take place on a slide, in a tube, in a column of a gel card, or even in a microwell plate when high-throughput is desired[47,48]. There is also paper-based testing available for ABO and RhD at bedside, which is fast and easy for interpretation. However, there is a trade-off for sensitivity for the bedside test, and the test has been found to be the major risk for ABO-incompatible transfusions[49].

There are serological testing methods available with potentially higher sensitivity, such as protein chip, surface plasmon resonance testing and flow cytometry. All three require specific instrumentation and sample preparation, but flow cytometry is most widely practiced among the three, and it is a very powerful technique when identifying dual cell populations, as in chimeras or transfused patients with blood from one or more donors in circulation.

The sensitivity of serologic testing is based on the avidity and affinity of the antibody-antigen binding, but also depends on the conditions used for interpretation. There are instruments such as a microscope for better observation, or reagents such as low-ionic strength solutions to facilitate the agglutination. There is always a risk that when an antigen is weakly expressed, or the affinity of the antibody binding is low, one might risk a false negative. Unbalanced proportions of RBCs and antibodies can give inaccurate results. In addition, serologic testing may also be compromised by cells that have already bound antibodies before testing, making them direct antiglobulin test (DAT) positive, and this can result in either falsely positive or negative results. Moreover, due to additional polymorphisms in the blood group genes, the antigen may display atypical reactions of both quantitative and qualitative nature. An example is the altered s antigen, where some reagents give positive results but some are weak or negative, as reported by a former Ph.D. student from our group in the Thai population[50]. All the above discrepancies in antigen typing pose the risk for transfusion of mismatched blood which could cause an adverse reaction.

Genotyping to predict antigen phenotype

Genotyping utilizes DNA or ribonucleic acid (RNA) as testing materials and looks for polymorphisms or variations as previously discussed, which contribute to antigen presentation. RNA testing is rare since it is technically challenging because RNA is prone to degradation unless stabilizing agents are added to the sample. However, reverse transcribing RNA to complementary DNA (cDNA) is a common approach when testing for transcripts. Therefore, RNA testing is rare for routine practices, but more and more widely used to solve intriguing and complex cases. One recent example of practical importance is the *RHD* transcript analysis that concluded that Asian-type Del individuals produce low amounts of full-length *RHD* transcript, and that these individuals typically would not be immunized to form anti-D. They can accordingly be safely transfused with RhD positive blood[51].

Blood group antigen genotyping methods started with testing according to specific SNVs or alleles at a low cost. These assays, which include polymerase chain reaction with restriction fragment length polymorphism (PCR-RFLP)[52-54], PCR with sequence-specific primers (PCR-SSP), and PCR with allele-specific primers (PCR-ASP)[55-57], are reliable but generally have relatively low sensitivity and are not quantifiable. A quantifiable method with improved sensitivity that focuses on SNVs or a specific target is the quantitative PCR (qPCR, also known as real-time PCR), which serves well for genotyping and for quantification of gene copies, such as determining *RHD* zygosity and typing for fetal blood groups in maternal plasma[58].

At a higher throughput we have the microarray or DNA chip for handling multiple SNVs in a single test. This is particularly useful and multiple commercial products were made available e.g., BloodChip, HEABeadChip, Luminex xMAP, ID CORE XT and LIFECODE. However, these products were designed for Western populations, or to some degree individuals of African ethnicity. The latter group often present variants in the Rh blood group system and have high prevalence of sickle cell disease which often form alloantibodies following transfusion[59]. Since the distribution of antigens and the underlying polymorphisms vary widely between populations, these commercial genotyping platforms may not be suitable for all populations. Nevertheless, a multinational study developed a universal donor genotyping platform based on arrays has been validated, and shows promise for high-throughput, affordable prediction of blood groups and other markers of relevance for transfusion medicine[60]. Even if concordance rates are high, the most discordant results were initially recorded in the Rh system, which was improved by devising a novel typing workflow[60,61].

The unbiased genotyping method regardless ethnicity would be sequencing, where Sanger sequencing has served as the golden standard for many decades. Sequencing facilitates the finding of novel alleles, has a fast turnaround time and is simple to interpret. However, the sensitivity and the risk of allele-drop out due to primer mismatch during PCR are some of the limitations of Sanger sequencing[62].

Massive parallel sequencing, or more commonly, next-generation sequencing (NGS), is a sequencing technology which sequences a massive number of small fragments of DNA, typically 150 to 300 bp, simultaneously[63]. It has become a powerful tool in many areas of medicine, in which transfusion and blood typing are no exception[64]. In 2016, Lane *et al.* published the first study utilizing whole-genome sequencing (WGS) to predict antigens from 45 blood group and six platelet genes in a single individual[65]. Others (including my previous work before this Ph.D. study) use whole exome sequencing (WES) or targeted blood group panels for a more focused approach[66,67]. NGS offers high sensitivity for fetal DNA typing and can resolve some SVs such as gene rearrangements or large indels in the Rh and MNS systems that were difficult to detect and define with Sanger sequencing[68-71].

Moreover, the less recognized potential of NGS lies with its contribution to the discovery of new blood group systems (Figure 1). Similarly to how the introduction of IAT in 1945 boosted the discovery of new blood group systems, the introduction of NGS also greatly contributed to the recent discovery of several new blood group genes and the establishment of the blood group systems. Starting with VEL (ISBT 034) and KANNO systems (ISBT 037), in which WES was used to confirm the variants, *SMIM1* c.64_80delAGCCTAGGGGCTGTGTC or *PRNP* c.655G>A, were found in the Vel- or KANNO-negative individuals, respectively[72,73]. Following KANNO, the discovery of SID (ISBT 038) comprised of studying the variant fitting for the profile in the NGS-generated 1000 Genomes Project (1000G, now known as The International Genome Sample Resource)[74,75]. The five systems immediately after SID, the discovery of the CTL2, PEL, MAM, EMM, and ABCC1 systems were all successes of WES[76-80], while the latest system, CD36, is a well-established antigen known to be expressed on platelets, now also recognized for its expression on mature RBCs and that anti-CD36 can cause hemagglutination and severe fetal anemia in immunized pregnant women[35].

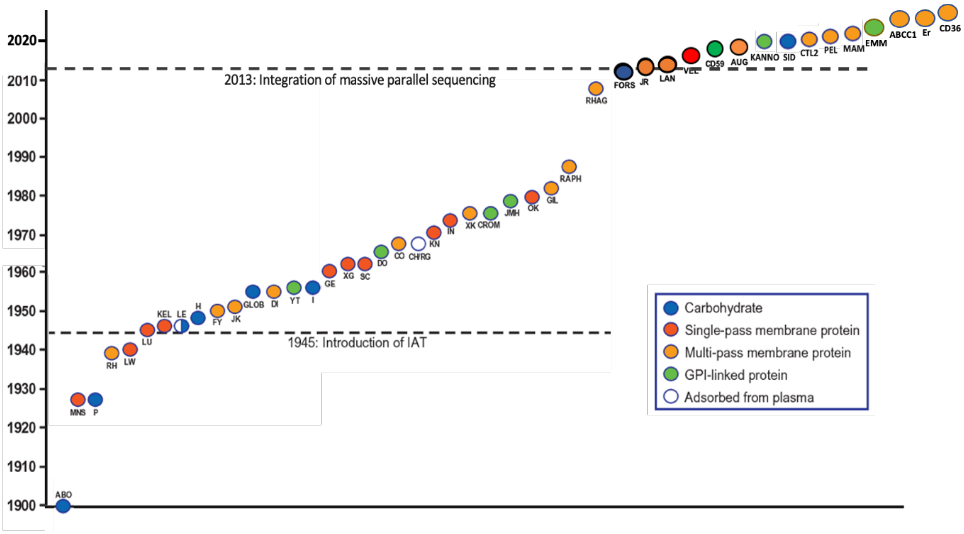
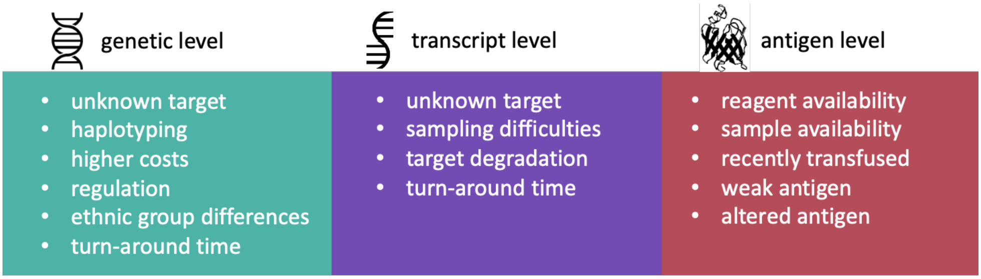


Figure 1. Discovery of blood group systems
Modified from Daniels and Reid, Transfusion 2010 with permission of the publisher John Wiley and Sons (license number 5743241390814)[81].

Recently, third-generation sequencing, also known as long-read sequencing, offers sequencing of single-strand DNA molecules of 20 to 50 kilobase pair (kbp) in size, or even up to 2000 kbp in the ultra-long technology with Nanopore[82]. This technology has become increasingly popular for blood group genotyping for its capability to decipher large SVs and gene hybrids[83,84]. Moreover, long-read sequencing offers haplotyping without additional treatment of the samples, thus resolving ambiguous haplotypes as seen e.g. in the *RHD* and *RHCE* genes and helps build up the reference allele to enable accurate interpretation of intronic regions for *ACKR1*[84-86]. Long-read sequencing is capable of sequencing cDNA in full length, as shown in a recent publication investigating Asian-Del type (*RHD* c.1227G>A) individuals. Their results showed low levels of full-length *RHD* cDNA transcripts among other truncated transcripts, which could both explain why they typed as Del and why they were not prone to make allo-anti-D[51].

Nevertheless, genetic testing can be overwhelming if one wants to sequence for all blood types and can also be costly. Furthermore, there are still 28 serologically defined blood group antigens for which the genetic or regulatory determinants are unknown, thus rendering testing for these antigens impossible until they have been resolved. Even with advanced NGS or third-generation sequencing, one can be left with many variants and needs to figure out the association between variants to each allele (haplotype) or their relationships to the phenotype. Some of the challenges that different kinds of blood group antigen typing are faced with are summarized in Figure 2 below.



genetic level

- unknown target
- haplotyping
- higher costs
- regulation
- ethnic group differences
- turn-around time

transcript level

- unknown target
- sampling difficulties
- target degradation
- turn-around time

antigen level

- reagent availability
- sample availability
- recently transfused
- weak antigen
- altered antigen

Figure 2. Difficulties in blood group antigen typing.

Phenotype/genotype discrepancies

Genotyping is becoming more popular as routine practice in some regions. This has resulted in a growing number of discrepancies between phenotype and genotype. These discrepancies can be classified into four main categories,

1. False positive phenotype, true negative genotype
2. False negative phenotype, true positive genotype
3. True negative phenotype, false positive genotype
4. True positive phenotype, false negative genotype

Interestingly, according to a study spanning from 2015 to 2017 in a molecular reference laboratory, only the first three categories of discrepancies were found[87]. They also summarized that most of the discrepancies were found in the Rh blood group system, followed by the Duffy system. The explanations for the discrepancies were recently transfused patients, difficulties in differentiation of allo- or autoantibodies, poor or unavailable antisera for testing, positive DAT, weak and partial types or silencing alleles not picked up by the molecular method used.

In the following paragraphs, information about the three blood group systems that are specifically investigated in this Ph.D. thesis is summarized.

The Rh system (ISBT 004)

The Rh system was discovered by Landsteiner and Wiener in 1940. They named the antibodies anti-rhesus after they injected RBCs from rhesus monkey (*Macacus rhesus*) RBCs into rabbits or guinea pigs[88]. Eventually they realized that human Rh antibodies are not the same but kept the name Rh. To date, the Rh system consists of 56 antigens, which is the highest number of antigens among all systems (with MNS in 2nd place with 50 antigens). The Rh proteins are erythroid-specific and are expressed on RBCs from cord blood.

Rh antigens and their carrier molecules

The antigens in the Rh blood group system are really fascinating with a complexity arising from the two homologous genes. The two genes, *RHD* and *RHCE* code for RhD and RhCE proteins, respectively. Both proteins are multi-pass transmembrane proteins with 12 transmembrane domains and six extracellular loops[89]. The most clinically important and routinely typed antigens are D, C, c, E and e, where C/c and E/e are antithetical antigens, but D simply denotes the presence of RhD. The notion historically was that "d" either meant the lack of RhD antigen or a hypothetical but not yet discovered antithetical antigen, which turned out not to exist (thus, the symbol "d" is no longer used). The presence or absence of RhD is often referred to as positive or negative blood type, e.g. "I am O positive", which means that "I am blood group O and RhD positive". There are eight possibilities concerning the combination of the five Rh antigens mentioned above. In the 1940s, in order to communicate them in an efficient way, Ronald Fisher and Robert Russell Race came up with the nomenclature for these antigens and Alexander Wiener developed the "Wiener shorthand"[90]. Table 4 is the summary of the two systems and is possibly the only place where D is placed in front of C and not alphabetically, which is otherwise the "norm" to blood bankers.

The Rh proteins (RhD and RhCE) form a core complex together with the Rh-associated glycoprotein (RhAG) in the RBC membrane. This Rh core complex interacts with GPA, GPB, LW (ICAM-4), CD47 and band 3, and is essential to the structure of RBC membrane. When RBCs lack both Rh proteins, they are not able to form the biconcave shape but become stomatocytic and spherocytic, which leads to hemolytic anemia[89,91,92]. The role of Rh proteins may be to transport ammonia over the cell membrane[93]. Recently, a study showed that a variant in the coding region of the RhCE protein is associated with the 2,3-biphosphoglycerate (BPG) levels in erythrocytes, suggesting its role in boosting the generation of BPG[94].

Table 4. Rh haplotypes and their prevalences in three populations. Prevalence data from AABB Technical Manual 21st ed[48].

Wiener shorthand	Fisher-Race haplotype	Prevalence (%)		
		White	Black	Asian
R_0	_Dce_	4	44	3
R_1	_DCe_	42	17	70
R_2	_DcE_	14	11	21
R_z	_DCE_	<0.01	<0.01	1
_r_	_dce_	37	26	3
_r'_	_dCe_	2	2	2
_r"_	_dcE_	1	<0.01	<0.01
r^y	_dCE_	<0.01	<0.01	<0.01

Genetics

The genes governing the Rh system are a pair of homologous genes located at 1p36.11*, _RHD_ and _RHCE_, coding for the RhD and RhCE protein, respectively (Figure 3). The _RHD_ and _RHCE_ genes are highly homologous with up to 97% identical sequences and are positioned in opposite orientation with their 3' ends facing each other[95,96]. Both genes contain 10 exons each, where the exon 2 of the _RHCE*C_ allele is identical to exon 2 of _RHD_. It has therefore been suggested that the C versus c differences come from non-reciprocal gene conversion of _RHD_ sequences to the _RHCE*ce_ allele[97]. This conversion or arrangement between _RHD_ and _RHCE_ is not limited to exon 2 but can be observed in many exons of _RHD_ and _RHCE_. This gives rise to the complexity of the Rh antigens[89]. The unequal gene conversion is also the mechanism behind the _RHD_ deletion responsible for the vast majority of all RhD negative phenotypes. Joining of the _RHD_ flanking regions – the so-called "Rhesus boxes", two 9-kbp homologous and identical orientation regions directly upstream and downstream of the _RHD_ gene, resulted in the complete deletion of the _RHD_ gene[53]. In addition to the traditional haplotype terminology in Table 4, there is a special guideline for naming _RH_ alleles, giving instructions of how normal D or partial, weak Ds should be numbered, and that instead of having just one reference allele, _RHCE_ has four common alleles, _RHCE*01_ for _ce_, _RHCE*02 for Ce_, _RHCE*03 for cE_, and _RHCE*04_ for the allele encoding _CE_[98].

*1p36.11, 1 denotes chromosome 1, p for the chromosome short arm, and the location of 36.11 is according to banding pattern, thus it is read three-six and not thirty-six, same for 11, read one-one, and not eleven.

At the time of writing this thesis, there are around three hundred alleles that cause weak or altered expression of RhD, and close to a hundred alleles that result in RhD negative phenotype. For *RHCE*, there are a total of 150 antigen-weakening or -altering alleles, and 34 null alleles. However, we also need to consider the *RHD* and *RHCE* alleles as haplotypes, that is, a certain *RHD* allele tends to be co-inherited together with a specific *RHCE* allele. This extended *RH* haplotype can create aberrant expression from both *RHD* and *RHCE*. Haplotypes are also a way to help blood bankers with strategies to differentiate true weak D from D negatives. Take weak D typing in Taiwan as an example, all RhD negative donor samples should go through so- called "weak D" testing to ensure that it is indeed a null phenotype and not weak expression of RhD. Since we know that *RHD* null that results from the whole gene deletion often travels with the *RHCE*ce* allele, and the common Asian *DEL* allele travels with *RHCE*Ce*, we then genotype those seemingly "RhD negative" phenotypes but positive for RhC antigen, and found that the majority of these cases indeed carry the Asian *DEL* (up to 1/3 of them initially type as RhD negatives), while very few carried other *RHD* weak or null alleles.

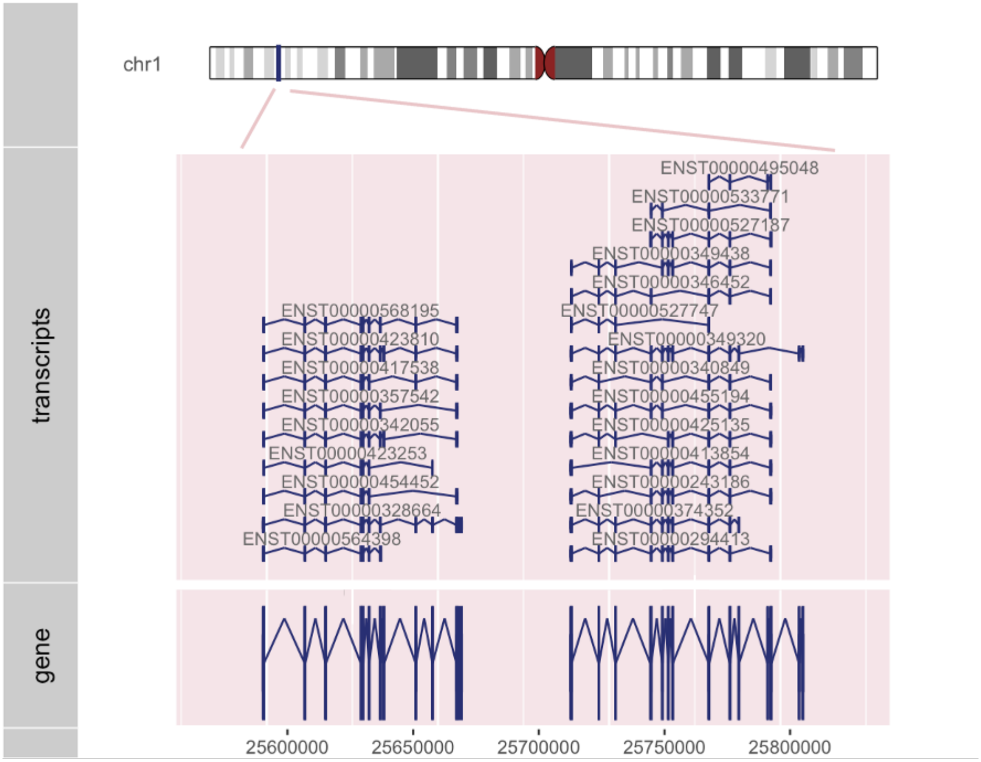


Figure 3. Chromosomal location, transcripts and genomic schematics of *RHD* and *RHCE* genes.

Clinical relevance

The Rh blood group system is clinically significant, and often regarded as the second most important blood group system after the ABO. Rh antibodies can cause immediate or delayed HTR and HDFN. Rh antigens are highly immunogenic, particularly anti-D, and can be induced by the residual number of RBCs in the platelet concentrate when transfused to an RhD negative patient[99]. Rh antibodies are also found in autoimmune hemolytic anemia (AIHA), especially in drug-induced AIHA[100-102]. Anti-D immunoglobulin prophylaxis is administered to RhD negative mothers carrying RhD positive fetuses to prevent HDFN. Nowadays, many countries have nation-wide protocols to type fetal *RHD* by screening cell-free fetal DNA (cffDNA) in maternal plasma in RhD negative pregnant women, including Denmark being the first to implement a nation-wide policy in 2010[103]. Thanks to these national genetic screening programs, administration of RhD prophylaxis will be delivered in a targeted manner, i.e. only to those who truly need them. However, in low-income countries, many of the pregnant women who need this protection are unable to receive it, thus the recently started Worldwide Initiative for Rh Disease Eradication (WIRhE) consortium aims to encourage collaborative projects by providing a centralized source of information about Rh disease for all relevant parties involved in protecting the mother and the baby[104].

Due to the complexity of gene variants and recombination between the *RHD* and *RHCE* genes, the RhD protein phenotypes can be further categorized as partial D, weak D and Del[89]. Generally speaking, the variants affecting the extracellular domain of the *RHD* gene tend to result in partial D (that can also be designated as one of multiple known D categories), where the RhD epitopes are affected and some but not all are missing. This means the patient or donor can in fact be viewed as both "RhD positive and negative" at the same time. In practice, this means a blood recipient or pregnant woman needs to be considered as RhD negative while a blood unit from a donor with partial RhD phenotype needs to be labelled as RhD positive. SNVs affecting transmembrane domains or intracellular loops can result in weak D phenotypes, where the RhD epitopes are not supposed to be affected, but in which the ability to integrate into the membrane is decreased and consequently results in reduced levels of RhD protein on the RBC surface. In these cases, an antiglobulin test is often required to get an approved result of the RhD blood group typing. Del is a particular variant of weak D that carries too few of the RhD polypeptides to allow hemagglutination to occur in routine methods but can be detected only by the adsorption-elution method. There are studies showing that the carriers of the most common weak D variants (type 1, 2, 3) in Caucasians are not at risk for forming anti-D. They do not need for Rh prophylaxis and can be safely transfused with RhD positive blood without complications[105]. However, there are reports that donors with weak D types can cause RhD immunization[106]. Thus, precision diagnostics for RhD can be important in both patients and donors.

