Identification and characterisation of SMIM1 variants determining the Vel blood group

Christophersen, Mikael Kronborg

2017

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

Link to publication

Citation for published version (APA):
Christophersen, M. K. (2017). *Identification and characterisation of SMIM1 variants determining the Vel blood group*. [Doctoral Thesis (compilation), Division of Hematology and Transfusion Medicine]. Lund University: Faculty of Medicine.

Total number of authors:
1

Identification and characterisation of *SMIM1* variants determining the Vel blood group

Mikael Kronborg Christophersen

DOCTORAL DISSERTATION
by due permission of the Faculty of Medicine, Lund University, Sweden.
To be defended at 13:00 on Thursday 30th March 2017
at Biomedicinsk Centrum, Segerfalksalen

Faculty opponent
Associate Professor Morten Hanefeld Dziegiel

Organization
LUND UNIVERSITY
Department of Laboratory Medicine
Division of Hematology and Transfusion Medicine
Author(s) Mikael Kronborg Christophersen

Document name
DOCTORAL DISSERTATION

Date of issue
March 30th 2017

Sponsoring organization

Title and subtitle Identification and characterisation of SMIM1 variants determining the Vel blood group

Abstract
The Vel blood group antigen is present on red blood cells from all humans except rare Vel-negative individuals, who can form antibodies to Vel in response to transfusion or pregnancy. It was first described in 1952 as a high incidence antigen, while the molecular background was recently discovered to be a 17-bp deletion in Small Integral Membrane Protein 1, that causes a frame-shift mutation and abolishes *SMIM1* expression, thus creating a Vel-negative phenotype.

This thesis contains one of the three original discovery studies reporting the Vel antigen-defining *SMIM1*-deletion. We analysed SNP microarray data from Vel-positive and -negative individuals and identified *SMIM1* as a potential gene. We then found the 17-bp deletion to only be present in Vel-negative individuals and sought to finally prove *SMIM1* as the genetic background for the Vel antigen by examining: 1) *SMIM1* mRNA sequence and expression levels, 2) presence of SMIM1 and the Vel antigen in red blood cell membranes from Vel-positive and -negative individuals and 3) anti-Vel antibody reactivity in erythroleukaemia cells expressing wild type and mutant SMIM1 protein. Our discovery allowed Vel to be officially recognised by the International Society of Blood Transfusion as blood group system 034.
The *SMIM1* deletion is the major determinant for Vel expression, yet even Vel-positive individuals (*i.e.* people carrying wild type *SMIM1*) show substantial variation in reactivity with anti-Vel antibody, creating a risk for Vel blood typing errors and transfusion reactions. We suspected the cause to be sequence variants in *SMIM1* and found rs143702418 (insertion, C>CGCA) and rs1175550 (A>G) to independently influence expression of *SMIM1*, potentially mediated by the erythroid transcription factor TAL1 that binds preferentially to the high-expressing rs1175550G allele.
Lastly, we examined historical Vel-negative samples and sought to retroactively reconcile historic Vel designations to our current knowledge of the *SMIM1* deletion variant. As such, we found the old Vel;–1,–2 designation, attributed to individuals phenotyped to have weak to no Vel antigen expression, to correspond to homozygosity for the *SMIM1* deletion, while persons designated Vel;1,–2, low Vel antigen expression, matched with being heterozygous carrier of the deletion.

The Vel antigen was one of the few remaining clinically significant antigens with unknown genetic background, that causes severe haemolytic transfusion reactions. This thesis assisted in characterising the molecular background of the Vel antigen, which paved the way for molecular Vel typing in the clinic, as well as the prospect of manufacturing a synthetic monoclonal anti-Vel antibody to be used in diagnostics, blood group typing and research.

Key words Blood groups, Vel blood group system, genetic variation, transcriptional regulation

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language English

ISSN and key title 1652-8220

ISBN 978-91-7619-428-7

Recipient's notes

Number of pages 94

Price

Security classification

Signature _____ **Date** February 22nd 2017

Identification and characterisation of *SMIM1* variants determining the Vel blood group

Mikael Kronborg Christophersen

LUND
UNIVERSITY

Faculty of Medicine
Department of Laboratory Medicine
Division of Hematology and Transfusion Medicine

Cover photos by Camilla Nehammer

ISSN 1652-8220
ISBN 978-91-7619-428-7
Lund University, Faculty of Medicine Doctoral Dissertation Series 2017:48

Printed in Sweden by Media-Tryck, Lund University
Lund 2017

Yet an experiment, were you to try it,
could free you from your cavil,
and the source of your arts' course
springs from experiment.

-Dante Alighieri: The Divine Comedy

Preface

The following PhD thesis details the results obtained through four years of research at the Division of Haematology and Transfusion Medicine, Lund University. It consists of three studies focusing on the Vel blood group system and one studying the disease multiple myeloma. Three studies have been published in *Nature Genetics*, *Scientific Reports* and *Blood* and one is an almost submission-ready manuscript.

The Vel blood group antigen is present on red blood cells from all humans except rare Vel-negative individuals who can form antibodies to Vel in response to transfusion or pregnancy. It was first described in 1952 as a high incidence antigen, while the molecular background was recently discovered to be a 17-bp deletion in Small Integral Membrane Protein 1, that causes a frame-shift mutation and abolishes *SMIM1* expression, thus creating a Vel-negative phenotype.

This thesis contains one of the three original discovery studies reporting the Vel antigen-defining *SMIM1*-deletion. We analysed SNP microarray data from Vel-positive and -negative individuals and identified *SMIM1* as a potential gene. We then found the 17-bp deletion to only be present in Vel-negative individuals and sought to finally prove *SMIM1* as the genetic background for the Vel antigen by examining: 1) *SMIM1* mRNA sequence and expression levels, 2) presence of SMIM1 and the Vel antigen in red blood cell membranes from Vel-positive and -negative individuals and 3) anti-Vel antibody reactivity in red blood cells from individuals homozygous or heterozygous for the 17-bp deletion and in erythroleukaemia cells expressing wild type and mutant SMIM1 protein. Our discovery allowed Vel to be officially recognised by the International Society of Blood Transfusion as blood group system 034.

The *SMIM1* deletion is the major determinant for Vel expression, yet even Vel-positive individuals show substantial variation in reactivity with anti-Vel antibody, creating a risk for Vel blood typing errors and transfusion reactions. We suspected the cause to be sequence variants in *SMIM1* and found rs143702418 (insertion, C>CGCA) and rs1175550 (A>G) to independently influence expression of *SMIM1*, potentially mediated by the erythroid transcription factor TAL1 that binds preferentially to the high-expressing rs1175550G allele.

Lastly, we examined historical Vel-negative samples and sought to retroactively reconcile historic Vel designations to our current knowledge of the *SMIM1* deletion variant. As such, we found the old Vel;–1,–2 designation, attributed to individuals phenotyped to have weak to no Vel antigen expression, to correspond to homozygosity for the *SMIM1* deletion, while persons designated Vel;1,–2, *i.e.* low Vel antigen expression, matched with being heterozygous carrier of the deletion.

This thesis assisted in characterising the molecular background of the Vel antigen, which paved the way for molecular Vel typing in the clinic as well as the prospect of manufacturing a synthetic monoclonal anti-Vel antibody to be used in diagnostics, blood group typing and research.

The final Study IV details the discovery of robust biomarkers for improved FACS sorting of multiple myeloma plasma cells in a bone marrow sample, a project I was involved in during the first year of my PhD.

The thesis is