The Kell system (ISBT 006)

The Kell blood group system was first discovered in 1946, a direct beneficiary of the introduction of antiglobulin test in 1945, and is named after Mrs. Kelleher, the first antibody producer[107]. To date, there are 38 antigens in the system, including 8 pairs of antithetical antigens*, and 12 low-prevalence antigens. The protein appears early in erythropoiesis and is also expressed on myeloid progenitor cells and megakaryocytes. Kell antigens are expressed on other tissues including fetal liver, testes, and with lesser amounts in heart, brain, lymphoid organs and skeletal muscle. Interestingly, the K antigen (KEL1) is found in about 9% Caucasians, and can be as high as 25% in Arabs, but is very rare in Asians, and has not been reported in the Taiwanese population (including aboriginals and Polynesians in Taiwan)[108,109]. Therefore, while the risk for immunization against KEL1/KEL2 (K/k) is substantial in Caucasians and all blood is typed for the presence of KEL1, antigen typing of KEL1 is not routinely performed in Taiwan and some other parts of the world where immunization does not constitute a clinical problem there.

Antigens and carrier molecule

The Kell glycoprotein is a type II single-pass membrane protein that consists of 732 amino acids. It serves as an endothelin-3-converting enzyme, which creates the bioactive vasoconstrictor endothelin-3 by cleaving big endothelin-3[110]. The Kell glycoprotein is covalently linked to XK, an integral RBC membrane protein, which carries the Kx antigen and encoded by the X-linked gene *XK*, via a single disulfide bond at Cys72 of Kell and Cys347 of XK protein[111]. The absence of XK protein causes the McLeod syndrome, a disorder that affects blood (hemolysis and acanthocytosis), brain and nerve system, muscles, and heart. Kell protein is only weakly expressed on the RBC membrane in McLeod individuals[112]. The absence of Kell glycoprotein is called the K_{null} (K_0) phenotype, and the corresponding antibody is named anti-Ku. A K_0 phenotype individual will have no Kell antigens, but an increased expression of Kx antigen compared to the common phenotype[113].

* KEL3/KEL4/KEL21 are also known as $Kp^a/Kp^b/Kp^c$, respectively, and are grouped together as antithetical antigens; the Kp^b is the high-prevalence antigen and is encoded by the reference allele with amino acid Arg at position 28, while in Kp^a it is replaced with Trp and Kp^c with Gln at the same position.

Genetics

KEL consists of 19 exons and spans over 21 kbp on 7q34 (Figure 4). Interestingly, most of the antigen-determining SNVs are located in exons 6 and 8, with a few located in other exons, but none in the first three or the last exon(s). At the time of writing, there are 71 known alleles, with various molecular mechanisms such as nonsense mutations, splice-site variations, insertions, deletions or a combination of indels, contributing to the K_0 phenotype. There are also known homozygous or compound heterozygous exonic variants leading to the $Kell_{mod}$ phenotype, which exhibits weak expression of Kell antigens. There are 25 alleles listed as $Kell_{mod}$ in the latest version of the KEL blood group allele table from ISBT (v8.0), but two alleles were marked as K_0 phenotype while one $Kell_{mod}$ allele shares the same variant as a K_0 allele. This perfectly demonstrates how difficult it maybe to discriminate weak expression from the null phenotype, in this case it also depends on the reagents and the techniques used for phenotyping. To complicate antigen testing even further, the weakened expression of the Kell antigens can be the result of a single SNV, such as *KEL*01.02* weakening the expression of K[114]; or it can be weakened due to the formation of another antigen, such as *KEL*01.03*, which encodes Kp^a but also has a *cis*-modifier effect that weakens the expression of KEL system antigens on the cell surface[115,116]. Moreover, the absence of an antigen may affect the expression level of other antigens even if they are not antithetical, such as in the case of KANT− individuals, who express the K11 antigen weakly and KETI antigen very weakly[117].

Clinical relevance

Kell is considered the most clinically significant blood group system after ABO and Rh, and alloimmunization in the Kell system is the third most common cause for HDFN. Kell antigens are developed early during erythropoiesis and the antibodies against Kell antigens can therefore not only cause lysis of RBCs but more importantly underlie severe HDFN due to and pronounced fetal anemia[118]. The antibodies can also cause mild to severe HTR (both acute and delayed). Due to the clinical significance and the risk for HDFN, Kell was one of the first antenatal genotyping targets of NGS using cffDNA[119]. Anti-K in pregnancy not only destroys RBCs but can also suppress erythropoiesis, that leads to severe anemia which may require intrauterine transfusions[120-122].

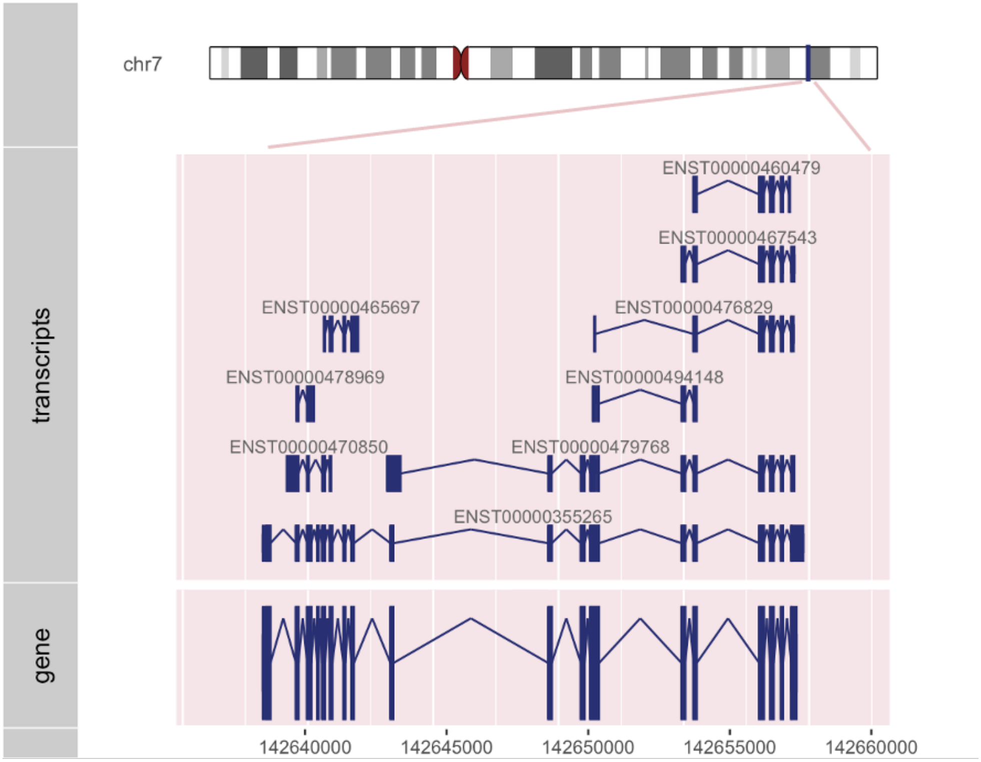


Figure 4. Chromosomal location, transcripts and genomic schematics of *KEL* gene.

The Knops system (ISBT 022)

The Knops blood group system was discovered in 1970. The first antigen was named Knops-Helgeson (Kn^a) after the patient (D. Knops), who had an antibody that was incompatible with all the units in the inventory, and the blood banker (Margret Helgeson) who eventually donated her own blood for the patient. Helgeson tested her own blood and turned out to be compatible and the patient was transfused without complication but the antibody titer increased post transfusion[123]. It was only after a half-century later (and the focus of **Paper I**) that we understood the mechanism of Helgeson's phenotype. To date, the system consists of 13 antigens with Complement Receptor 1 (CR1; CD35) as the carrier protein. CR1 expression is common on blood cells and exists at low levels as a soluble form in the plasma. Other tissues expressing CR1 include glomerular podocytes, dendritic cells in spleen and lymph nodes and peripheral nerve fibers.

Antigens and carrier molecule

The 13 antigens recognized by ISBT are classified as 5 pairs of antithetical antigens, and three independent antigens, Yk^a, Sl3, and KNMB. However, the Sl3 antigen is very dependent on the Sl1 antigen, where the amino acids at position 1601 and 1610 determine the Sl1, Vil (previously Sl2) and Sl3 antigen specificity. The molecular model of Moulds *et al.*[124] showed that the proximity of p.1601Arg and p.1610Ser would create the conformational epitope of Sl3, thus changing the amino acid at either position will cause loss of the Sl3 antigen as described in Table 5.

Table 5. Relationships of the Sl1, Vil (Sl2), and Sl3 phenotypes.

Phenotype	p.1601	p.1610	Note
Sl:1, –2, 3	Arg	Ser	Common in Caucasians
Sl: –1, 2, –3	Gly	Ser	Common in African American
Sl: 1, –2, –3	Arg	Thr	Only in one Caucasian

CR1 is a single-pass type I membrane glycoprotein, which binds C3b and C4b for the purpose of processing and clearance of complement-opsonized immune complexes. It also has an inhibitory effect on both the classical and alternative complement pathways. Human CR1 exhibits three types of polymorphism: a) peptide sequence variants from DNA sequence differences contributing to the antigens of the Knops system[125], b) four structural polymorphisms (allotypes A to

D) with molecular weight ranging from 160 kDa to 250 kDa* due to the number of long homologous repeat (LHR)[126], and c) expression levels, i.e. number of CR1 molecules per RBC, also known as the CR1-E† levels[127,128] (Figure 5). The number of CR1 molecules on RBC can vary 10-fold among normal individuals, with most people having a range of 100-1,000 CR1 molecules per cell, whilst a fraction of people may carry just 20-100 molecules per cell[127,129]. The latter is said to exhibit the "Helgeson phenotype", where the few CR1 molecules are too distant to allow agglutination and are considered a phenotypically "null" phenotype. It is named after the blood banker M. Helgeson who contributed to the discovery of the system for carrying such a phenotype[129]. Again, antigen determination is never easy and straight forward. Besides the hereditary factor that could result in the (very) weak expression of CR1, CR1 can be consumed due to diseases, causing an acquired Helgeson phenotype, and the level of CR1 on RBCs is decreased upon storage and aging of RBCs[130,131].

Genetics

The gene encoding the Knops antigens is *CR1*, located at 1q32.2 (Figure 6), a region that harbours other immunoregulatory genes, such as complement component 4 binding protein alpha (*C4BPA*), decay-accelerating factor (DAF; *CD55*; Cromer blood group system, ISBT 021), complement receptor type 2 (*CR2*), complement receptor 1-like (*CR1L*), membrane co-factor protein (*CD46*). Based on the allotypes, the length of *CR1* varies from 130 to 160 kbp, and the number of exons also varies[125,132]. To facilitate communication for the following discussion, the exon numbering here and onwards will be based on the CR1*1 (A allotype), which consists of 39 exons. Despite the large number of exons, the variations altering the antigen expression clusters on exon 29, with few on exon 22 and exon 26.

As early as in 1986, Wilson *et al*. associated *Hin*dIII RFLP to the level of CR1 on RBCs[127]. However, this restriction site, rs11118133A>T in intron 27 of *CR1*, proven to be a great predictor for Caucasians, was a poor predictor in Africans[133,134]. Later, other SNVs (which were strongly associated to the *Hin*dIII restriction site) were proposed to be predictors for low CR1 expression, such as rs2274567A>G (p.His1208Arg) and rs3811381C>G (p.Pro1827Arg)[135-138]. However, they were not better predictors than the *Hin*dIII site for the Helgeson phenotype outside of the Caucasian population and nor do they explain functionally why CR1 expression varies. Interestingly, rs2274567 was recognized as the SNV governing a new pair of antithetical antigens DACY/YCAD in the Knops system during the course of this study[139].

*under non-reducing conditions

† E for erythrocytes

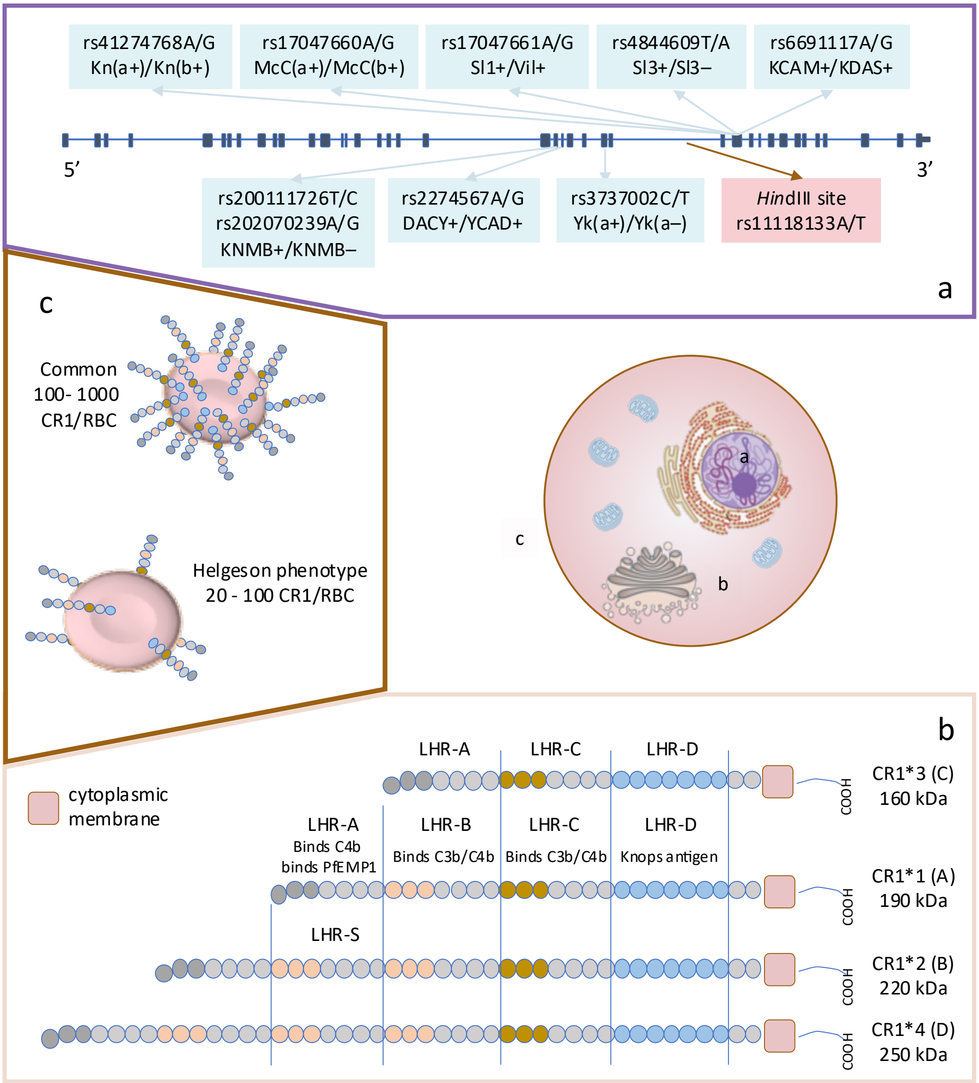


Figure 5. Three types of CR1 polymorphism.

a) DNA sequence variants with its rs number contributing to the 13 antigens of the Knops system are shown in light blue box, most of the variants are located in exon 29 as shown above the gene. The *Hind*III restriction site is shown in pink box, which was used to associate with high or low copies of CR1 on RBCs mainly in Caucasians. **b**) The structural polymorphisms with allotypes CR1*1 to CR1*4 (also A to D), with CR1*1 being the most prevalent one, and was set as an example for labeling the LHR regions and the main functions of each region. Most of the Knops blood group antigens are situated in LHR-D region. **c**) The copies of CR1 molecules on each RBC. In common individuals there are 100 to 1000 copies of CR1 per RBC, but in individuals with the Helgeson phenotype, it is found to be only 20 to 100 copies of CR1 molecules per cell.

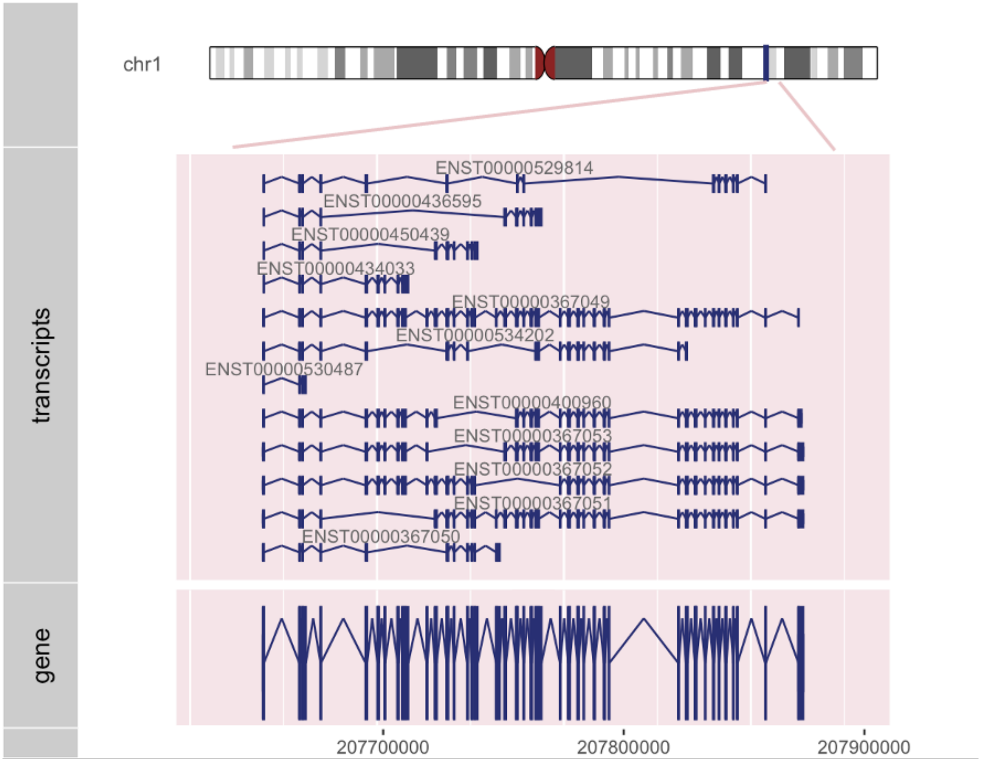


Figure 6. Chromosomal location, transcripts and genomic schematics of *CR1* gene.

Clinical relevance and disease association

As mentioned previously, the antibodies against the Knops antigens are not considered clinically significant since they do not cause HTR nor HDFN. However, they may complicate the antibody identification process being found in multispecific sera or reacting as weak pan-agglutinins with all panel cells[140]. Therefore, it is still common practice to identify their specificity in order to make sure they are not mimicking other antibodies, and that the patient can be safely transfused with Knops-incompatible blood. A recent addition to the serological toolbox for defining suspected antibodies against Knops antigens is to add to the patient plasma recombinantly expressed, soluble CR1 (sCR1) protein corresponding to the epitopes for common antigens[141]. If anti-Kn^a is present, the sCR1 will neutralize the antibodies and turn a previously positive reaction negative, or at least weakened.

Nevertheless, CR1 is associated with numerous diseases due to its role in immune responses. Low CR1 levels have been correlated with sarcoidosis[135], systemic lupus erythematosus (SLE)[129], and Alzheimer's disease[142]. Moreover, CR1 is a receptor for the invasion of *Plasmodium falciparum* via PfRh4[143]. CR1 also serves as a ligand for the *Plasmodium. falciparum* erythrocyte membrane protein 1 (PfEMP1) expressed on infected RBCs[144]. This interaction of PfEMP1 with CR1 on uninfected RBCs results in RBC rosette and contribute to infection and disease severity[145,146]. RBCs with low CR1 copy number, such as the Helgeson phenotype, rosette poorly, and thus this phenotype confers protection against severe malaria[147].

Erythropoiesis

"That's here. That's Home. That's us."
Carl Sagan, Pale Blue Dot, 1994

Erythropoiesis is a part of hematopoiesis describing the differentiation and maturation process of RBCs, also known as erythrocytes. In the early stage of the development of embryo, primitive erythropoiesis takes place in the yolk sac, where large nucleated erythroid cells are produced. Later in development, definitive erythropoiesis takes place first in the fetal liver, then later transits to bone marrow, resulting in small enucleated erythrocytes. In adult erythropoiesis, it can be characterized by the several intermediate progenitor stages from hematopoietic stem cell (HSC) to mature RBCs (Figure 7).

In the classical model, differentiation to the erythroid lineage begins with the HSC, the progenitor of all blood cells, to differentiate to the erythroid lineage. HSCs first differentiate to a stage called colony-forming unit-granulocyte, erythrocyte, monocyte and megakaryocyte (CFU-GEMM) or common myeloid progenitor (CMP). It is a multi-potent progenitor cell that can give rise to the cells mentioned in its name. Further, CFU-GEMMs take one step toward the erythroid lineage by differentiating into CFU-EM, or megakaryocyte-erythrocyte progenitors (MEP). CFU-EM commit to the first erythroid progenitor as burst-forming unit-erythroid (BFU-E), which can be identified by their ability to generate hemoglobinized progeny and proliferating capacity. BFU-E also exhibit a limited self-renewal ability. The next differentiation stage is called the colony-forming unit-erythroid (CFU-E). CFU-Es cannot self-renew, and their proliferative ability is limited. They also lack the cell surface marker, CD34, that is common to the progenitors. However, they express Rh proteins and glycophorin A (GPA) that are characteristic of erythroblasts. Next, CFU-E differentiate into the proerythroblast, the earliest morphologically recognizable erythroid cell. Starting with this stage, erythroid cells increase their hemoglobin synthesis, acquire and store large amounts of iron for heme synthesis. Proerythroblasts continue to differentiate into basophilic erythroblasts (basoE), and at this stage, become slightly smaller in cell and nucleus size, the cytoplasm increases with abundance of RNA, which make them intensely blue with Giemsa stain. Next, in the lineage comes the polychromatophilic erythroblast (polyE). The word "*poly chroma*" refers to the lighter-greyish cytoplasm by both acidic and basic stains. The polyE differentiate into orthochromatophilic erythroblast (orthoE). At this stage, the cell size continues to decrease, and the nucleus becomes a dark opaque nucleus. The extrusion of the nucleus, a process also known as enucleation, occurs at this stage, which is crucial to the maturation of the RBCs. Once polyE extrude the nucleus, they develop into reticulocytes and begin to migrate from bone marrow into the circulation. Reticulocytes are named for the mesh-like reticular network of ribosomal RNA, thus the cytoplasm shows reddish-blue color, and the size continues to decrease. Finally, the mature RBC is formed, where RNA is lost, and the abundance of hemoglobin turns it bright red in color. The form turns into its iconic biconcave shape, to facilitate oxygen exchange.

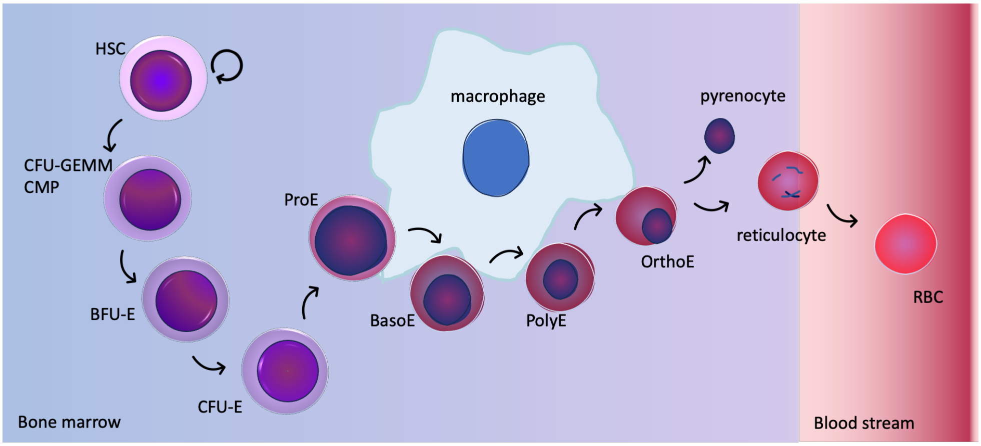


Figure 7. The process of erythropoeisis from hematopoietic stem cell (HSC) to mature RBC.

From birthplace to grave site of erythrocytes

The erythropoiesis process in adult humans takes place primarily in a specialized microenvironmental compartments in the bone marrow, called the erythroblastic islands (EBI)[148]. EBIs consist of a central macrophage, surrounded by erythroid progenitors, ranging from CFU-E to young reticulocytes[149]. Studies suggested that the macrophages provide nutrients to the erythroid progenitors, and also give signals for proliferation and survival, as well as help with the extrusion of nuclei at the end of the erythropoietic process[149]. After enucleation, the young reticulocytes will detach from EBI within 24-48 hours and move into the circulation, where they expel the remaining organelles and remodel the membrane for the next 24-48 hours until RBC maturation. The RBCs have a life span of 120 days, and old cells will be cleared by in the liver and spleen. About 200 billion erythrocytes are formed in the bone marrow each day, about 2 million per second[150]. The erythrocytes exceed by far the number of other cell types. In fact, it was estimated to constitute 84% of all human cells[151].

Regulation of human erythropoiesis

The extrinsic regulation of erythropoiesis

Erythropoiesis is regulated by several cytokines and growth factors, including the stem cell factor (SCF), interleukin-3 (IL-3) and the most essential erythropoietin (EPO)[152]. EPO is produced by renal stromal cells in adults and low levels are constantly being secreted to maintain the normal renewal of RBCs. The production of EPO can be boosted in response to cellular hypoxia, and up-regulated erythropoiesis. CFU-E are highly responsive to EPO and the binding of EPO through the EPO receptor (EPOR) prevents apoptosis and enhances subsequent proliferation and differentiation. The binding of EPO to EPOR activates the intra-cellular JAK2 signalling cascade and initiates the STAT5, PIK3, and Ras MAPK pathways[153]. The activation of these pathways promotes cell survival and proliferation through transcriptional control in the nucleus as the intrinsic regulation of erythropoiesis, where the transcription factors (TF) are critical in this process.

The intrinsic regulation of erythropoiesis

TFs, coregulators, and non-coding RNAs (ncRNAs) are involved in the intrinsic regulation of erythropoiesis[154,155]. TF are proteins that recognize specific sequences called motifs to allow binding to DNA, which controls the transcription processes. The main DNA-binding TFs involved in erythropoiesis include GATA1, GATA2, KLF1, NFE2, TAL1, BCL11A, PU1 and RUNX1. The interplay of the TFs governs the process at specific stages and propels erythropoiesis forward in an orderly fashion. The TFs are also interconnected as one TF can be the driver or inhibitor for the transcription of another TF, as seen with GATA2 which activates GATA1 transcription, and GATA1, which enhances KLF1 but inhibits GATA2[156]. A selection of TFs (Figure 8) is described later together with their binding partners such as erythroid co-regulators FOG1, LMO2, LDB1 and p300/CBP. Noncoding RNAs have been shown to be a modulator of erythropoiesis, including microRNAs (miRNAs, 21 to 23 nucleotides) and long ncRNAs (lncRNAs, >200 nucleotides)[152,157,158]. It has been shown that GATA1 regulates several erythroid-specific miRNAs (miRNA144/145). Further, miRNA15a can bind to the 3'UTR of TF c-Myb gene (*MYB*) to regulate γ globin, and miRNA96 binds to the coding region of γ globin mRNA to decrease its expression[159]. Recently, a study showed that the GATA2 antisense (GATA2AS) lncRNA is able to activate *GATA2* and *HBG*, and interact with erythroid TFs including KLF1 and is essential for maintaining the chromatin regulatory landscape during erythropoiesis[160].

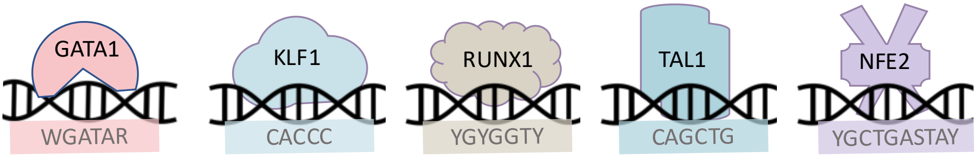


Figure 8. Erythroid transcription factors included in this study and their binding motifs.

GATA1

GATA1 belongs to the family of zinc-finger proteins and binds to the consensus DNA sequence, WGATAR, and is named after the center of its own motif[161]. GATA1 and GATA2 both recognize the same binding sites but their role during erythropoiesis is temporally and functionally different. GATA2 is expressed at an earlier stage of erythropoiesis and is found at high levels in HSCs, CMPs, and MEPs. However, when the cell lineage is committed to the erythroid lineage as erythroblast, the expression of GATA2 decreases and GATA1 increases. This is known as the GATA switch[156]. While GATA2 positively regulates both *GATA1* and *GATA2*, GATA1 only auto-activates but represses *GATA2*.

GATA1 shows binding sites in many erythroid gene promoters, including other erythroid transcription factors, such as KLF1, TAL1, GATA1 and GATA2, therefore, it is considered as a master regulator of erythropoiesis. It also promotes the maturation of erythroblasts into mature RBCs and is involved in stimulating essential genes for oxygen-carrying components including the α- and β-globin genes, and heme that is indispensable for RBCs. GATA1 also contributes to the maturation of megakaryocytes and the shedding of platelets into blood[162].

GATA1 can work independently or together with its partner protein, friend of GATA1 (FOG1). FOG1 is a co-regulator and can assist GATA1 in selective binding to specific WGATAR sites[163]. The GATA1-FOG1 complex also recruits other gene expression-regulating complexes such as the chromatin modulator Mi-2/NuRD complex or histone deacetylase, CTBP1, or other regulating proteins. GATA1 also works with TAL1 and is involved in the SCL complex (described more in the TAL1 section). The interactions between these complexes and co-regulators shape the interaction of GATA1 in a gene-specific manner.

GATA1 is located on the X chromosome, therefore, inactivating mutations of *GATA1* cause X-linked recessive diseases, predominantly in males. Since GATA1 plays a critical role for hematopoiesis, a complete lack of GATA1 is lethal at the embryonic stage. However, non-lethal mutations in *GATA1* can lead to a spectrum of RBC and megakaryocyte disorders depending on the type of the mutation. The genetic background of the individuals can also affect the manifestation, e.g. the mutation of GATA1 in individuals with trisomy 21 will cause a transient myeloproliferative disorder and acute megakaryoblastic leukemia characteristic of Down’s syndrome[164].

Variations in *GATA1* can cause the X-linked Lu(a–b–) blood group phenotype[165]. It was first hypothesized that an unidentified X-linked suppressor gene (named as *XS2*) caused the phenotype. Sequencing of the proband revealed the *GATA1* stop codon was replaced by a missense mutation (p.Ter414Argext*41). The proband also showed weak P1 expression, but normal CD44, GPA, GPC, CD59, Fy3, and Co^a antigens. The hereditary pattern and the antigen expression profiles of this Lu(a–b–) case separate itself from the In(Lu) phenotype caused by another TF discussed in the following section. There are also a number of blood group genes where GATA1 binding has been demonstrated. The loss of a GATA1 binding site in these genes could cause significant downregulation of blood group expression levels. This will be further expanded in a later section.

KLF1

KLF1 also named EKLF, short for erythroid Krüppel-like factor, binds to the CACCC DNA motif. The expression of KLF1 is limited to the erythroid lineage and is essential for erythropoiesis. During CFU-E and pro-E stages, KLF1 is mainly located in the cytoplasm, but imported into the nucleus when differentiating from pro-E to Baso-E[166]. KLF1 is involved in globin switching from embryonic stage (ζ and ε) to fetal stage (α and γ), and then to adult (α and β). KLF1 interacts with the promoter of the β globin gene, and activate the promoter of *BCL11A*, a repressor protein for the fetal γ globin gene. Besides having the transactivation domain to bind to DNA, KLF1 also has a chromatin-remodelling domain enabling it to recruit chromatin remodelling complexes and components to shape the epigenetic landscape for basal transcription.

KLF1 is located on chromosome 19, and more than 100 variants have been reported that can lead to a diverse spectrum of erythroid phenotypes[167], including defects in RBC membrane proteins, persistent elevated HbF (HPHF)[168], hemolytic anemia[169,170], congenital dyserythropoietic anemia (CDA)[171], and can cause hydrops fetalis due to severe anemia[172]. Variants in the *KLF1* gene has been reported to cause weakening expression of several blood group antigens. The most well-known is the In(Lu) phenotype, where Lu(a–b–) is observed despite a fully functional *LU* (*BCAM*) gene caused by haploinsufficiency of *KLF1*, i.e. an individual carrying one functional *KLF1* allele and one defective allele[173]. The variants leading to the In(Lu) phenotype include a variant in the *KLF1* promoter, which disrupts the binding of GATA1. Others result in premature stop codons or missense variants in the zinc-finger domains that affect the binding to the motif. A decreased level of CD44 has been observed in individuals with the In(Lu) phenotype[171,173,174]. Other blood group systems have been reported to be affected by *KLF1* mutations, including Colton (*AQP1*)[171,173], Dombrock (*ART4*)[173], Scianna (*ERMAP*)[173], Ok (*BSG*)[173], Duffy (*ACKR1*)[173] Diego (*SLC4A1*)[173], and LW (*ICAM4*)[171]. Interestingly, previous studies with ChIP-seq analysis of KLF1 in mice revealed KLF1 binding to many erythroid genes, including blood group genes mentioned above, *Ermap*, *Cd44*, *Gypc*, but not

Bcam. Another study with two CDA patients carrying the *KLF1* mutation, NM_006563.3: c.973G>A, NP_006554.1: p.Glu325Lys, showed decreased expression of CD44, AQP1 and ICAM4 in both individuals, but BCAM was only decreased in one of the two patients[171]. To this day, it is not fully understood how KLF1 affects LU expression. Adding to the mysteries yet to be explained, in a previous study of the P_1 and P_2 (i.e. the lack of P1 antigen) phenotypes, even though a KLF1 motif was found in the responsible gene, *A4GALT*, KLF1 did not appear to affect the *A4GALT* transcript levels. Furthermore, no other connection with KLF1 was observed to explain why P1 would be low in the In(Lu) phenotype[175].

RUNX1

RUNX1 belongs to the family of Runt-related transcription factors and is also known as the acute myeloid leukemia 1 protein (AML1) or core-binding factor subunit alpha-2 (CBFA2). It forms a heterodimer with CBFβ, which stabilizes the complex and increases the ability of DNA binding. RUNX1 recognizes the YGYGGTY motif and regulates genes involved in hematopoietic differentiation, definitive erythropoiesis, cell cycle, and ribosome biogenesis[176]. Similar to GATA1, RUNX1 also autoregulates itself through binding to the enhancer in the intronic region of *RUNX1*. Interestingly, *RUNX1* has two enhancers, which permit binding for lymphoid or erythroid regulatory proteins in a tissue-specific manner. *RUNX1* is located at 21q22.12, spanning 261 kbp and is involved in over 50 chromosomal translocations found mostly in acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL)[177]. Mutations disrupting the DNA-binding domain of RUNX1 can be found in around 18% of T-ALL patients[178].

RUNX1 binds to the +5.8 kbp of *ABO* intron 1, and disruption of the RUNX1 motif result in weakening expression of A/B antigens[179,180]. In proximity to this RUNX1 motif are two GATA motifs, and they are all part of the LTR15 sequence, which is a type of the transposable element containing long terminal repeats[181]. These transposable elements have the ability to replicate and insert themselves into the genome and have been shown to be capable of gene regulation[182]. Moreover, RUNX1 has shown to be the major player in determining the P_1/P_2 phenotype that could not be explained by KLF1 above. By knocking down RUNX1, *A4GALT* expression decreased significantly, and a SNV at the rs5751348 locus contributes to binding or loss of binding of RUNX1, resulting in the P_1 or P_2 phenotypes[175].

TAL1

T-cell acute lymphocytic leukemia protein 1 (TAL1), also called stem cell leukemia (SCL) factor, is a helix-loop-helix protein containing basic amino acid residues to enable DNA binding[183]. TAL1 participates in the SCL complex where it works together with other TFs and co-factors such as GATA1, LMO2, and LDB1. The complex is required for erythropoiesis especially for GATA1-activated genes, but is not present for GATA1-repressed genes[184]. CAGCTG is the preferred binding motif for TAL1 in erythroid cells, and is a so-called E-box sequence (CANNTG). TAL1 is also known to work together with GATA1, where the two motifs are conserved and spaced by 9 to 12 nucleotides (WGATAN$_{9\text{-}12}$CANNTG) in the regulatory regions of many erythroid genes. However, only a small number of erythroid-specific genes have been identified as direct targets of TAL1, and include both *GATA1* and *KLF1*. TAL1 is also suggested to be involved in regulating histone modification at erythroid-specific enhancer and promoter elements by interacting with histone acetyltransferases p300/CBP, pCAF and histone deacetylases (HDACs)[185]. It is also involved in DNA looping together with GATA1 to recruit LDB1 for dimerization to bring enhancers and promoters closer in proximity to stimulate gene transcription[186].

NFE2

NFE2, nuclear factor, erythroid 2* is a heterodimer complex with a broad Maf (muscular aponeurotic fibrosarcoma) protein and an erythroid-specific subunit, p45, which is involved in erythropoiesis and megakaryopoiesis. NFE2 recognizes a T/CGCTGAC/GTAT/C sequence[187]. NFE2 is involved in the DNase hypersensitivity sites (DHS) of the locus control region (LCR) region for β globin synthesis. A previous study also suggests that NFE2 directly binds to distal regulatory elements and indirectly to proximal promoter region, as a supporting element[187]. The β globin promoter does not contain NFE2 binding sites[188].

Transcription factors and blood group antigen expression

It is well-established that the primary function of TFs is to bind to their motifs in regulatory regions such as promoters or enhancers, and to control the rate of the transcription of their target genes[189]. The binding of TF to the promoter region helps stabilize the transcription initiation complex, and the binding to the enhancers can either up-regulate or repress the transcription. The gene regulation by TF is an intricate process concerning spatial and temporal dynamics to ensure that the right set of the genes are expressed during a certain stage of cell development and differentiation. The loss of TF binding either by mutation of the TF or the disruption of the motif will affect the gene transcription as depicted in Figure 9.

*Nuclear factor, erythroid 1 is the former name of GATA1.

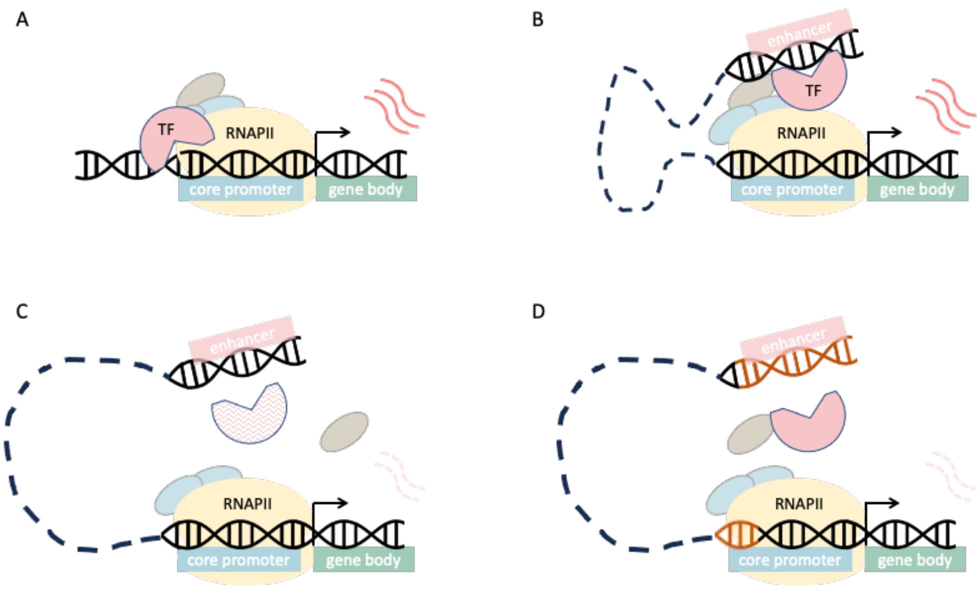


Figure 9. Transcription is enhanced by the binding of TFs to the motifs in promoters and enhancers.
The transcription machinery highlighting the mechanism of TF binding. **A.** TF binds to the motif in the proximal promoter to enhance gene transcription. **B.** TF binds to the motif in the enhancer to facilitate looping of DNA to interact with the RNA pol II machinery at the core promoter with the cofactors and mediators, and enhances gene transcription. **C.** A mutation in the TF gene generates defective TF, which loses its ability to bind to the motifs or to its co-factors, thus no loops are formed and gene transcription is not enhanced. **D.** TF is functional, but variants in the DNA motif disable TF binding at the enhancers or promoters, thus gene transcription is not enhanced. TF, transcription factor; RNAPII, RNA polymerase II.

As mentioned earlier, variants in both *GATA1* and *KLF1* can result in the Lu(a–b–) phenotype, XS2 Lu-mod and In(Lu), respectively, and the variants leading to these phenotypes are summarized in ISBT allele tables[190,191]. Another mechanism for TFs to affect blood group antigen expression is through the loss or gain of TF binding in the regulatory regions. The most well-known disruption of TF binding is the variant defining the Duffy-negative phenotype, Fy(a–b–), primarily in individuals of African ethnicity[192]. In these individuals a SNV in the promoter region of *ACKR1* c.1–67C>T alters a GATA1-binding site such that transcription is not initiated and no ACKR1 protein is presented at the RBC surface[192]. This variant is very common in West Africa and has been shown to protect the individual against RBC invasion by *Plasmodium vivax*, i.e. constitutes a resistance factor against malaria[193]. Interestingly, this null phenotype is also correlated to so-called benign ethnic neutropenia, and typing of this regulator is therefore used for diagnostic purposes[194]. It has also been shown how the Duffy blood group protein regulates leukocyte production and serves as a promiscuous chemokine receptor[195].

Recent progress in sequencing technology and epigenetic studies have shown to be an efficient way to understand how erythroid (and other) TFs govern expression of blood group molecules on erythroid (and other) cells. For instance, the GATA1 site at +5.8 kbp in *ABO* intron 1 was revealed by studying the DHS of *ABO,* and then the disruption of the binding variants was confirmed in individuals showing weakened expression of A and B antigens[196,197]. Soon after, a RUNX1 binding site in the proximal region of GATA1 binding sites of *ABO* intron 1 was also shown to regulate *ABO* expression[180]. Mutations of TF binding motifs in *ABO* intron 1 result in weak A or B subtypes such as A_m, B_m, A_3 or B_3, and lead to discrepancies in the blood bank laboratory testing. Variants in the proximal promoter region of *ABO* were also shown to associate with weakened expression of ABO antigens[198-200]. Although the relationship of these variants to TF binding is not established, a KLF motif was proposed[198].

The erythroid master regulator, GATA1 has also been shown to bind to the proximal promoter of *RHD*. A variant that disrupts the motif was found in an individual with decreased strength of agglutination, showing a mixed-field pattern when tested with various anti-D reagents[201]. GATA1 was also shown to co-regulate both *CD99* and *XG* by a single binding motif situated between the two genes[202,203]. The complete *XG* gene is carried by the X chromosome, while the Y chromosome only carries the first 3 exons of the gene and cannot make the full Xg protein product. Thus, GATA1 binding to motifs on both X chromosomes in females homozygous for the wild-type motif exhibit higher expression levels of Xg^a antigen than males also homozygous for the wild-type motif since only one copy of the X chromosome is present. Furthermore, GATA1 was associated with the expression levels of *SMIM1* through massive parallel reporter assay to screen erythroid GWAS variants[204]. The SNV rs1175550 in *SMIM1* intron 2 was identified to be linked to *SMIM1* mRNA and Vel expression levels, but GATA1 did not show direct binding to rs1175550, but rather its frequent working partner TAL1 may be the key player here[205].

The P1 antigen has been shown to have different expression levels among individuals, and the altered start codon was hypothesized to be the cause for the varies expression at first. However, Westman *et al.* demonstrate that it was the loss of binding to RUNX1 by the variant rs5751348G>T that decreases the expression of *A4GALT*, while Yeh *et al*. instead observed that the variable expression is caused by EGR1 binding or not binding to rs5751348 motif[175,206].

Epigenetics in erythroid gene regulation

Epigenetics was first used to describe changes in phenotype through the interactions between genes and their products without changes in the genotype. The prefix *epi* comes from Greek origin, meaning "over" or "outside of". Therefore, epigenetics refers to something that is on top of, or in addition to, the genetic sequences. Today, epigenetics in eukaryotes is composed of the studies of DNA methylation, RNA methylations, ncRNAs, histone modifications, and nucleosome positioning to the three-dimensional configuration of the genome (Figure 10).

DNA methylation

The methylation of DNA is facilitated by DNA methyltransferases to add a methyl (CH_3) group to the cytosine residue[207]. The methylation often occurs to the cytosine preceding guanine (CpG), but a high density of CpG called CpG islands are often free from methylation, and the majority of promoters and house-keeping genes are found embedded in CpG islands[208]. The methylation of CpG islands in promoters or gene body repress gene expression. The transposable elements are often hypermethylated in order to inactivate their replication and insertion which can cause harm to the genome. The hypomethylation of these transposable elements lead to genome instability and can be oncogenic[209].

The expression of A, B, and H antigens on endothelial cells and RBCs can have a drastic decrease in patients with tumors and hematopoietic malignancies[210-212]. It was hypothesized that the methylation of the *ABO* promoter plays a role in *ABO* expression by studying the degree of methylation of cell lines and their expression levels of *ABO*[213,214]. Later, it was confirmed that loss of antigen expression by methylation of the promoter is commonly found in solid tumor and leukemia patients[215-217]. The loss of the ABH antigens is transient and recovery of antigens is usually observed after treatment[218,219].

Today, both DNA sequencing and the detection of DNA methylation can be performed in a single experiment with long-read sequencing technology, both offered by the major players in the field, PacBio and Oxford Nanopore Technology.

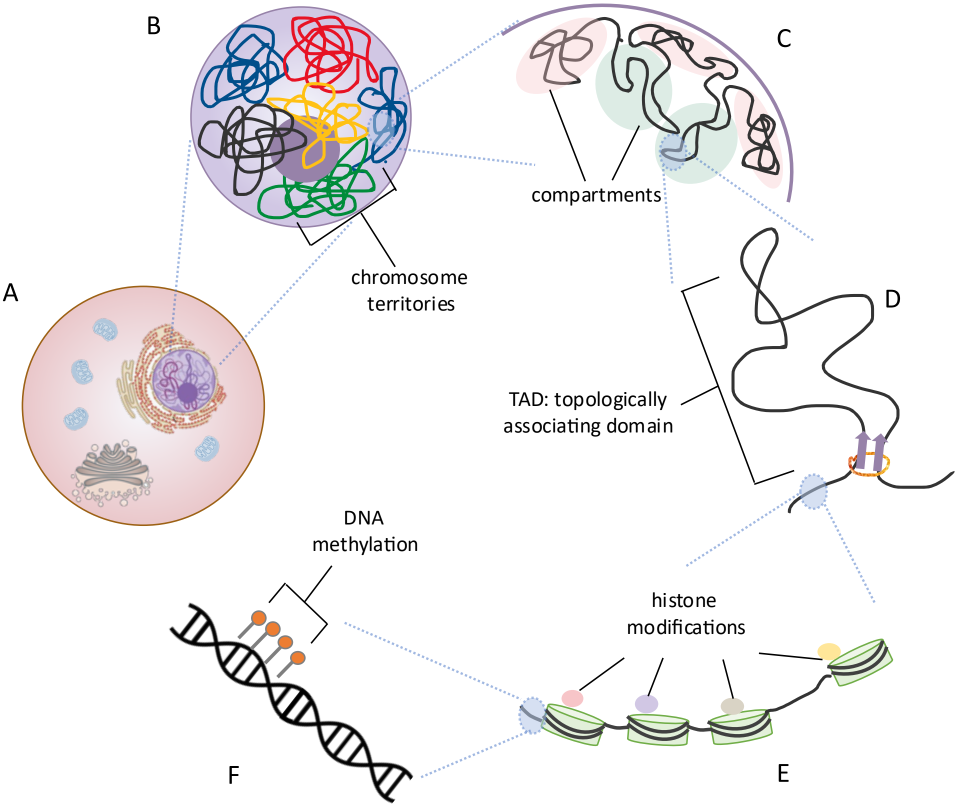


Figure 10. Schematic of nuclear architechture and epigenetic mechanisms concerning gene regulation
A. The human genome consists of 23 pairs of chromosomes, which are located in the nucleus of the cell. **B.** Each chromosome is occupying specific regions of the nuclues called chromosome territories. **C.** The genome is divided into two spatial compartments, A and B, where the chromatin in the A compartment tends to be more open and active and interact preferentially with the chromatin in the A regions, and B being more inactive chromatin and associate preferably with B compartment-associated regions. **D.** TAD, topologically associating domain, is a self-interacting region of the genome. TAD boundaries are separated by CCCTC-binding factor (CTCF) and cohesins. **E.** DNA is wrapped around histones which form the nucleosomes, and the histones can undergo different types of modifications depending on the active or repressed state of the gene. **F.** DNA methylation represses the gene and takes place on the cytosine nucleotide, and are often seen in a cluster as the CpG island.

Histone modifications

The histone is an octamer protein and DNA is wrapped around each histone about 1.65 turn and forms a structure called the nucleosome. The octamer consists of H1, H2A, H2B, H3 and H4 subunits. The H1 subunit functions as the linker and the rest belong to the core superfamily. The modifications of histones include methylation, acetylation, phosphorylation, ubiquitination and sumoylation[220]. While acetylation usually marks the activation of gene transcription, methylation can be either active or repressive depending on the level of methylation, location of the histone subunit, and protein residue of the modification[221].

Among the variety of histone modifications, H3K4me3, H3K4me1 and H3K27ac are constantly marked with positive regulatory elements, promoters and enhancers[222,223]. H3K4me3 modifications are found to be associated with active promoters, and transcriptional activation; H3K4me1 is a mark for both enhancers and promoters, involved in enhancer priming and transcription activation; H3K27ac marks active enhancers and to a less extent, promoters[224]. The identification of the *ABO* intron 1 regulatory region involved the concurrence of H3K4me1 together with H3K9ac to narrow down the window of search for the enhancer. It resulted in a successful application example of histone markers in blood group research[197].

Chromatin accessibility

Chromatin accessibility is the degree of openness of the chromatin to allow physical contact and binding with the nuclear molecules such as TFs or the transcription machinery itself. In a similar way that the binding of TFs and histone modifications work, gene regulation concerning chromatin accessibility is also a temporal and spatial dynamic process. Chromatin accessibility is largely controlled by chromatin remodelers. Their functions include sliding the histone octamer across DNA, changing conformation of nucleosomal DNA and histone octamer composition[225].

The measuring of chromatin accessibility began a half century ago with the use of DNA endonucleases (DNase) to fragment chromatin, revealing the DHSs that are more accessible to these enzymes[226]. Now incorporating the sequencing technology and the enzymatic approaches, many assays have been developed to measure the chromatin accessibility either through cleavage-based methods, including DNase-seq[227], FAIRE-seq (formaldehyde-assisted isolation of regulatory elements with sequencing)[228], MNase-seq (micrococcal nuclease sequencing)[229], and ATAC-seq (assay for transposase-accessible chromatin with sequencing)[230]; or by methyltransferase-based method such as NOMe-seq (nucleosome occupancy and methylome sequencing)[231], Fiber-seq[232], etc[233].

One of the most well studied chromatin accessibility regions is the LCR of the globin genes, designated DHSs. It was found that these erythroid-specific five 5' hypersensitivities are brought closer in proximity to the globin genes by DNA looping[234,235]. DHS also marked *ABO* intron 1 +5.8 kbp as a regulatory region before the GATA1 and RUNX1 sites were discovered[197]. The region marked with DHS was identified as a positive regulatory element through reporter assays in the K562 cell line, and the deletion of the region was found in B_m phenotype individuals.

Three-Dimensional genome interactions

Although we often refer to DNA as a long stretch of string and focus on its sequence and linearity, it is actually very much a three-dimensional complex of folding, looping and interactions between different segments of the genome. The looping and the conformation of the DNA chromatin can bring two segments that are sequence-wise thousands of bps apart, in physically close proximity to interact with one another. As described in Figure 10, the looping of DNA forms a topologically associated domain (TAD) with boundaries set by CCCTC-binding factor (CTCF) sites arranged in convergence as depicted with purple arrows. The looping and the DNA segments brought in close proximity can be measured by what can be called the "many Cs" methods, starting with 3C (chromosome conformation capture), 4C (chromosome conformation capture-on-Chip), 5C (chromosome conformation capture carbon copy) and HiC[236].

Aims of this work

The overall aim of this thesis was to study blood group antigen regulation systematically, with a specific focus on the essential erythroid transcription factor binding sites, and the variants disrupting these binding sites that could lead to weakened or abolished gene transcription and the subsequent decrease of blood group antigen expression on RBCs.

To achieve the overarching aim, the project has broken down into smaller goals as stated below.

1. As proof of principal with GATA1 as the prime example, to identify and characterize TF binding sites within or in proximity to the blood group genes via *in silico* analysis, and to evaluate the GATA1 binding sites and their roles in the blood group antigen expression *in vitro*.

   Paper I

2. To expand our focus beyond GATA1 to multiple erythroid TFs, for which there is data available in the public databases, allowing us to study the interplay of the TFs, and to incorporate epigenetic information as addition layers in the analysis for identifying potential promoters and enhancers.

   Papers II and III

3. To apply our knowledge of TF binding sites to solve discrepant cases in the clinic, where genotype does not mirror the phenotype, and the possibility of studying TF binding and its role in the homologous *RHD* and *RHCE* genes.

   Paper III

Methods

The studies reported in this thesis were carried out as a mix of *in silico* bioinformatics and *in vitro* laboratory experiments. Tables 6 and 7 summarize the tools and methods used for these two aspects of the work.

Table 6. Selection of main softwares and packages for the *In silico* analysis included in this thesis.

Software/Package	Paper	Short description
nextflow (nf) pipeline (ChIPseq, ATACseq)	I, II, III	Prepackaged analysis pipeline for processing raw sequencing data to annotated peaks for ChIP-seq, ATAC-seq experiments, respectively.
MACS2	I	The "callpeak" function to call peaks from merged BAM files.
BEDtools	I, II, III	The "Intersect", "sort", and "getfasta" were used for tasks to intersect and sort BED files or to get fasta sequences from BED files, respectively.
MEME	I	Use function "FIMO" to locate motif with specific TF logotypes.
OpenRefine	I, II, III	For tabular table handling.
Rstudio	I, II, III	Varies packages and libraries for peak analysis , data handling, statistic analysis and figure making (ChIPseeker, tidyverse, dplyr, rtracklayer, Rsamtools, ggplot2, BiocManager, ggrepel, ggpubr, doBy, broom, AICcmodavg, DescTools, ggbio, GenomicRanges, Homo.sapiens, biovizBase, devTools, corrplot, jtools, TxDb.Hsapiens.UCSC.hg38.knownGene, bedr).
Statistical analysis	I, II	Parametric T-test and non-parametric Kruskal-Wallis test were selected to compare mean or median, and distributions between groups. Multiple linear regression analysis was performed to compared two or more independent variable to predict the outcome of a dependent variable.

Table 7. Selection of main *in vitro* assays included in this thesis.

Assay	Paper	Short description
RBC (ghost) membrane preparation	I	RBCs are mixed 1:10 in cold lysis buffer with protease inhibitor and centrifuged to remove hemoglobin. The membranes are washed until the color of pellet turns pale.
Western blot	I, A*	Denatured proteins are separated in tris-glycine gels, and incubated with primary antibody and visualized with HRP-conjugated secondary antibody.
Flow cytometry	I, III, A	Cells either incubated with primary/secondary antibodies carrying specific fluorochrome or transfected with fluorescence protein and measured on flow cytometer.
Fluorescence microscope	I, A	Cells observed and documented directly under microscope with the light source for GFP or RFP on the EVO microscope.
Transfection of plasmid DNA into cells by electroporation	I, II, III, A	Cells are washed and incubated with the plasmids in the cuvette on ice, and electroporation with the BioRad electroporator and quickly placed back in warm medium for recovery.

Assay	Paper	Short description
Transfection of plasmid DNA into cells by lipofectamine	A	DNAs were mixed together with lipofectamine and then added to wells containing cells directly drop-wise in the plate. Change into complete growth medium at 6-24 hours after transfection.
Real-time quantitative PCR/gene expression	I	cDNA was added to the gene expression probe (multiplex for target and control) and measured on the qPCR system QuantStudio 3, Ct value and RQ were compared.
Real-time quantitative PCR/*RHD* zygosity	III	*RHD* exon 7 and *ALB* gene were amplified using gDNA as template, the amplification ratio between the two products was used to determine the zygosity of *RHD*.
Amplification and Sanger sequencing of DNA	I, II, III, A	Template DNA and specific primer pairs were mixed together with polymerase reagents and amplified in thermocycler and visualized on tris-acetate-EDTA/tris-borate-EDTA gels.
Taqman allelic discrimination assay	I, A	Sample DNA was mixed together with Taqman allele probes and assessed on the qPCR system QuantStudio 3 with controls for different allele groups.
EMSA	I, II, III, A	Customized DNA probes were mixed together with nuclear extract and antibody for demonstrating the specific binding of transcription factor.
LC-MS/MS and proteomic analysis	II	Customized DNA probes were mixed together with nuclear extract and assessed by LC-MS/MS on mass spectrometer and the intensity of the peptides measured with label-free Quantification.
DNA cloning	I, II, III, A	A DNA inserts were amplified with customize-designed primers with restriction enzyme sites, both insert and vector were digested by restriction enzyme and then ligated with T4 ligase and transformed into *E. coli* competent cells.
Cell culture	I, II, III, A	Cells are replaced with fresh medium at each passage and incubated in 37 Celsius degrees with 5% carbon dioxide. Cell counts were performed with Countess counter.
Site-directed mutagenesis by overlap extension	I, II, III, A	Template DNA was amplified with customized mutagenic primers introducing mutations in independent, nested PCR, and then the final products were combined together for an extended PCR.
PCR-RFLP	I	PCR is followed by the digestion of restriction enzymes and the fragments are analysed on agarose gel electrophoresis.
gDNA extraction	I, III	Genomic DNA was extracted from peripheral whole blood with DNA columns or magnetic reagents to bind DNA, wash, and then elute into a clean tube.
Total RNA extraction	I	Cells are lysed in trizol overnight. After adding chloroform, the aqueous phase containing RNA is moved to a filter column for washing and collection of the eluate.
Reverse transcription	I	Extracted total RNA is reverse transcribed with rt polymerase to cDNA with a thermocycler.
Limited dilution	A	The transfected cells were diluted to a concentration of 0.8 cell/well and seeded into a 96-well plate for clonal expansion.
Single-cell sorting	A	The cells stained with antibody used for selection of certain populations to be sorted with cell sorter into a 96-well plate using the “single cell” setting.

*A indicates the additional work presented at the end of this thesis.

Summary of results

Paper I highlights

Background

While the genetic basis underlying presence or absence of virtually all clinically relevant blood groups is now known after decades of hard work, our knowledge about quantitative interindividual variation is still very limited. The disruption of a TF binding site leading to weakened or abolished antigen expression has been studied in a few of the blood group systems, such as ABO, Duffy, XG and P1PK. The identification of these TF binding sites was performed independently and often manually, with a particular phenotype in focus. Thus, a systematic approach concerning blood group gene regulation at large has been lacking. Also, previous systematic studies on erythroid gene regulation have focused mostly on erythropoiesis and the globin genes and not on blood groups, or simply lacked phenotype correlation.

Aim

This study aimed to elucidate blood group regulation with a holistic, systematic approach by analyzing available TF ChIP-seq datasets to identify TF binding sites *in silico*, and validate *in vitro* the binding of TF and the effect the TF exerts on blood group phenotype.

In silico *approach and findings*

In this paper, we first set out a search for human GATA1 ChIP-seq datasets. Much considerations were given to the selection and inclusion of datasets. We found many experiments in the Cistrome and the ENCODE databases that were done on the nearly triploid K562 cell line and not on primary cells,[237] and there were duplicates of experiment entries. In order to select only high-quality datasets which would best represent a true human erythroid lineage and which would allow investigation of GATA1 binding in the blood group genes, we included solely the datasets performed on primary erythroblast cells, pass the quality of the sequencing data, and pass our filter for approved positive controls, i.e. contain ChIP-seq peaks at the known GATA1 binding sites in the *ABO* and *ACKR1* genes. The experiment with the most peaks were used as the reference, and all other experiments were intersected with the reference peaks, and then filtered down to the blood group genes, where the TF-binding motifs were located with motif scores.

By combining different packages and their functions for analyzing GATA1 ChIP-seq datasets, we identified 193 potential regulatory sites in close proximity to or within 33 blood-group genes. This combinatory dataset includes some key parameters that enable us to set priorities while examining them. Parameters include:
peak scores - denoting the intensities of the ChIP-seq at a location over the background;
overlaps - how many experiments contains peaks which overlap at the same location;
location - the peak's location in relation to the nearest blood group gene; *motif score* - scores of the sequence compared to the GATA1 motif logotype, higher scores denoting the preferred sequence of GATA1 binding.

In vitro *validation and cohort study*

We applied the GATA1 ChIP-seq result to real blood banking problems and agreed that one of the most variable blood group phenotypes from a quantitative view must be CR1. We therefore set out to find the mechanism for the Helgeson phenotype, the most low-expressing complement receptor 1 (CR1) type on erythrocytes. Among the six GATA1 binding sites in *CR1*, the binding sites in intron 4 showed the highest metrics collectively in the afore-mentioned parameters and was given the highest priority for validation. Through EMSA and luciferase assays, we showed that both candidate motifs in intron 4 bind GATA1 and function as enhancers to drive transcription. Their enhancement abilities are abolished by naturally-occurring SNVs disrupting the GATA1 motifs.

With the inclusion of two different populations study, a Swedish cohort and a Thai cohort, we were able to show that the transcript and protein levels of erythroid *CR1* correlate with genotype of the SNV, rs11117991, in a dose-dependent manner. Linkage disequilibrium (LD) analysis of the SNV rs11117991 with previously proposed genetic markers for Helgeson in the 1000G dataset showed a high LD in Europeans but least in Africans, explaining the poor prediction of previous markers in Africans.

In summary, this study solved the genetic basis and molecular mechanism underlying the Helgeson phenotype, a long-standing enigma in immunohematology. The result enables genetic typing as a possible way to predict this phenotype. Since low CR1 on RBCs has been associated with disease, it is possible that rs11117991 typing may become part of disease susceptibility and/or severity testing.

Paper II highlights

Background

Alongside GATA1, the master regulator in erythropoiesis, many other erythroid TFs have been associated with blood group antigen expression, such as KLF1, RUNX1, TAL1. Many TFs have also been shown to work together, shaping and fine-tuning gene regulation during erythropoiesis. Moreover, TF may enhance gene transcription through the binding to promoters and enhancers. These regulatory regions, promoters and active enhancers are marked with specific histone modifications and are displayed as open chromatin regions.

Aim

We aimed to study datasets allowing us to incorporate multiple TFs so that their effect, alone and/or together when binding to blood group genes could be evaluated. To increase the stringency, we also wanted to include other epigenetic markers to further identify possible enhancer and promoter regions for blood group genes.

In silico *approach and findings*

We expanded our analysis to include ChIP-seq data on adult primary cells of erythroid relevant TFs, KLF1, RUNX1, TAL1, NFE2, and overlay them with epigenetics datasets of ChIP-seq with histone markers, H3K27ac, H3K4me1, indicator for promoters and enhancers, and ATAC-seq for open chromatin regions. We learned that TFs can work interactively, hence, we looked for co-occupancy of TFs using the *mergepeaks* function in Homer. All peaks were filtered to known blood group genes with its genomic coordinates of hg38 extending 10 kbp up- or downstream at each end.

The key findings from this analysis are:

1. Identifying the binding sites of each TF within 10 kbp of blood group genes.
2. Identifying co-occupancy among the TFs analysed.
3. Identifying histone markers for promoter and enhancer regions.
4. Identifying the open chromatin region according to the erythropoiesis stages.

To our surprise, TAL1, the TF that has been shown to interact with GATA1 did not result in any peaks after filtering for blood group genes. We examined the original peak files of TAL1 and found that it only contains 25 peaks across the whole human genome. This is significantly lower than for other TFs, which may indicate a poor experiment output. Thus, this dataset was excluded from the downstream analysis. As a result, a total of 814 potential regulatory sites in 47 blood-group-related genes showed binding for one or more erythroid TFs. Among these sites, intronic regions of *CR1*, *EMP3*, *ABCB6* and *ABCC4* showed co-occupancy of the four remaining TFs, GATA1/KLF1/RUNX1/NFE2. These regions were also identified as enhancer regions by histone markers and open chromatin regions by ATAC-seq. Interestingly, the peak scores of these regions from ATAC-seq across erythropoiesis increases as the cells differentiate into mature RBCs.

In vitro *validation and findings*

We then performed functional analysis of these co-occupancy sites with their presumed enhancer and open chromatin regions at *CR1*, *EMP3*, *ABCB6* and *ABCC4*. The functional study using dual-reporter luciferase assay showed inconsistent results across the four presumed enhancers. *CR1* intron 37 and *ABCC4* intron 1 showed increased gene transcription activity when the intronic regions were analysed compared to the respective promoter region, indicating a functional enhancer. *ABCB6* intron 1 showed no enhancement nor down regulated transcriptional activity. *EMP3* intron 4 showed a slight increase when placed in the forward orientation and conversely a slight decrease when placed in the reverse direction.

The proximal promoter region of *KEL* showed co-occupancy of GATA1 and KLF1. Previous studies reported that GATA1 and Sp1 bind to the *KEL* promoter[238,239], but identification of KLF1 binding was new, at least to our knowledge. TF binding scores were calculated with JASPAR and when naturally-occurring variants were introduced into GATA1 and KLF1 motifs, the binding energy scores decreased compared to the wild-type motifs. However, Sp1 showed a lower-than-default threshold score at the wild-type motif, indicating less likelihood of binding.

We dissected the *KEL* promoter region in a detailed way, making constructs of varying lengths containing different compositions of the motifs, a combination of motifs in their wild-type or disrupted motifs with SNVs. We then confirmed that two of the three GATA1 sites and the KLF1 site are able to drive transcription when in their native states, and the disruption of each single one of them will decrease the promoter's ability to drive gene transcription, and when all the sites were disrupted, transcription dropped to levels similar to the background. We showed increased binding of GATA1 and KLF1 to the wild-type motif over the disrupted motif by EMSA and mass spectrometry (MS).

In summary, by expanding the analysis to include other erythroid-important TFs, and overlaying with epigenetic information of histone modifications and open chromatin regions, we have a list of putative regions to which TFs (co-)bind. The co-occupancy of the TFs allow us to investigate TF interactions with the gene, and a detailed dissection of the clinically significant KEL blood group promoter revealed that co-binding of KLF1 and GATA1 is important to drive transcription from the gene.

Paper III highlights

Background

The Rh blood group system is clinically significant and consists of two highly homologous and polymorphic genes, *RHD* and *RHCE*. Many of the variations have been associated with weakening or atypical expression of RhD, even the RhC/Rhc phenotype is associated with the expression level of RhD (the so-called Ceppellini effect[240]). However, some samples display weakened expression of RhD despite having a normal coding sequence for *RHD*. The weakened expression of antigens could be a result of gene regulation at the transcript level, and possibly regulated by TF binding. GATA1 is the master regulator in erythroid cells and governs the expression of many blood group antigens. It was therefore hypothesized that diminished or abolished binding of GATA1 to a regulatory region in *RHD* could be the reason for the weak expression of RhD antigen in a cohort of 13 samples.

Aim

Here, we aimed to select candidate regulatory region in *RHD* to validate their functionality in binding of GATA1 and subsequently investigate these binding sites in a small cohort of weak D samples with normal coding sequence.

In silico *approach and findings*

The GATA1 binding sites were overlayed with the open chromatin ATAC-seq regions in *RHD* gene, revealed binding to the proximal promoter, intron 1 and intron 2 regions. One GATA1 binding site was located in intron 1 but had no ATAC-seq region and was therefore excluded from further investigation. The ATAC-seq data from the erythroid lineage showed that these regions in *RHD* become open chromatin state beginning at the CFU-E stage and are being kept open throughout differentiation until enucleation. Investigating these ATAC-seq regions further, we found a SNV, rs675072G/A linked to the C/c phenotype residing in intron 2.

In vitro *validation and weak D sample investigations*

We validated the open chromatin region with GATA1 binding in intron 1 as a possible enhancer via luciferase assay. The inclusion of this region enhances gene transcription activity in both the forward and reverse direction. However, for the *RHD* intron 2 region, the C/c-associated SNV, rs675072G/A, influences transcription activity. In the construct containing rs675072G (C phenotype), the luciferase activity was similar to the activity of the promoter regardless of orientation, but when rs675072A is present (c phenotype), both forward and reverse direction of this region enhances the transcription. Thus, the luciferase results in intron 2 appear to mirror the Ceppellini effect.

In one of the weak D samples examined, a novel c.1–110A>C variant was found in the proximal promoter region of *RHD*. This variant disrupted the GATA1 motif located in the promoter, the same GATA1 motif which was disrupted in another weak D sample carrying c.1–115A>C variant reported in a previous study[201]. The luciferase results from the promoter region containing wild-type or the two variants showed that the construct containing c.1–115A>C exhibited similar transcription activity as the wild-type, while the construct of the novel c.1–110A>C showed decreased levels (73%) compared to the wild-type (100%). The loss of GATA1 binding due to c.1–110A>C was confirmed with EMSA.

The sequencing result of the remaining samples showed no variant in the open chromatin regions that would disrupt GATA1 sites in promoter, intron 1 or intron 2 regions. However, *RHD* zygosity test showed atypical results in two samples indicating possible chimeras, each sample has a very small fraction of the RBCs carrying the *RHD* gene. Flow cytometry showed that one of the samples contained both RhD+ (78.4%) and RhD– (20.7%) populations, confirming chimera.

In summary, by studying the GATA1 binding sites and the open chromatin region of *RHD*, we were able to discover a novel variant leading to a Del phenotype, and the intron 2 open chromatin region has greater enhancer ability when containing the SNV of the c haplotype, which may be a lead to how the Ceppellini effect works. In addition, two weak D samples were identified as chimeras, which can be confused with weak D status. This leaves the majority of the cohort, 10 samples of 13, unexplained, requiring future work to solve.

Additional work not included in the papers

Exploration of other possible regulatory regions in *CR1* to explain the Helgeson phenotype in Africans

Background

Even though the SNV rs11117991T>C explaining the Helgeson phenotype is established, it is only found at a low prevalence in the African population. Also, both mRNA and CR1 protein exhibit variable expression levels in individuals carrying the same rs11117991 genotype, hinting that other as yet unidentified regulators may also be at play here as modulating factors.

Aim

We aimed to investigate possible regulatory regions outside of the two GATA1 sites in *CR1* intron 4 to find modulator(s) for gene expression and the genetic background underlying the Helgeson phenotype in the African population.

Main findings

We received 31 African samples from Dr. Joann Moulds from her previous studies on malaria and CR1. A total of 21 out of the 31 samples were successfully genotyped with Taq SNP genotyping assay or by sequencing of a redesigned shorter segment suitable for fragmented DNA. None of these African samples carried the rs11117991T>C variant, as expected since the allele frequency is very low in the African population. We looked into our expanded pipeline from Paper II and identified several possible regulatory regions associated with TF binding in *CR1*, including intron 1 (GATA1 and KLF1 binding sites), intron 4 (GATA1 binding sites, Paper I, and KLF1 binding site), intron 35 (GATA1 sites), and intron 37 (GATA1, KLF1, RUNX1, and NFE2 binding sites, Paper II). Outside the intron 4 GATA1 sites and intron 37 sites for 4 TFs, which was investigated in Paper I and Paper II, we summarized a list of SNVs that could potentially disrupt the TF binding sites in intron 1 and 35 (Table 8). Most minor alleles of the SNVs disrupt the motif, hypothesizing a loss of binding and decreased transcriptional activity. To our surprise, there were two intron 1 SNVs that altered the GATA1 motif and showed increased binding scores, indicating a potential for up-regulating transcript activity. However, all SNVs in intron 1 and 4 had a very low minor allele frequency (MAF), and thus are unlikely to be the common reason for the Helgeson phenotype in the African population. On the contrary, the SNVs in intron 35 showed very high prevalence in the African population compared to all the other super populations. The binding score also suggested the decrease in TF binding to the minor allele variant. However, when we tested these binding sites and the variants with EMSA, we could not validate GATA1 binding, as suggested by the low score for the wild-type, and even lower score for the variant. A possible alternative is to replace EMSA with MS for detecting the binding to the intron 35 sequence. Thus, the quest continues.

Table 8. Possible regulatory TF binding sites in *CR1* and the SNVs within the respective motifs.

Int	TF	WT sequence	Score* -WT	SNV	Score* -SNV	Allele count#	
						ALL	African
1	GATA1	AAT**G**AGATAACAATAC	0.8465	rs951944362G>A	0.8641	2/152198	2/41452
1	GATA1	AATGA**G**ATAACAATAC	0.8465	rs1324888149G>T	0.7464	2/152204	0/41454
1	GATA1	AATGAGATAA**C**AATAC	0.8465	rs1006038858C>G	0.9433	2/152128	0/41414
1	GATA1	TCTCCTTATCAGAAAT	0.8645	NA	NA	NA	NA
1	KLF1	TGGGC**G**GAA	0.8119	rs1430208140G>A	0.6884	3/152174	0/41430
1	GATA1	CCA**A**AGATAAGTGG	0.9273	rs1325151959A>C	0.9175	1/152172	0/41426
4	KLF1	CTCCTC**C**CA	0.8993	rs1274543584C>T	0.7758	2/152170	2/41438
35	GATA1	ATTACG**T**ATCATTGAG	0.7433	rs74153331T>C	0.6384	1229/152290	1160/41554
35	GATA1	GAGCAGATGAGC**A**AAG	0.8071	rs56105884A>G	0.7869	9581/152240	8947/41476

* Number indicating the relative binding energy score calculated by JASPAR 2024 using the TF motif logotype (MA0035.3 for GATA1 and MA0493.2 for KLF1). Last accessed March 1st, 2024.

# Allele count numbers retrieved from gnomAD v4. Last accessed March 1st, 2024.

Abbreviation: Int, intron; WT, wild-type, SNV, single nucleotide variant; ALL, all populations.

The nucleotide in the motif sequence altered by the SNV is highlighted in **underlined bold dark red**.

The alteration of the binding motif from wild-type to the SNV minor allele resulted in the increase of binding scores are highlighted in red.

KEL promoter study with overexpression of GATA1 and KLF1

Background

GATA1 and KLF1 motifs are found in the proximal promoter of *KEL*. Naturally-occurring variants disrupt the motifs, which resulted in decreased transcriptional activity of the promoter region. Since, the GATA1 and KLF1 motifs are shared with other GATA and KLF family TFs, there are possibly other binding motifs that we have not identified. The direct effects of GATA1 and KLF1 bound to the *KEL* promoter should therefore be investigated.

Aim

We aimed to investigate the role of GATA1 and KLF1 in the *KEL* promoter region by over-expression of the TFs to observe the ability of the promoter to drive the gene.

Brief method

Various *KEL* promoter constructs with a reporter gene were used to transfect the HL60 cell line, which expresses minimal innate GATA1 or KLF1. By co-transfecting GATA1 and/or KLF1 vectors, the effect of GATA1 and KLF1 on promoter constructs, both wild-type and variants, was studied by measuring the changes in reporter gene signals.

Main findings

We tested erythroid precursor cell lines which display minimal expression of innate GATA1 and KLF1 and selected HL60, which expressed only low levels of GATA1 and KLF1 (Figure 11A,B). However, HL60 showed low transfection efficiency regardless if they were transfected via electroporation or lipofectamine, with or without albumin deprivation. HEK293 was then selected for its high transfection rate and the expression of innate GATA1 or KLF1 is very minimal, thus potentially ideal for co-transfection.

KEL promoter constructs were made with either wild-type, or natural variants disrupting two GATA1 motifs and one KLF1 motif. The constructs included a segment of scaffold-matrix attachment region (SMAR) and were cloned into a vector containing ZsGreen (a green fluorescent protein) reporter gene. The inclusion of SMAR transforms the vector into becoming an episomal vector and is therefore able to sustain the transfection during cell division/proliferation. As a pilot test, HEL cells were used for monitoring the ability of SMAR to generate stably transfected cells by measuring ZsGreen under the fluorescence microscope and by flow cytometry. Cells transfected with the SMAR vector showed stable transfection during the observation period spanning over three months (Figure 11C).

The cells that were batch-transfected showed a wide distribution of fluorescence intensity, which may be a confounding factor in the GATA1 and KLF1 overexpression analysis. The clonal expansion after limiting dilution is able to mitigate the wide distribution of fluorescent intensity and should be established for each *KEL* promoter construct (Figure 11D) in future studies.

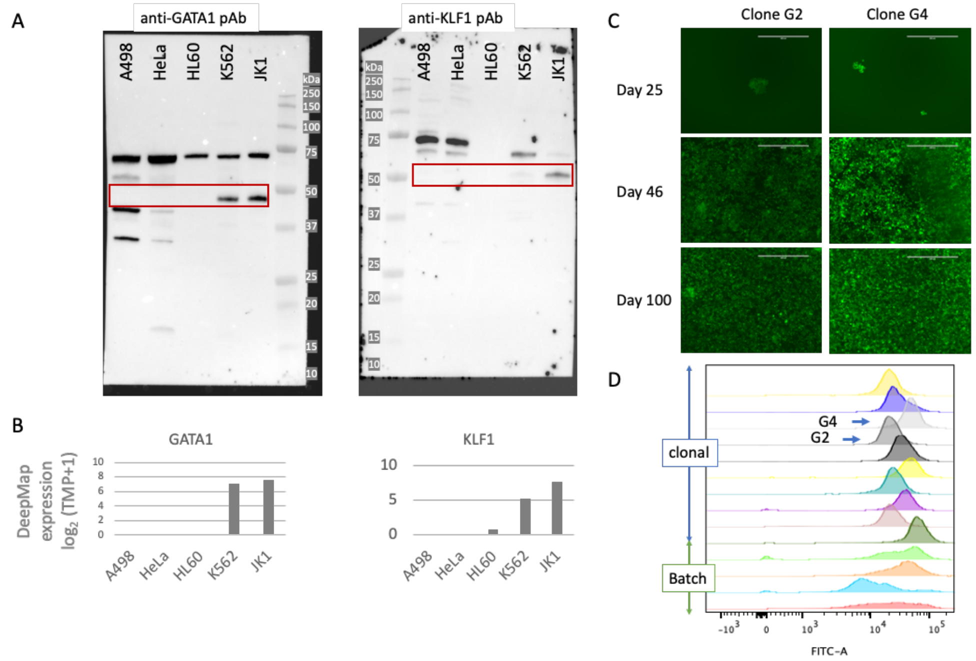


Figure 11. Validation of the essential steps for overexpression of GATA1 and KLF1 for investigation of their effects on the *KEL* promoter.

A. Western blots showing no detectable innate GATA1 or KLF1 expression level in the HL60 cell line. A498 and HeLa cell lines serve as negative controls for both proteins and K562 and JK1 cell lines serve as positive controls. **B.** The RNA expression in selected cell lines (based on data retrieved from DeepMap) correspond to the Western blot results for GATA1 and KLF1. **C.** The image taken with the fluorescence microscope on clonal expansion after transfection of the episomal vector expressing ZsGreen. The days given on the left indicate the day post batch transfection. Although transfected in the same batch, clone G4 shows higher intensity over clone G2. **D.** Flow cytometry analysis of batch and clonal expression of ZsGreen intensity measured on the FITC channel. The batch transfection shows a wide colour intensity distribution, while the individual clones showed much narrower channel distribution. In the histogram indicated for G4 and G2 clones, we can see that G4 has higher intensity over G2, reminiscent of what was observed under the fluorescence microscope (in **C**).

General discussion

A systematic data analysis approach for investigating the blood group regulome

In the cells in our body, the central dogma goes from DNA to RNA to protein. However, since proteins represent the phenotypes that we first observe, our investigation of blood group antigen expression has usually gone in the opposite direction, namely from protein back to DNA, that is, from first observing the phenotypic differences and then investigating the underlying reason in the genotype, often as a research project outside the clinical laboratory and not seldom at considerable cost. With DNA sequencing technology becoming reliable and affordable, it is now increasingly simple and inexpensive to sequence a number of samples of same unusual phenotype and then to compare them to the so called "control" group with another, perhaps more common, phenotype and try to identify the variants associated with the altered antigen expression.

As an innovative approach, the work presented in my thesis goes from genotype to phenotype, by systematically studying the variants in potential regulatory regions of the blood group genes, and investigating if these variants affect antigen expression. Whilst most work up until now has been concentrated on qualitative aspects of blood groups, my work focused on quantitative aspects, i.e. what determines how much of a certain blood group is expressed on the RBCs.

We started to test this approach using the master regulator GATA1 as a model. The reason for choosing GATA1 were at least three-fold: 1) we knew that other blood group phenotypes had been shown to depend on GATA1; 2) GATA1 is one of the most important erythroid TFs; and 3) it was possible to find a reasonable number of high-quality datasets to kickstart the study. These reasons paved the way to Paper I, in which we were able to study the GATA1-driven blood group regulome. We identified many of the blood group genes to have GATA1-binding sites within or in close proximity, confirming the pluripotent and dominant role of GATA1 in erythropoiesis. The proof of principle analysis of GATA1 ChIP-seq data contained many matrices and served as a great database (information center), into which our gene of interest could simply be plugged in. It displays the possible GATA1 sites and their scores concerning peaks, and motifs.

Nevertheless, we understand that there are many other TFs involved in the erythropoiesis process and that they have an intricate interplay with GATA1 and yet other TFs and related nuclear proteins. Some others, not typically identified as erythroid TF, such as RUNX1, were also shown to have a regulating effect on blood group antigens. Although it is better known for megakaryocyte development, RUNX1 has been shown to regulate P1 antigen expression on RBCs. Furthermore, we speculated if we could narrow down further the list of 193 GATA1 targets, to those sites that show indications of promoters or enhancer activity. Thus, we expanded our search to include other key erythroid or known blood group TFs, to overlay them with histone markers for promoters and enhancers, and finally, to overlay the open chromatin region by ATAC-seq on the erythroid lineage. This extension of the work from Paper I led to the studies presented in Papers II and III.

We have now a long list of regulatory region candidates and the binding sites of the selected TFs for the known blood group genes. This enables us to study individual TFs for a specific gene, TFs active on multiple genes, multiple TFs working together in one gene, and even multiple TFs active on multiple genes. The beauty of this approach is the plasticity and flexibility it offers. The analysis can be expanded rapidly as new data are made available, and can be customized to any gene set, not exclusively to just blood group genes. As the new blood group genes and TFs are coming to light, we expect this systematic approach will be a good gateway for us to study gene regulation, especially for blood groups. At the same time, we realize that blood groups for which synthesis happens in non-erythroid cells (e.g. the Lewis and Chido-Rodgers blood group systems) cannot be investigated with exactly the same pipeline.

The molecular mechanism underlying the Helgeson phenotype

The molecular basis underlying the Helgeson phenotype remained an enigma for half a century and we were able to use the output of our data analysis to provide a mechanism for the regulatory aspect for this phenotype. It also became the first proof-of-principle that our pipeline could solve real-world problems. The matrices (peak score, motif scores, number of experiments overlapping at this region, etc.) from our analysis serves as great indicators and led us to *CR1* intron 4.

With the functional *in vitro* assays, I was able to demonstrate that this intron 4 region serves as a powerful enhancer to *CR1* gene transcription in erythroid cells. Interestingly, the two GATA1 motifs in this region tend to be indispensable for each other: when separating the motifs and examining each motif individually, not much of transcript enhancement was observed. However, when the two motifs were examined together, we saw a great boost in transcriptional activity. Moreover, when mutating the motifs to their variant forms, which disrupt GATA1 binding, the transcript levels decreased. The enhancement was totally abolished when both of the motifs were mutated, suggesting the two motifs may be the only essential players in this enhancer region.

Of the two motifs, the SNV rs11117991T>C that disrupted one of the GATA1 binding sites caught our attention based on its allele frequency in the population, which is compatible with the observed prevalence of the Helgeson phenotype. Through linkage disequilibrium analysis, we were able to show the high LD of rs11117991 with previous predictive markers in Caucasians but that it decreased outside this population group. We were then able to conclude both transcriptionally and protein phenotypically that rs11117991T>C is the cause for the low expression of CR1, and not the previously proposed markers. CRISPR-Cas9 technology was explored to mutate rs11117991T>C at an early erythroblast stage, prior to the expression of CR1, and to differentiate the cells to mature RBCs to evaluate the expression of CR1. However, due to the repetitive nature of the *CR1* gene encoding the LHR regions, no sufficiently specific guide RNA sites were available without high off-target effects for this particular intron 4 SNV and this idea was abandoned.

Interestingly, when we examined the LD relationship between rs11117991 and the SNVs determining the Knops antigens, we were able to identify 18 haplotypes across the five super populations (EUR, European; EAS, East Asian; SAS, South Asian; AFR, African; AMR, Ad Mixed American), and that the minor allele causing the Helgeson phenotype, rs11117991C, is only observed in the haplotypes with rs41274768G for Kn(a+), rs17047660A for McC(a+), rs17047661A for Sl1+, and rs4844609T for Sl3+ (Paper I, supplementary figure 1). Since the Helgeson phenotype is defined by very low CR1 expression, it would not be easy to type these antigens serologically, but they can be readily predicted genetically. While matching of the antigens for transfusion purposes is not necessary, since the antibodies against the Knops antigens are typically not clinically significant, the specific Knops antigen phenotype of test RBCs can be very useful information for assisting antibody identification in the reference laboratory.

In addition, even within a group of individuals carrying the same rs11117991 genotype, there is variation both in mRNA levels and protein levels of CR1 expression. This indicates that the level of CR1 on RBCs is not solely controlled by rs11117991, and might be affected by other regulatory sites as well, as we have identified several regions in the expanded analysis beyond Paper I. Moreover, the expression of CR1 on RBCs decreases upon maturation and storage of the cells and is also associated with several disease states upon removal of immune complexes such as SLE and Alzheimer's disease[129,142]. These acquired kind of mechanisms could also explain the wide distribution of CR1 expression within the genotype.

Here, we acknowledge that rs11117991T>C is only one explanation for the Helgeson phenotype and is the prevalent mechanism in Caucasians and Asians. We hypothesize that there are other regulatory mechanisms in the African population that account for the Helgeson phenotype. We have looked into the regulatory sites and identified some SNVs within these regions that could change the TF-binding motifs. We should also be reminded that Helgeson can be acquired as well as inherited, as in some of the examples mentioned above.

The co-occupancy of TFs does not automatically make result in enhanced enhancers

We know that TFs can work solely, as dimers or even collaborate together with other TFs, such as TAL1 and GATA1 in the SCL complex. Therefore, the mapping of TF co-occupancy can potentially reveal the interaction and the synergistic effect exerted on gene transcription. Intuitively, TF co-occupancy in regions marked with enhancer histone markers suggested that the region will act as positive enhancers and increase gene transcription. To our surprise, this was not observed in the four intronic regions we found in *ABCC4*, *EMP3*, *ABCB6* and *EMP3* showing co-occupancy of GATA1, KLF1, RUNX1, and NFE2.

Gene regulation is a continuous process including temporal and spatial aspects. The co-occupancy of the four TFs we included here were taken from different experiments not necessarily reflecting the same time-point or stage of the erythroblast. In other words, it is not necessarily true that all four TFs will exist and bind to the region and regulate at the same time. Moreover, the different factors or co-factors that a specific TF binds to may influence the ability of a TF to upregulate or downregulate. It is also possible that there are other regulators binding in these regions that were not identified with our analysis but that could contribute to the apparently discrepant results.

When comparing the gene expression of *CR1* intron 4 (Paper I) with its two GATA1 binding sites to intron 37 (Paper II) with four TF binding sites, we observed that GATA1 binding in intron 4 resulted in a higher transcript level than that in intron 37, suggesting that the transcription activity is not directly proportional to the number of the TFs binding to the region. It has also been shown that TFs do not necessarily bind on top of each other to maximize the enhancer's abilities to regulate a gene, instead they bind in a sequence with some distance in between. For example, a stretch of 24 kpb in the mouse α-globin gene super-enhancer region contains five separate enhancer regions[241]. It would be interesting to include the approach of stitching a cluster of regulatory regions to identify super-enhancers. However, as pointed out in the same study, the super-enhancers for the α- and β-globin genes are ranked high in the list of super-enhancers they identified, and that none of the five enhancer regions within the super-enhancer region for the α-globin gene is critical for globin expression. Since the globin genes are essential for survival, it is logical that many regulatory machineries may be involved in their expression. As for blood groups, it is not guaranteed that the same super-enhancer-like machinery will exist for all genes, since many blood group molecules are not essential to survival and can be completely absent from an individual without causing any apparent physiological changes or symptoms, such as in RhD-negative individuals.

KEL promoter, TF binding and its expression regulating the master regulator?

The *KEL* promoter region was identified previously, highlighting three GATA1 and one Sp1 binding sites[238,239]. In our study, we concluded that one KLF1 and two GATA1 motifs up-regulate the expression in the proximal promoter region. A recent study revealed that TAL1 is also involved in *KEL* expression with GATA1[242]. Interestingly, in the latter study, the expression level of *KEL* was associated with intermediate or poor risk assessment of AML patients. Knock-down or overexpression of *KEL* in myeloid cell lines affects a specific H3K27ac marker in the proximal promoter of the *GATA1* gene. Therefore, the interaction of TFs, the promoter and the gene expression may be more complicated than our current understanding. Is this through the negative feedback loop or can *KEL* really serve as a modulator and contribute in gene regulation? An interesting question that requires further study.

Investigation of *RHD* and *RHCE* co-regulation

We started the study of Paper III with the focus to investigate the co-regulation of *RHD* and *RHCE*, especially with the ambition to explain the Ceppellini effect, a phenomenon in which RhD antigen is expressed at a lower level when RhC antigen is expressed [240]. The mechanism underlying the Ceppellini effect still remains an enigma to us. From our study of the *RHD* intron 2 ATAC-seq region that showed different levels of transcription enhancement by an R^1- or R^2-associated SNV, we have good ground to hypothesize that the Ceppellini effect may involve gene regulation with respect to variants in the R^1 or R^2 alleles, and potentially differential binding of the TFs. Moreover, HiC data from K562 cells suggested looping of the proximal region upstream of *RHCE* at a kilobase resolution[243], possibly consolidating the hypothesis that the DNA sequence of *RHD* and *RHCE* can be brought to close proximity and possibly involved in the gene regulation to each other.

The expression levels of RhD linked to *RHD*/*RHCE* genotype has been studied with flow cytometry and RhD was shown to be expressed in a dosage-dependent manner to the number of R^2 alleles and was expressed highest in phenotypic R_2R_2 samples. While RhAG does not seem to correlate with RhD expression[244] to our knowledge, no studies have presented work on the transcript levels. However, when we investigated *RHD* and *RHCE* gene expression by qPCR assays in small cohorts consisting of different R^1, R^2, *r* compositions, we could not see a correlation that reflected the Ceppellini effect. Since qPCR is best at detecting gene expression levels exceeding two-fold changes (otherwise the Ct values are too similar), one explanation is that the differences in expression levels of RhD due to R^1 or R^2 allelic differences are below two-fold change, thus is not picked up by this assay. It could also be that the Ceppellini effect has nothing to do with transcript level of *RHD*, and thus explains our negative findings.

The hypothesis that the SNVs associate with the C/c phenotype could alter gene transcription can be further explored by CRISPR editing. We have several SNVs linked to C/c phenotype that are possible candidates for us to edit, changing only one base pair of SNV from a C- to c-associated nucleotide in a R_1r cell line and then to compare the outcome of gene and protein expression levels before and after editing. At the same time, we can also compare the chromatin accessibility, histone modifications and possible transcription factors bindings associated with this base-editing approach.

Methodological considerations

Data selection

The beauty of the experiments used for *in silico* analysis in this thesis is that they are retrieved from publicly available sources, enabling *de novo* analysis of the existing datasets to create new value tailored to investigate our specific aims. The benefit of this reuse of data is that we save time, resources, and are able to select quality datasets if they are available. However, the downside is not to be neglected. Our primary concerns were the availability of the datasets, the type, quality and content of the datasets, the information (or lack thereof) about the primary erythroblast (the stage of the erythroblast, individual's age, sex, health, ethnic groups, etc) to name a few. The exclusion of the TAL1 ChIP-seq data in Paper II was an example of poor data quality, having too few peaks after analysis.

Some recently developed assays such as CUT&RUN-seq and CUT&Tag-seq serve a similar purpose as ChIP-seq[245]. These assays greatly reduce the background noise, and thus require far fewer cells as starting material, as well as being easier to process and more economical. These assays can greatly boost the quality of data when working with erythroid primary cells where cell numbers are normally a limitation. Although there are not many publicly available datasets as yet working specifically with primary erythroblast cells and our TFs of interest within these new assays, we can expect data from these assays would be useful for similar analysis in future studies.

Analysis pipeline inclusion and data management

The *in silico* analysis in this thesis required handling different types of sequencing data and transforming them into sensible peak information for regions concerning TF binding, histone modifications, and chromatin accessibility[246]. There are great resources from the whole bioinformatics community, particularly in Sweden where I received lectures and consultations, such as the National bioinformatics infrastructure in Sweden (NBIS), the SciLifeLab, and the European life-science infrastructure for biological information (ELIXIR). Through the resource experts, I was led to the Nextflow core pipelines containing highly standardized and streamlined analysis that permitted manipulation of some parameter settings to obtain useful data. The downside is that whenever the analysis pipeline breaks, it requires careful examination of the source code, and sometimes the troubleshooting is not so straightforward for someone without a formal bioinformatics training. There were also some datasets such as HiC experiments for identifying 3D genome interaction that I would like to incorporate in our analysis, but I did not have the ability to troubleshoot within the given time. This is really a limitation on my part. Nevertheless, during each analysis process, I was able to try and play with some widely used packages for handling NGS data, sometimes in a trial-and-error spirit, to develop the whole process that would tailor the need to focus on blood group gene regulation.

Other limitations

A great limitation of the analysis in this study is that use of bioinformatic tools with codes derived from various places is not always cohesive and sometimes without proper documentation. This may hinder the sharing and reproducibility of the work. It was not until the second half of my Ph.D. study when we were about to publish the first paper that I realized that record keeping could have been more stringent, and that the lack of version control in some cases impeded the process for publishing in high-impact journals that strive to promote transparency and reproducibility. Having said that, we were able to satisfy the tough requirements of the Nature publishing group, as noted in the Nature portfolio Reporting Summary that accompanies the paper online. We also published the code used in a GitHub deposition in line with reviewer comments received.

Our current approach just scrapes the surface of blood group gene regulation, which is of course a very complex topic to study. Our focus has particularly been on selected erythroid TFs and how they bind or not in or close to the blood group genes. We have not included any of other factors concerning gene regulation as we have mentioned previously, such as DNA methylation, miRNAs, or 3D genome interactions, nor have we integrated genomic, transcriptomic or phenotype data in our analysis. Therefore, although the term blood group regulome is used, it is far from complete. Nevertheless, our study is the first attempt at systematically studying how and why blood group expression varies quantitatively between individuals. As is hinted in the title of this thesis, we have started a process that will have to continue for some time: "elucidating" is to start shining a light on something and we have taken a couple of first steps towards understanding more but there are hundreds of candidate regulator sites to investigate in more detail, only from our two first attempts, so it feels like we have only begun investigating what may well end up being life-long quest.

Future Perspectives

This thesis work is only the beginning and initial stage of the elucidation of the blood group regulome project. The overall aim is to continuously explore the regulome and we had some potential targets laid down in front of us. One of my great interests for studying TFs and blood group antigen expression has to be the mechanism of KLF1 haploinsufficiency resulting in the In(Lu) blood type. As mentioned before, there is no apparent KLF1-binding motif near or within the *BCAM* gene, leading us to hypothesize there is a long-range regulatory element or a secondary effect with unknown intermediate agents at play here. The distal regulatory element(s) can be possibly located by studying the 3D genome to map out possible targets in the same TAD region with *BCAM*. The variant, *KLF1* c.304T>C, previously thought to cause the In(Lu) phenotype has now been invalidated, suggesting there are other mechanisms, not necessarily KLF1 related, that could alter the expression of the Lutheran antigens[247].

A more comprehensive way to study the blood group regulome would be to take advantage of the recent development of the multi-omics at a single cell level[248]. Multi-omics combines the study of the genome, epigenome, transcriptome, proteome and metabolome and each can be performed individually or even simultaneously, for example, genome and transcriptome sequencing (G&T-seq)[249], or simultaneous high-throughput ATAC and RNA expression with sequencing (SHARE-seq), which profiles chromatin accessibility and gene expression in the same single cell[250]. As we know that gene regulation is a temporal and spatial process, it is therefore desirable to integrate the process of the different stages during the continuum of erythropoiesis to identify time-sensitive information for regulatory sites. With combination of the blood group phenotype data from proteomics, these techniques could contribute to solving the genetic background and the carrier protein for the remaining orphan antigens in the 700 and 901series or collections.

To address future perspectives concerning computation and analysis approaches, it is likely that we will see integration of artificial intelligence (AI) specifically within deep learning of neuron networks to study the pattern of gene regulation. This could be general or even have a directed focus on erythroid-specific genes and the model could be applied to solve unknown regulatory regions[251]. Such deep learning analysis methods for different -omic studies are constantly under development. Some methods for identifying regulatory regions are also available, e.g. Deepbind for prediction of DNA-protein binding[252]; DeepSEA for prediction of DHS, TF binding and histone modifications[253]; Expecto to expand on DeepSEA to predict gene expression level[254]; Enformer for gene expression and epigenetic predictions[255]; Akita, GraphReg and Orca to predict 3D genome structure with the latter two including DHS and histone modifications profiling[256-258]. To summarize it in short, we can simply expect a boost in the improvement of AI-assisted methods as -omics data accumulates.

However, as shown by the successful combination of *in silico* and *in vitro* experiments in this thesis, experimental evidence will still be needed to validate and formally prove that computer-aided predictions are indeed correct.

In conclusion, there are still many exciting developments regarding technologies, methods and data that would make the research in the blood group gene regulation much more efficient and cost-effective. With the greater use of personal genomics data in the clinic, these finding can be easily translated and applied to the field of transfusion medicine and other aspects of precision medicine to facilitate identification and matching the correct blood type between patients and donors to ensure transfusion safety.

Acknowledgements

There are many many many people who have made this journey possible. Without all the support and care, I would not be here at all. I would certainly like to take this chance to express my gratitude to many and special thanks to

The **Invisible One** behind all things, I know it is Your sovereign hand that leads me here, you are the way, the reality and the life.

Martin, my main supervisor, for all your support and sharp expertise in the field. Thank you for the invaluable opportunities you have provided me and for allowing me to learn from brilliant minds that are rare to find elsewhere.

Jill, my mentor, the walking immunohematology dictionary (that's how I referred to you to my colleagues back home), you are the one who leads me in the field. Our conversations started with your kindness and generosity, offering your expertise to assist me when at the time I had minimal understanding about blood group genes. Over the course of the years, I appreciate more and more about your caring and loving nature and your remarkable problem-solving abilities. Thank you for standing by my side throughout the entire process.

Magnus, **Yan Quan** and **Karina**, my formal and informal co-supervisors, thank you all for the care towards my growth and well-being, and for generously sharing your expertise with me. Your guidance and mentorship have helped me develop in various ways. I feel very fortunate to have such great talents around me.

The past and current members of the MLO research group, **Annika**, **Åsa**, **Bahram**, **Jennifer**, **Linn**, **Anja**, **Mattias**, **Melissa**, **Amanda**, **Pat**, **Abdul Ghani**, **Stephan**, **Maria**, **Ann-Marie**, **Nysa** and **William**, thank you all for the love and care and the diverse talents you each have to allow the place to be thriving. Thank you also for the fun and the memories that we have shared together during camping, canoeing, BBQs, and much more.

The BMC C14 corridor, especially **John** (and **Liz**), and **Diana** for your roles as next door PIs and for having such wisdom to share, to care for and guide me through the journey. **Johan**, for being an experienced researcher and sharing with me your path and for being a wonderful friend that shares the same passion for badminton. **Geneviève**, for the great caring personality and the encouragement I received from you throughout, you are truly amazing. **Amal**, **Ashmita**, **Kalle**, and **Zahra,** thank you for all the support and the enjoyable times together. I do miss the Wednesday padel, and travel around within Skåne with the gals are wonderful adventures. I am so grateful to have shared those memories with you. **Eva**, **Suvi**, **Magnus**, **Alexandra**, **Ton**, thank you for all nice and friendly chit-chat in the kitchen.

The collaborators **Sudip** and **Jenny**, **Genghis** and **Catherine**, thank you for your expertise that took the work to a higher level. The Stem Cell Center, especially the ones in the **PDP**, thank you for sharing the passion for medical research, it is really inspiring to know you and all your work. I enjoyed the writing sessions we had together and the workshop activities, and special thanks for **Juliane** for keeping me company when writing this thesis. **Niklas** and **Maggie**, whom I shared the early journey of PhD student life with, thank you for being physical friends to me during the digital COVID period. **Anna** and **Zhi** for sharing your wonderful expertise in the FACS core unit. Thank you for being so warm and inviting. The **MentLife** mentees and mentors, especially my mentor from both PDP and MentLife, **Deniz** and **Roni**, thank you for sharing with me your life experiences, it is wonderful to have a conversation partner who has "been there, done that."

To the friends I got to know here in Lund, the ones who kept and keep my physical body healthy, my great badminton gang, **Au**, **Thong**, **Tony**, **Adde**, 王曉,周曄,小包 ,常總, thank you for all the great fun on court and outside the court. The ones who care for me and feed me often, 虹姐, **Jill**, **Emily**, **Yuko**, thank you for all the great conversations, helping me to establish a life here, and always satisfying my desire for good food. 小邱 and 韻庭, thank you for the friendship you have given me, all the care and love for the past four and half years, and thank you for sharing life journeys together, I feel so blessed to have you two as my great friends.

To my friends, colleagues and mentors back home, 瑞川,奐綺,明鴻,蕾芳,秀芳,淑惠姐,慶哥,耀哥,佩璇,家菩,么么,志信,嫵茹,齡苡,伊婷,翠微,凱楝,前輝,玥玲, I enjoyed so much working with you and it was not easy to leave that, I am truly thankful to have you not just as my colleagues but as my friends who have cheered me on along the way, being so supportive to me even that we are a whole eurasia continent apart, and showered me with love whenever I return. 陳沛隆醫師,陳定平老師,白組長, who encouraged me to pursue the study even before it all began.

My dear friends, 丹丹丹,代容,庭安, thank you for always caring for my life and well-being, continuously sending your love and good wishes, and being my solid support when I needed it. 賴爸,賴媽,周媽 thank you for taking care of me throughout the years and treating me as your own daughter, you are really my family.

My family in the church, my companion and the one who served me even before I set foot in Sweden, **Karen** and **Wei** and the three little ones, **Johan**, **Caleb** and **Bethany**. Thank you for being a solid family standing firmly in the Lord, and it was really through your persistent prayers that paved the way for the Lord to move me here.小燕,傳帥,旗旗,祝林,家榕, **Xiamara**, **Esther**, and **Ziyu**, thank you for sharing your life here in Lund with me and pursuing the Lord together, I enjoyed all the hymns we sang together, all the love feasts we had and the prayers for each other. The dear ones in **Church in Copenhagen**, thank you all and special thanks to **Priscilla** (and **Peter**), **Phoebe**, **Payal**, **Tanya**, **Camelia** and **Stefan**, **Ioana** and **Dan**, **Ploy**, who faithfully stayed with the Lord every week for all the bible reading, life study reading and prayers. I cannot survive without eating these spiritual foods regularly with you all. Thank you for lifting me up with your prayers whenever I am weak and down, and for standing with me to fight the spiritual warfare as one body. Thanks to **John** and **Penny** in Linköping, thank you for supplying and strengthening me with your visits, chats, and prayers. Brothers and sisters in **Church in Stockholm**, and **Church in Uppsala**, especially brother **Håkan** and sister **Margit**, brother **Chin-Fu** and sister **Yvonne**, brother **Dave** and sister **Cathy**, brother **Bosse** and sister **Eivor**, thank you for caring for me long before I was married to Tomas, and displaying wonderful patterns to serve the Lord as a household for us to follow. Thank you for pouring out life on us, particularly sister **Yvonne**, thank you for sustaining me with our weekly fellowship, nourishing and shepherding me according to God's ordained way, and thank you for all the fellowship concerning family life, academic life, which has helped me overcome many difficulties during the process. Thanks to all the brothers and sister back home in Taiwan, 柏如得明,婉珺邦哥,靖茹忠霖,怡璇,芳生,易萱,妍君,秀芬,皖婷,厚恩,思婷 and many that I cannot name them all, thank you for sending me love packages, recording your prayers for me, writing encouraging cards to me, and sustaining me with your prayers. A special thank you to 茗先姐, who passed away during the study, who had served me and took care of me as a big sister, thank you for showing me that life can be short but full of purpose if you just love the Lord.

My family back home in Taiwan, US, and Japan, to my biggest and firmest supporter in life, my brother **David**, always supporting me in every possible way, you are the best older brother one can ask for. My **grandma** and **dad**, it was not easy at first for you to be comfortable with me leaving my stable job and pursuing something up in the air, but thank you for turning around (which I understand it is totally not easy to do when you get older) and being the loving and supportive (grand)parent figure to me. Mom **Ellen**, **Catherine**, **Richard**, thank you for the love and care you sent me always, and for accepting me for who I am, allowing me to have a place to rest and call home whenever coming back to the US. My **uncles**, **aunts**, **cousins** and their **spouses**, thank you all so much for being always so welcoming and caring, and being such a pattern living according to the principle of life for me. I see all the great human virtues expressed on each one of you, perseverance, loving, caring, faithful, willing to serve, always rather pouring out wine than to keep wine. Special thanks to aunt **Ruby -** this study in Sweden would not have happened without you encouraging me to care for the need of others. Thank you for always caring for my inner and outer life, leading me to the tree of life and being so positive and encouraging all the time. Special memory to my uncle 清海叔叔, who also passed away during this study, thank you for being a great pattern of loving the Lord, caring for the family and saints, and always striving to share truth and life with us. I thank the Lord that I was able to visit you and had the chance to hear you sharing about the end of this Age not long before He took you to be by His side. You are dearly missed.

Last but not least, my husband **Tomas**, thank you for being everything that I am not, which allows the Lord to transform me and to complement me. My new found family in Sweden, **Marianne**, **Matti**, **Stina**, **Jenny**, **Kenneth**, **Alvar**, **Tuva**, **Björn**, and **Leo**, thank you for receiving me into the family warmly and openly.

References

1. Ribatti, D. William Harvey and the discovery of the circulation of the blood. *J Angiogenes Res* **1**, 3 (2009).
2. Roux, F.A., Sai, P. & Deschamps, J.Y. Xenotransfusions, past and present. *Xenotransplantation* **14**, 208-16 (2007).
3. 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* **10**, 296-311 (1819).
4. Waller, C. Case of Uterine Hemorrhage, in Which the Operation of Transfusion Was Successfully Performed. *Lond Med Phys J* **54**, 273-277 (1825).
5. Hughes-Jones, N.C. & Gardner, B. Red cell agglutination: the first description by Creite (1869) and further observations made by Landois (1875) and Landsteiner (1901). *Br J Haematol* **119**, 889-93 (2002).
6. Landsteiner, K. Zur Kenntnis der antifermentativen, lytischen und agglutinierenden Wirkungen des Blutserums und der Lymphe. *Centralblatt für Bakteriologie, Parasitenkunde und Infektionskrankheiten* (1900).
7. Reuben Ottenberg, D.J.K. Accidents in Transfusion -Their prevention by preliminary blood examination: Based on an experience of one hundred twenty-eight transfusions. *JAMA* **61**, 2138-2140 (1913).
8. Sharma, S., Sharma, P. & Tyler, L.N. Transfusion of blood and blood products: indications and complications. *Am Fam Physician* **83**, 719-24 (2011).
9. Szczepiorkowski, Z.M. & Dunbar, N.M. Transfusion guidelines: when to transfuse. *Hematology Am Soc Hematol Educ Program* **2013**, 638-44 (2013).
10. Hervig, T.A. *et al.* Re-introducing whole blood for transfusion: considerations for blood providers. *Vox Sang* **116**, 167-174 (2021).
11. Mark H. Yazer, J.N.S., Andrew Beckett, Darrell J. Triulzi, Philip C. Spinella. Rebirth of the cool: the modern renaissance of low titer group O whole blood for treating massively bleeding civilian patients. *Annals of Blood* **7**(2022).
12. Burgstaler, E.A. Blood component collection by apheresis. *J Clin Apher* **21**, 142-51 (2006).
13. Cho, H.J. *et al.* COVID-19 transmission and blood transfusion: A case report. *J Infect Public Health* **13**, 1678-1679 (2020).
14. Bassil, J., Rassy, E. & Kattan, J. Is blood transfusion safe during the COVID-19 pandemic? *Future Sci OA* **6**, FSO626 (2020).
15. Vamvakas, E.C. & Blajchman, M.A. Transfusion-related mortality: the ongoing risks of allogeneic blood transfusion and the available strategies for their prevention. *Blood* **113**, 3406-17 (2009).
16. Goel, R., Tobian, A.A.R. & Shaz, B.H. Noninfectious transfusion-associated adverse events and their mitigation strategies. *Blood* **133**, 1831-1839 (2019).

17. Taleghani, B.M. & Heuft, H.G. Hemovigilance. *Transfus Med Hemother* **41**, 170-1 (2014).

18. de Jonge, L.L., Wiersum-Osselton, J.C., Bokhorst, A.G., Schipperus, M.R. & Zwaginga, J.J. Haemovigilance: current practices and future developments. *Annals of Blood* **7**, 23-23 (2022).

19. Edgren, G. *et al.* The new Scandinavian Donations and Transfusions database (SCANDAT2): a blood safety resource with added versatility. *Transfusion* **55**, 1600-6 (2015).

20. Edgren, G. & Hjalgrim, H. Epidemiology of donors and recipients: lessons from the SCANDAT database. *Transfus Med* **29 Suppl 1**, 6-12 (2019).

21. Zhao, J., Rostgaard, K., Hjalgrim, H. & Edgren, G. The Swedish Scandinavian donations and transfusions database (SCANDAT3-S) - 50 years of donor and recipient follow-up. *Transfusion* **60**, 3019-3027 (2020).

22. Blood safety and availability. Vol. 2024 (World Health Organisation, 2023).

23. Raykar, N.P. *et al.* Assessing the global burden of hemorrhage: The global blood supply, deficits, and potential solutions. *SAGE Open Med* **9**, 20503121211054995 (2021).

24. Liu, W.J. *et al.* An imbalance in blood collection and demand is anticipated to occur in the near future in Taiwan. *J Formos Med Assoc* **121**, 1610-1614 (2022).

25. Goldstein, J., Siviglia, G., Hurst, R., Lenny, L. & Reich, L. Group B erythrocytes enzymatically converted to group O survive normally in A, B, and O individuals. *Science* **215**, 168-70 (1982).

26. Liu, Q.P. *et al.* Bacterial glycosidases for the production of universal red blood cells. *Nat Biotechnol* **25**, 454-64 (2007).

27. Pellegrin, S., Severn, C.E. & Toye, A.M. Towards manufactured red blood cells for the treatment of inherited anemia. *Haematologica* **106**, 2304-2311 (2021).

28. Dias, J. *et al.* Generation of red blood cells from human induced pluripotent stem cells. *Stem Cells Dev* **20**, 1639-47 (2011).

29. Trakarnsanga, K. *et al.* An immortalized adult human erythroid line facilitates sustainable and scalable generation of functional red cells. *Nat Commun* **8**, 14750 (2017).

30. Thompson, A.A. *et al.* Gene Therapy in Patients with Transfusion-Dependent beta-Thalassemia. *N Engl J Med* **378**, 1479-1493 (2018).

31. Mirmiran, A. *et al.* Erythroid-Progenitor-Targeted Gene Therapy Using Bifunctional TFR1 Ligand-Peptides in Human Erythropoietic Protoporphyria. *Am J Hum Genet* **104**, 341-347 (2019).

32. K. Landsteiner, P.L. Further Observations on Individual Differences of Human Blood. *J Exp Med* **24**(1927).

33. NobelPrize.org. The Nobel Prize in Physiology or Medicine 1930 Vol. 2024 (Nobel Prize Outreach AB 2024).

34. Jonathan De Oliveira Rios, A.S., Romain Duval, Alexandre Raneri, Thomas Poyot, Jérome Babinet, Mariane De Montalembert, Claudia Regina Bonini Domingos, Caroline Le Van Kim, Marc Romana, Thierry Peyrard, Slim Azouzi. The Cs a and Cs b Red Cell Antigens of the Cost Blood Group Collection Correspond to the HNA-3a and HNA-3b Neutrophil Antigens: Unexpected Twins with Implications for Sickle Cell Anemia. *Blood* **142**, 698 (2023).

35. Alattar, A.G., Storry, J.R. & Olsson, M.L. Evidence that CD36 is expressed on red blood cells and constitutes a novel blood group system of clinical importance. *Vox Sang* (2024).
36. Yamamoto, F.-i., Clausen, H., White, T., Marken, J. & Hakomori, S.-i. Molecular genetic basis of the histo-blood group ABO system. *Nature* **345**, 229 (1990).
37. Daniels, G. The molecular genetics of blood group polymorphism. *Hum Genet* **126**, 729-42 (2009).
38. Kominato, Y., Sano, R., Takahashi, Y., Hayakawa, A. & Ogasawara, K. Human ABO gene transcriptional regulation. *Transfusion* **60**, 860-869 (2020).
39. Wertheimer, S.P. & Barnwell, J.W. Plasmodium vivax interaction with the human Duffy blood group glycoprotein: identification of a parasite receptor-like protein. *Exp Parasitol* **69**, 340-50 (1989).
40. Reich, D. *et al.* Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. *PLoS Genet* **5**, e1000360 (2009).
41. Leffler, E.M. *et al.* Resistance to malaria through structural variation of red blood cell invasion receptors. *Science* **356**(2017).
42. Cserti, C.M. & Dzik, W.H. The ABO blood group system and Plasmodium falciparum malaria. *Blood* **110**, 2250-8 (2007).
43. Hsu, K. *et al.* Expedited CO_2 respiration in people with Miltenberger erythrocyte phenotype GP.Mur. *Sci Rep* **5**, 10327 (2015).
44. Palma, J., Tokarz-Deptula, B., Deptula, J. & Deptula, W. Natural antibodies - facts known and unknown. *Cent Eur J Immunol* **43**, 466-475 (2018).
45. Poole, J. & Daniels, G. Blood group antibodies and their significance in transfusion medicine. *Transfus Med Rev* **21**, 58-71 (2007).
46. Morath, C., Zeier, M., Dohler, B., Opelz, G. & Susal, C. ABO-Incompatible Kidney Transplantation. *Front Immunol* **8**, 234 (2017).
47. Li, H.Y. & Guo, K. Blood Group Testing. *Front Med (Lausanne)* **9**, 827619 (2022).
48. Claudia Cohn, M.D., Susan T. Johnson, Louis M. Katz, Joseph (Yossi) Schwartz. *AABB Technical Manual*, (AABB, 2023).
49. Ahrens, N., Pruss, A., Kiesewetter, H. & Salama, A. Failure of bedside ABO testing is still the most common cause of incorrect blood transfusion in the Barcode era. *Transfus Apher Sci* **33**, 25-9 (2005).
50. Jongruamklang, P. *et al.* Characterization of *GYP*Mur* and novel *GYP*Bun*-like hybrids in Thai blood donors reveals a qualitatively altered s antigen. *Vox Sang* **115**, 472-477 (2020).
51. Ji, Y. *et al.* Patients with Asian-type DEL can safely be transfused with RhD-positive blood. *Blood* **141**, 2141-2150 (2023).
52. Fukumori, Y., Ohnoki, S., Shibata, H., Yamaguchi, H. & Nishimukai, H. Genotyping of ABO blood groups by PCR and RFLP analysis of 5 nucleotide positions. *Int J Legal Med* **107**, 179-82 (1995).
53. Wagner, F.F. & Flegel, W.A. RHD gene deletion occurred in the Rhesus box. *Blood* **95**, 3662-8 (2000).
54. Tanaka, M. *et al.* RHC/c genotyping based on polymorphism in the promoter region of the RHCE gene. *Leg Med (Tokyo)* **3**, 205-12 (2001).

55. Ugozzoli, L. & Wallace, R.B. Application of an allele-specific polymerase chain reaction to the direct determination of ABO blood group genotypes. *Genomics* **12**, 670-4 (1992).
56. Olsson, M.L. & Chester, M.A. A rapid and simple ABO genotype screening method using a novel B/O2 versus A/O2 discriminating nucleotide substitution at the ABO locus. *Vox Sang* **69**, 242-7 (1995).
57. Olsson, M.L. *et al.* Genomic analysis of clinical samples with serologic ABO blood grouping discrepancies: identification of 15 novel A and B subgroup alleles. *Blood* **98**, 1585-93 (2001).
58. Daniels, G., Finning, K., Martin, P. & Summers, J. Fetal RhD genotyping: a more efficient use of anti-D immunoglobulin. *Transfus Clin Biol* **14**, 568-71 (2007).
59. Pirenne, F., Floch, A. & Habibi, A. How to avoid the problem of erythrocyte alloimmunization in sickle cell disease. *Hematology Am Soc Hematol Educ Program* **2021**, 689-695 (2021).
60. Gleadall, N.S. *et al.* Development and validation of a universal blood donor genotyping platform: a multinational prospective study. *Blood Adv* **4**, 3495-3506 (2020).
61. Lane, W.J. *et al.* Automated typing of red blood cell and platelet antigens: a whole-genome sequencing study. *Lancet Haematol* **5**, e241-e251 (2018).
62. Blais, J. *et al.* Risk of Misdiagnosis Due to Allele Dropout and False-Positive PCR Artifacts in Molecular Diagnostics: Analysis of 30,769 Genotypes. *J Mol Diagn* **17**, 505-14 (2015).
63. Goodwin, S., McPherson, J.D. & McCombie, W.R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**, 333-51 (2016).
64. Orzińska, A. Next generation sequencing and blood group genotyping: a narrative review. *Annals of Blood* **8**, 4-4 (2023).
65. Lane, W.J. *et al.* Comprehensive red blood cell and platelet antigen prediction from whole genome sequencing: proof of principle. *Transfusion* **56**, 743-54 (2016).
66. Schoeman, E.M., Roulis, E.V., Perry, M.A., Flower, R.L. & Hyland, C.A. Comprehensive blood group antigen profile predictions for Western Desert Indigenous Australians from whole exome sequence data. *Transfusion* **59**, 768-778 (2019).
67. Wu, P.C. *et al. ABO* genotyping with next-generation sequencing to resolve heterogeneity in donors with serology discrepancies. *Transfusion* **58**, 2232-2242 (2018).
68. Rieneck, K., Clausen, F.B. & Dziegiel, M.H. Noninvasive Antenatal Determination of Fetal Blood Group Using Next-Generation Sequencing. *Cold Spring Harb Perspect Med* **6**, a023093 (2015).
69. Tounsi, W.A., Madgett, T.E. & Avent, N.D. Complete RHD next-generation sequencing: establishment of reference RHD alleles. *Blood Adv* **2**, 2713-2723 (2018).
70. Wheeler, M.M. *et al.* Genomic characterization of the RH locus detects complex and novel structural variation in multi-ethnic cohorts. *Genet Med* **21**, 477-486 (2019).

71. Lopez, G.H. *et al.* Frequency of Mi(a) (MNS7) and Classification of Mi(a)-Positive Hybrid Glycophorins in an Australian Blood Donor Population. *Transfus Med Hemother* **47**, 279-286 (2020).
72. Cvejic, A. *et al.* SMIM1 underlies the Vel blood group and influences red blood cell traits. *Nat Genet* **45**, 542-545 (2013).
73. Omae, Y. *et al.* Integrative genome analysis identified the KANNO blood group antigen as prion protein. *Transfusion* **59**, 2429-2435 (2019).
74. Stenfelt, L. *et al.* Missense mutations in the C-terminal portion of the *B4GALNT2*-encoded glycosyltransferase underlying the Sd(a-) phenotype. *Biochem Biophys Rep* **19**, 100659 (2019).
75. Zhang, W. *et al.* Comparing genetic variants detected in the 1000 genomes project with SNPs determined by the International HapMap Consortium. *J Genet* **94**, 731-40 (2015).
76. Koehl, B. *et al.* Lack of the human choline transporter-like protein SLC44A2 causes hearing impairment and a rare red blood phenotype. *EMBO Mol Med* **15**, e16320 (2023).
77. Azouzi, S. *et al.* Lack of the multidrug transporter MRP4/ABCC4 defines the PEL-negative blood group and impairs platelet aggregation. *Blood* **135**, 441-448 (2020).
78. Thornton, N. *et al.* Disruption of the tumour-associated *EMP3* enhances erythroid proliferation and causes the MAM-negative phenotype. *Nat Commun* **11**, 3569 (2020).
79. Lane, W.J. *et al.* PIGG defines the Emm blood group system. *Sci Rep* **11**, 18545 (2021).
80. H Sugier, C.V., R Duval, C Le Van Kim, C Arnoni, T Vendrame, F Latini, R De Medeiros, L Castilho, S Azouzi, T Peyrard. Null allele of *ABCC1* encoding the multidrug resistance protein 1 defines a novel human blood group system. *Vox Sang* **115 Suppl 1**, 5-396 (2020).
81. Daniels, G. & Reid, M.E. Blood groups: the past 50 years. *Transfusion* **50**, 281-9 (2010).
82. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K.F. Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* **39**, 1348-1365 (2021).
83. Tounsi, W.A. *et al.* Rh Blood Group D Antigen Genotyping Using a Portable Nanopore-based Sequencing Device: Proof of Principle. *Clin Chem* **68**, 1196-1201 (2022).
84. Zhang, Z. *et al.* Accurate long-read sequencing allows assembly of the duplicated RHD and RHCE genes harboring variants relevant to blood transfusion. *Am J Hum Genet* **109**, 180-191 (2022).
85. Srivastava, K. *et al.* ACKR1 Alleles at 5.6 kb in a Well-Characterized Renewable US Food and Drug Administration (FDA) Reference Panel for Standardization of Blood Group Genotyping. *J Mol Diagn* **22**, 1272-1279 (2020).
86. Fichou, Y. *et al.* Defining Blood Group Gene Reference Alleles by Long-Read Sequencing: Proof of Concept in the ACKR1 Gene Encoding the Duffy Antigens. *Transfus Med Hemother* **47**, 23-32 (2020).

87. Menegati, S.F.P., Santos, T.D., Macedo, M.D. & Castilho, L. Discrepancies between red cell phenotyping and genotyping in daily immunohematology laboratory practice. *Transfus Apher Sci* **59**, 102585 (2020).

88. Landsteiner, K. & Wiener, A.S. STUDIES ON AN AGGLUTINOGEN (Rh) IN HUMAN BLOOD REACTING WITH ANTI-RHESUS SERA AND WITH HUMAN ISOANTIBODIES. *J Exp Med* **74**, 309-20 (1941).

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

90. Edwards, A.W. R. A. Fisher's 1943 unravelling of the Rhesus blood-group system. *Genetics* **175**, 471-6 (2007).

91. Westhoff, C.M. The structure and function of the Rh antigen complex. *Semin Hematol* **44**, 42-50 (2007).

92. Sturgeon, P. Hematological observations on the anemia associated with blood type Rhnull. *Blood* **36**, 310-20 (1970).

93. Westhoff, C.M., Ferreri-Jacobia, M., Mak, D.O. & Foskett, J.K. Identification of the erythrocyte Rh blood group glycoprotein as a mammalian ammonium transporter. *J Biol Chem* **277**, 12499-502 (2002).

94. D'Alessandro, A. *et al.* Genetic polymorphisms and expression of Rhesus blood group RHCE are associated with 2,3-bisphosphoglycerate in humans at high altitude. *Proc Natl Acad Sci U S A* **121**, e2315930120 (2024).

95. Cherif-Zahar, B. *et al.* Localization of the human Rh blood group gene structure to chromosome region 1p34.3-1p36.1 by in situ hybridization. *Hum Genet* **86**, 398-400 (1991).

96. Suto, Y., Ishikawa, Y., Hyodo, H., Uchikawa, M. & Juji, T. Gene organization and rearrangements at the human Rhesus blood group locus revealed by fiber-FISH analysis. *Hum Genet* **106**, 164-71 (2000).

97. Carritt, B., Kemp, T.J. & Poulter, M. Evolution of the human RH (rhesus) blood group genes: a 50 year old prediction (partially) fulfilled. *Hum Mol Genet* **6**, 843-50 (1997).

98. ISBT. Guidelines Naming RH Alleles. v2.0 edn (2014).

99. Villalba, A. *et al.* Anti-D Alloimmunization after RhD-Positive Platelet Transfusion in RhD-Negative Women under 55 Years Diagnosed with Acute Leukemia: Results of a Retrospective Study. *Transfus Med Hemother* **45**, 162-166 (2018).

100. Vos, G.H., Petz, L.D., Garratty, G. & Fudenberg, H.H. Autoantibodies in acquired hemolytic anemia with special reference to the LW system. *Blood* **42**, 445-53 (1973).

101. Dacie, J.V. Autoimmune hemolytic anemia. *Arch Intern Med* **135**, 1293-300 (1975).

102. Gehrs, B.C. & Friedberg, R.C. Autoimmune hemolytic anemia. *Am J Hematol* **69**, 258-71 (2002).

103. Clausen, F.B. *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* **52**, 752-8 (2012).

104. WiRhE. Worldwide Initiative for Rh Disease Eradication (WIRhE). in *Worldwide Initiative for Rh Disease Eradication (WIRhE)* Vol. 2024 (Worldwide Initiative for Rh Disease Eradication (WIRhE), 2022).

105. Sandler, S.G. *et al.* It's time to phase in RHD genotyping for patients with a serologic weak D phenotype. College of American Pathologists Transfusion Medicine Resource Committee Work Group. *Transfusion* **55**, 680-9 (2015).

106. Flegel, W.A., Khull, S.R. & Wagner, F.F. Primary anti-D immunization by weak D type 2 RBCs. *Transfusion* **40**, 428-34 (2000).

107. Marsh, W.L. & Redman, C.M. The Kell blood group system: a review. *Transfusion* **30**, 158-67 (1990).

108. Yung, C.H., Chow, M.P., Hu, H.Y., Mou, L.L. & Lyou, J.Y. [Blood group phenotyping and their application in Taiwan]. *Zhonghua Yi Xue Za Zhi (Taipei)* **43**, 345-54 (1989).

109. Lin, M. & Broadberry, R.E. Immunohematology in Taiwan. *Transfus Med Rev* **12**, 56-72 (1998).

110. Lee, S. *et al.* Proteolytic processing of big endothelin-3 by the kell blood group protein. *Blood* **94**, 1440-50 (1999).

111. Lee, S., Russo, D. & Redman, C. Functional and structural aspects of the Kell blood group system. *Transfus Med Rev* **14**, 93-103 (2000).

112. Wimer, B.M., Marsh, W.L., Taswell, H.F. & Galey, W.R. Haematological changes associated with the McLeod phenotype of the Kell blood group system. *Br J Haematol* **36**, 219-24 (1977).

113. Redman, C.M. *et al.* Biochemical studies on McLeod phenotype red cells and isolation of Kx antigen. *Br J Haematol* **68**, 131-6 (1988).

114. Poole, J. *et al.* A KEL gene encoding serine at position 193 of the Kell glycoprotein results in expression of KEL1 antigen. *Transfusion* **46**, 1879-85 (2006).

115. Kormoczi, G.F., Scharberg, E.A. & Gassner, C. A novel KEL*1,3 allele with weak Kell antigen expression confirming the cis-modifier effect of KEL3. *Transfusion* **49**, 733-9 (2009).

116. Yazdanbakhsh, K., Lee, S., Yu, Q. & Reid, M.E. Identification of a defect in the intracellular trafficking of a Kell blood group variant. *Blood* **94**, 310-8 (1999).

117. Velliquette, R.W. *et al.* Molecular basis of two novel and related high-prevalence antigens in the Kell blood group system, KUCI and KANT, and their serologic and spatial association with K11 and KETI. *Transfusion* **53**, 2872-81 (2013).

118. Ohto, H. *et al.* Three non-classical mechanisms for anemic disease of the fetus and newborn, based on maternal anti-Kell, anti-Ge3, anti-M, and anti-Jr(a) cases. *Transfus Apher Sci* **59**, 102949 (2020).

119. Rieneck, K., Clausen, F.B. & Dziegiel, M.H. Next-Generation Sequencing for Antenatal Prediction of KEL1 Blood Group Status. *Methods Mol Biol* **1310**, 115-21 (2015).

120. Vaughan, J.I. *et al.* Erythropoietic suppression in fetal anemia because of Kell alloimmunization. *Am J Obstet Gynecol* **171**, 247-52 (1994).

121. Vaughan, J.I. *et al.* Inhibition of erythroid progenitor cells by anti-Kell antibodies in fetal alloimmune anemia. *N Engl J Med* **338**, 798-803 (1998).

122. Lakhwani, S. *et al.* Kell hemolytic disease of the fetus. Combination treatment with plasmapheresis and intrauterine blood transfusion. *Transfus Apher Sci* **45**, 9-11 (2011).

123. Helgeson, M., Swanson, J. & Polesky, H.F. Knops-Helgeson (Kn^a), a high-frequency erythrocyte antigen. *Transfusion* **10**, 137-8 (1970).

124. Moulds, J.M. *et al.* Expansion of the Knops blood group system and subdivision of Sl(a). *Transfusion* **42**, 251-6 (2002).
125. Moulds, J.M. The Knops blood-group system: a review. *Immunohematology* **26**, 2-7 (2010).
126. Holers, V.M. *et al.* Human complement C3b/C4b receptor (CR1) mRNA polymorphism that correlates with the *CR1* allelic molecular weight polymorphism. *Proc Natl Acad Sci U S A* **84**, 2459-63 (1987).
127. Wilson, J.G. *et al.* Identification of a restriction fragment length polymorphism by a CR1 cDNA that correlates with the number of CR1 on erythrocytes. *J Exp Med* **164**, 50-9 (1986).
128. Moulds, J.M., Moulds, J.J., Brown, M. & Atkinson, J.P. Antiglobulin testing for CR1-related (Knops/McCoy/Swain-Langley/York) blood group antigens: negative and weak reactions are caused by variable expression of CR1. *Vox Sang* **62**, 230-5 (1992).
129. Wilson, J.G., Wong, W.W., Schur, P.H. & Fearon, D.T. Mode of inheritance of decreased C3b receptors on erythrocytes of patients with systemic lupus erythematosus. *N Engl J Med* **307**, 981-6 (1982).
130. Liu, D.H., Yao, Y.T., Li, L.H. & Huang, C.M. Effects of Ulinastatin on In Vitro Storage Lesions of Human Red Blood Cells. *Clin Lab* **63**, 833-838 (2017).
131. Ripoche, J. & Sim, R.B. Loss of complement receptor type 1 (CR1) on ageing of erythrocytes. Studies of proteolytic release of the receptor. *Biochem J* **235**, 815-21 (1986).
132. Vik, D.P. & Wong, W.W. Structure of the gene for the F allele of complement receptor type 1 and sequence of the coding region unique to the S allele. *J Immunol* **151**, 6214-24 (1993).
133. Herrera, A.H., Xiang, L., Martin, S.G., Lewis, J. & Wilson, J.G. Analysis of complement receptor type 1 (CR1) expression on erythrocytes and of CR1 allelic markers in Caucasian and African American populations. *Clin Immunol Immunopathol* **87**, 176-83 (1998).
134. Rowe, J.A. *et al.* Erythrocyte CR1 expression level does not correlate with a *Hin*dIII restriction fragment length polymorphism in Africans; implications for studies on malaria susceptibility. *Genes Immun* **3**, 497-500 (2002).
135. Zorzetto, M. *et al.* Complement receptor 1 gene polymorphisms in sarcoidosis. *Am J Respir Cell Mol Biol* **27**, 17-23 (2002).
136. Xiang, L., Rundles, J.R., Hamilton, D.R. & Wilson, J.G. Quantitative alleles of *CR1*: coding sequence analysis and comparison of haplotypes in two ethnic groups. *J Immunol* **163**, 4939-45 (1999).
137. Cooling, L. Blood Groups in Infection and Host Susceptibility. *Clin Microbiol Rev* **28**, 801-70 (2015).
138. Prajapati, S.K. *et al.* Complement Receptor 1 availability on red blood cell surface modulates Plasmodium vivax invasion of human reticulocytes. *Sci Rep* **9**, 8943 (2019).
139. Grueger, D. *et al.* Two novel antithetical KN blood group antigens may contribute to more than a quarter of all KN antisera in Europe. *Transfusion* **60**, 2408-2418 (2020).
140. Reid, M.E., Lomas-Francis, C. & Olsson, M.L. *The blood group antigen factsbook*, xii, 745 pages (Elsevier/AP, Amsterdam, 2012).

141. Seltsam, A. *et al.* Recombinant blood group proteins facilitate the detection of alloantibodies to high-prevalence antigens and reveal underlying antibodies: results of an international study. *Transfusion* **54**, 1823-30 (2014).
142. Lambert, J.C. *et al.* Genome-wide association study identifies variants at *CLU* and *CR1* associated with Alzheimer's disease. *Nat Genet* **41**, 1094-9 (2009).
143. Tham, W.H. *et al.* Complement receptor 1 is the host erythrocyte receptor for Plasmodium falciparum PfRh4 invasion ligand. *Proc Natl Acad Sci U S A* **107**, 17327-32 (2010).
144. Rowe, J.A. *et al.* Mapping of the region of complement receptor (CR) 1 required for *Plasmodium falciparum* rosetting and demonstration of the importance of CR1 in rosetting in field isolates. *J Immunol* **165**, 6341-6 (2000).
145. Rowe, J.A., Moulds, J.M., Newbold, C.I. & Miller, L.H. *P. falciparum* rosetting mediated by a parasite-variant erythrocyte membrane protein and complement-receptor 1. *Nature* **388**, 292-5 (1997).
146. Miller, L.H., Baruch, D.I., Marsh, K. & Doumbo, O.K. The pathogenic basis of malaria. *Nature* **415**, 673-9 (2002).
147. Cockburn, I.A. *et al.* A human complement receptor 1 polymorphism that reduces *Plasmodium falciparum* rosetting confers protection against severe malaria. *Proc Natl Acad Sci U S A* **101**, 272-7 (2004).
148. Bessis, M., Mize, C. & Prenant, M. Erythropoiesis: comparison of in vivo and in vitro amplification. *Blood Cells* **4**, 155-74 (1978).
149. Manwani, D. & Bieker, J.J. The erythroblastic island. *Curr Top Dev Biol* **82**, 23-53 (2008).
150. Palis, J. Primitive and definitive erythropoiesis in mammals. *Front Physiol* **5**, 3 (2014).
151. Sender, R., Fuchs, S. & Milo, R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol* **14**, e1002533 (2016).
152. Hattangadi, S.M., Wong, P., Zhang, L., Flygare, J. & Lodish, H.F. From stem cell to red cell: regulation of erythropoiesis at multiple levels by multiple proteins, RNAs, and chromatin modifications. *Blood* **118**, 6258-68 (2011).
153. Zhang, Y. *et al.* Erythropoietin action in stress response, tissue maintenance and metabolism. *Int J Mol Sci* **15**, 10296-333 (2014).
154. Kim, S.I. & Bresnick, E.H. Transcriptional control of erythropoiesis: emerging mechanisms and principles. *Oncogene* **26**, 6777-94 (2007).
155. Wells, M. & Steiner, L. Epigenetic and Transcriptional Control of Erythropoiesis. *Front Genet* **13**, 805265 (2022).
156. Grass, J.A. *et al.* GATA-1-dependent transcriptional repression of GATA-2 via disruption of positive autoregulation and domain-wide chromatin remodeling. *Proc Natl Acad Sci U S A* **100**, 8811-6 (2003).
157. Mei Zhan, C.-Z.S. MicroRNA and Erythroid Differentiation. in ***Current Perspectives in microRNAs (miRNA)*** (ed. Ying, S.-Y.) 97-117 (Springer, 2008).
158. Xu, C. & Shi, L. Long non-coding RNAs during normal erythropoiesis. *Blood Sci* **1**, 137-140 (2019).
159. Listowski, M.A. *et al.* microRNAs: fine tuning of erythropoiesis. *Cell Mol Biol Lett* **18**, 34-46 (2013).

160. Liu, G., Kim, J., Nguyen, N.H., Zhou, L. & Dean, A. Long noncoding RNA GATA2AS influences human erythropoiesis by transcription factor and chromatin landscape regulation. *Blood* (2024).

161. Evans, T., Reitman, M. & Felsenfeld, G. An erythrocyte-specific DNA-binding factor recognizes a regulatory sequence common to all chicken globin genes. *Proc Natl Acad Sci U S A* **85**, 5976-80 (1988).

162. Vyas, P., Ault, K., Jackson, C.W., Orkin, S.H. & Shivdasani, R.A. Consequences of GATA-1 deficiency in megakaryocytes and platelets. *Blood* **93**, 2867-75 (1999).

163. Pal, S. *et al.* Coregulator-dependent facilitation of chromatin occupancy by GATA-1. *Proc Natl Acad Sci U S A* **101**, 980-5 (2004).

164. Hitzler, J.K., Cheung, J., Li, Y., Scherer, S.W. & Zipursky, A. GATA1 mutations in transient leukemia and acute megakaryoblastic leukemia of Down syndrome. *Blood* **101**, 4301-4 (2003).

165. Singleton, B.K. *et al.* A novel GATA1 mutation (Stop414Arg) in a family with the rare X-linked blood group Lu(a-b-) phenotype and mild macrothrombocytic thrombocytopenia. *Br J Haematol* **161**, 139-42 (2013).

166. Shyu, Y.C. *et al.* Tight regulation of a timed nuclear import wave of EKLF by PKCtheta and FOE during Pro-E to Baso-E transition. *Dev Cell* **28**, 409-22 (2014).

167. Borg, J., Patrinos, G.P., Felice, A.E. & Philipsen, S. Erythroid phenotypes associated with KLF1 mutations. *Haematologica* **96**, 635-8 (2011).

168. Borg, J. *et al.* Haploinsufficiency for the erythroid transcription factor KLF1 causes hereditary persistence of fetal hemoglobin. *Nat Genet* **42**, 801-5 (2010).

169. Xu, L. *et al.* Compound Heterozygosity for KLF1 Mutations Causing Hemolytic Anemia in Children: A Case Report and Literature Review. *Front Genet* **12**, 691461 (2021).

170. Viprakasit, V. *et al.* Mutations in Kruppel-like factor 1 cause transfusion-dependent hemolytic anemia and persistence of embryonic globin gene expression. *Blood* **123**, 1586-95 (2014).

171. Arnaud, L. *et al.* A dominant mutation in the gene encoding the erythroid transcription factor KLF1 causes a congenital dyserythropoietic anemia. *Am J Hum Genet* **87**, 721-7 (2010).

172. Magor, G.W. *et al.* KLF1-null neonates display hydrops fetalis and a deranged erythroid transcriptome. *Blood* **125**, 2405-17 (2015).

173. Singleton, B.K., Burton, N.M., Green, C., Brady, R.L. & Anstee, D.J. Mutations in EKLF/KLF1 form the molecular basis of the rare blood group In(Lu) phenotype. *Blood* **112**, 2081-8 (2008).

174. Helias, V. *et al.* Molecular analysis of the rare in(Lu) blood type: toward decoding the phenotypic outcome of haploinsufficiency for the transcription factor KLF1. *Hum Mutat* **34**, 221-8 (2013).

175. Westman, J.S. *et al.* Allele-selective RUNX1 binding regulates P1 blood group status by transcriptional control of *A4GALT*. *Blood* **131**, 1611-1616 (2018).

176. Sood, R., Kamikubo, Y. & Liu, P. Role of RUNX1 in hematological malignancies. *Blood* **129**, 2070-2082 (2017).

177. De Braekeleer, E. *et al.* RUNX1 translocations and fusion genes in malignant hemopathies. *Future Oncol* **7**, 77-91 (2011).

178. Grossmann, V. *et al.* Prognostic relevance of RUNX1 mutations in T-cell acute lymphoblastic leukemia. *Haematologica* **96**, 1874-7 (2011).

179. Ying, Y. *et al.* A novel mutation +5904 C>T of RUNX1 site in the erythroid cell-specific regulatory element decreases the ABO antigen expression in Chinese population. *Vox Sang* (2018).

180. Takahashi, Y. *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* **106**, 167-75 (2014).

181. Vincent P Schulz, K.L.-G., Peiying Shan, Julien Papoin, Mohandas Narla, Laurie A. Steiner, Lionel Blanc, James Palis, Patrick G Gallagher. Identification of a Novel Gene Regulatory Element in Human Erythroid Progenitor Cells. *Blood* **142**(2023).

182. Chuong, E.B., Elde, N.C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* **18**, 71-86 (2017).

183. Lecuyer, E. & Hoang, T. SCL: from the origin of hematopoiesis to stem cells and leukemia. *Exp Hematol* **32**, 11-24 (2004).

184. Tripic, T. *et al.* SCL and associated proteins distinguish active from repressive GATA transcription factor complexes. *Blood* **113**, 2191-201 (2009).

185. Zhang, J., Kalkum, M., Yamamura, S., Chait, B.T. & Roeder, R.G. E protein silencing by the leukemogenic AML1-ETO fusion protein. *Science* **305**, 1286-9 (2004).

186. Deng, W. *et al.* Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**, 1233-44 (2012).

187. Andrews, N.C. *et al.* The ubiquitous subunit of erythroid transcription factor NF-E2 is a small basic-leucine zipper protein related to the v-maf oncogene. *Proc Natl Acad Sci U S A* **90**, 11488-92 (1993).

188. Sawado, T., Igarashi, K. & Groudine, M. Activation of beta-major globin gene transcription is associated with recruitment of NF-E2 to the beta-globin LCR and gene promoter. *Proc Natl Acad Sci U S A* **98**, 10226-31 (2001).

189. Lambert, S.A. *et al.* The Human Transcription Factors. *Cell* **172**, 650-665 (2018).

190. ISBT. Names for (ISBT_102) GATA1 Alleles. (2021).

191. ISBT. Names for Transcription Factor KLF1 Alleles. (2021).

192. Tournamille, C., Colin, Y., Cartron, J.P. & Le Van Kim, C. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nat Genet* **10**, 224-8 (1995).

193. Livingstone, F.B. The Duffy blood groups, vivax malaria, and malaria selection in human populations: a review. *Hum Biol* **56**, 413-25 (1984).

194. Duchene, J. *et al.* Atypical chemokine receptor 1 on nucleated erythroid cells regulates hematopoiesis. *Nat Immunol* **18**, 753-761 (2017).

195. Pogo, A.O. & Chaudhuri, A. The Duffy protein: a malarial and chemokine receptor. *Semin Hematol* **37**, 122-9 (2000).

196. Nakajima, T. *et al.* Mutation of the GATA site in the erythroid cell-specific regulatory element of the *ABO* gene in a B_m subgroup individual. *Transfusion* **53**, 2917-27 (2013).

197. Sano, R. *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* **119**, 5301-10 (2012).

198. Takahashi, Y. *et al.* Presence of nucleotide substitutions in transcriptional regulatory elements such as the erythroid cell-specific enhancer-like element and the ABO promoter in individuals with phenotypes A3 and B3, respectively. *Vox Sang* **107**, 171-80 (2014).

199. Isa, K. *et al.* Presence of nucleotide substitutions in the *ABO* promoter in individuals with phenotypes A_3 and B_3. *Vox Sang* **110**, 285-7 (2016).

200. Hellberg, A. *et al.* A novel single-nucleotide substitution in the proximal *ABO* promoter gives rise to the B_3 phenotype. *Transfusion* **59**, E1-E3 (2019).

201. Fennell, K. *et al.* Effect on gene expression of three allelic variants in GATA motifs of ABO, RHD, and RHCE regulatory elements. *Transfusion* **57**, 2804-2808 (2017).

202. Möller, M. *et al.* Disruption of a GATA1-binding motif upstream of *XG*/*PBDX* abolishes Xg^a expression and resolves the Xg blood group system. *Blood* **132**, 334-338 (2018).

203. Yeh, C.C. *et al.* The molecular genetic background leading to the formation of the human erythroid-specific Xg^a/CD99 blood groups. *Blood Adv* **2**, 1854-1864 (2018).

204. Ulirsch, J.C. *et al.* Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* **165**, 1530-1545 (2016).

205. Christophersen, M.K. *et al.* *SMIM1* variants rs1175550 and rs143702418 independently modulate Vel blood group antigen expression. *Sci Rep* **7**, 40451 (2017).

206. Yeh, C.C. *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* **58**, 1054-1064 (2018).

207. Moore, L.D., Le, T. & Fan, G. DNA methylation and its basic function. *Neuropsychopharmacology* **38**, 23-38 (2013).

208. Saxonov, S., Berg, P. & Brutlag, D.L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* **103**, 1412-7 (2006).

209. Jin, B., Li, Y. & Robertson, K.D. DNA methylation: superior or subordinate in the epigenetic hierarchy? *Genes Cancer* **2**, 607-17 (2011).

210. Van Loghem, J.J., Jr., Dorfmeier, H. & Van Der Hart, M. Two A antigens with abnormal serologic properties. *Vox Sang* **2**, 16-24 (1957).

211. Bianco, T., Farmer, B.J., Sage, R.E. & Dobrovic, A. Loss of red cell A, B, and H antigens is frequent in myeloid malignancies. *Blood* **97**, 3633-9 (2001).

212. Daniels, G. *Human blood groups*, xxx, 737 p. (Blackwell Science, Oxford ; Cambridge, Mass., USA, 1995).

213. Kominato, Y. *et al.* Expression of human histo-blood group ABO genes is dependent upon DNA methylation of the promoter region. *J Biol Chem* **274**, 37240-50 (1999).

214. Kominato, Y. *et al.* Alternative promoter identified between a hypermethylated upstream region of repetitive elements and a CpG island in human ABO histo-blood group genes. *J Biol Chem* **277**, 37936-48 (2002).

215. Chihara, Y. *et al.* Loss of blood group A antigen expression in bladder cancer caused by allelic loss and/or methylation of the *ABO* gene. *Lab Invest* **85**, 895-907 (2005).

216. Gao, S. *et al.* Genetic and epigenetic alterations of the blood groupABO gene in oral squamous cell carcinoma. *International Journal of Cancer* **109**, 230-237 (2004).

217. Bianco-Miotto, T., Hussey, D.J., Day, T.K., O'Keefe, D.S. & Dobrovic, A. DNA methylation of the *ABO* promoter underlies loss of *ABO* allelic expression in a significant proportion of leukemic patients. *PLoS One* **4**, e4788 (2009).

218. Miola, M.P., de Oliveira, T.C., Guimaraes, A.A.G., Ricci-Junior, O. & de Mattos, L.C. ABO discrepancy resolution in two patients with acute myeloid leukemia presenting the transient weak expression of A antigen. *Hematol Transfus Cell Ther* (2022).

219. Jeong, I.H., Seo, J.Y., Choi, S., Kim, H.Y. & Cho, D. ABO Blood Group Antigen Changes in Acute Myeloid Leukemia and No Significant Association With RUNX1 and GATA2 Somatic Variants. *Ann Lab Med* **43**, 635-637 (2023).

220. Bannister, A.J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res* **21**, 381-95 (2011).

221. Kimura, H. Histone modifications for human epigenome analysis. *J Hum Genet* **58**, 439-45 (2013).

222. Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol Cell* **49**, 825-37 (2013).

223. Heintzman, N.D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311-8 (2007).

224. Creyghton, M.P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-6 (2010).

225. Tyagi, M., Imam, N., Verma, K. & Patel, A.K. Chromatin remodelers: We are the drivers!! *Nucleus* **7**, 388-404 (2016).

226. Hewish, D.R. & Burgoyne, L.A. Chromatin sub-structure. The digestion of chromatin DNA at regularly spaced sites by a nuclear deoxyribonuclease. *Biochem Biophys Res Commun* **52**, 504-10 (1973).

227. Boyle, A.P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311-22 (2008).

228. Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R. & Lieb, J.D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* **17**, 877-85 (2007).

229. Cui, K. & Zhao, K. Genome-wide approaches to determining nucleosome occupancy in metazoans using MNase-Seq. *Methods Mol Biol* **833**, 413-9 (2012).

230. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-8 (2013).

231. Kelly, T.K. *et al.* Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res* **22**, 2497-506 (2012).

232. Stergachis, A.B., Debo, B.M., Haugen, E., Churchman, L.S. & Stamatoyannopoulos, J.A. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science* **368**, 1449-1454 (2020).

233. Mansisidor, A.R. & Risca, V.I. Chromatin accessibility: methods, mechanisms, and biological insights. *Nucleus* **13**, 236-276 (2022).

234. Grosveld, F., van Assendelft, G.B., Greaves, D.R. & Kollias, G. Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell* **51**, 975-85 (1987).

235. Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F. & de Laat, W. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell* **10**, 1453-65 (2002).

236. Han, J., Zhang, Z. & Wang, K. 3C and 3C-based techniques: the powerful tools for spatial genome organization deciphering. *Mol Cytogenet* **11**, 21 (2018).

237. Naumann, S., Reutzel, D., Speicher, M. & Decker, H.J. Complete karyotype characterization of the K562 cell line by combined application of G-banding, multiplex-fluorescence in situ hybridization, fluorescence in situ hybridization, and comparative genomic hybridization. *Leuk Res* **25**, 313-22 (2001).

238. Lee, S., Zambas, E., Green, E.D. & Redman, C. Organization of the gene encoding the human Kell blood group protein. *Blood* **85**, 1364-70 (1995).

239. Camara-Clayette, V. *et al.* Transcriptional regulation of the KEL gene and Kell protein expression in erythroid and non-erythroid cells. *Biochem J* **356**, 171-80 (2001).

240. Ceppellini, R., Dunn, L.C. & Turri, M. An Interaction between Alleles at the Rh Locus in Man Which Weakens the Reactivity of the Rh(0) Factor (D). *Proc Natl Acad Sci U S A* **41**, 283-8 (1955).

241. Hay, D. *et al.* Genetic dissection of the alpha-globin super-enhancer in vivo. *Nat Genet* **48**, 895-903 (2016).

242. Liu, W. *et al.* Functional Evaluation of KEL as an Oncogenic Gene in the Progression of Acute Erythroleukemia. *Oxid Med Cell Longev* **2022**, 5885342 (2022).

243. Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-80 (2014).

244. E Meyer, Y.M., C Gassner, Y Song, S Meyer, C Engström, B Frey. EXPRESSION OF RHD IS LINKED TO RHD/RHCE GENOTYPE. *Vox Sang* **114 Suppl 1**, 5-240 (2019).

245. Kaya-Okur, H.S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun* **10**, 1930 (2019).

246. Nakato, R. & Sakata, T. Methods for ChIP-seq analysis: A practical workflow and advanced applications. *Methods* **187**, 44-53 (2021).

247. Keller, J. *et al.* Novel mutations in KLF1 encoding the In(Lu) phenotype reflect a diversity of clinical presentations. *Transfusion* **58**, 196-199 (2018).

248. Baysoy, A., Bai, Z., Satija, R. & Fan, R. The technological landscape and applications of single-cell multi-omics. *Nat Rev Mol Cell Biol* **24**, 695-713 (2023).

249. Macaulay, I.C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods* **12**, 519-22 (2015).

250. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103-1116 e20 (2020).

251. Li, Z. *et al.* Applications of deep learning in understanding gene regulation. *Cell Rep Methods* **3**, 100384 (2023).

252. Alipanahi, B., Delong, A., Weirauch, M.T. & Frey, B.J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* **33**, 831-8 (2015).

253. Zhou, J. & Troyanskaya, O.G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* **12**, 931-4 (2015).

254. Zhou, J. *et al.* Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet* **50**, 1171-1179 (2018).

255. Avsec, Z. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* **18**, 1196-1203 (2021).

256. Fudenberg, G., Kelley, D.R. & Pollard, K.S. Predicting 3D genome folding from DNA sequence with Akita. *Nat Methods* **17**, 1111-1117 (2020).

257. Karbalayghareh, A., Sahin, M. & Leslie, C.S. Chromatin interaction-aware gene regulatory modeling with graph attention networks. *Genome Res* **32**, 930-944 (2022).

258. Zhou, J. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nat Genet* **54**, 725-734 (2022).

Printed by Media-Tryck, Lund 2024 NORDIC SWAN ECOLABEL 3041 0903

About the author

Trained as a medical technologist, and after serving at the Taipei Blood Center in Taiwan, I embarked on a new chapter in a distant land driven by my passion for blood groups and an aim to improve healthcare. Even though gene expression is meticulously governed, my quest pursuing it has not been without twists and turns. In this thesis work, I invite you to join me on a journey exploring how blood group genes are regulated, how we all are different and unique, and how this study can lead to safer transfusions.

LUND UNIVERSITY

FACULTY OF MEDICINE

Department of Laboratory Medicine

Lund University, Faculty of Medicine
Doctoral Dissertation Series 2024:49
ISBN 978-91-8021-542-8
ISSN 1652-8220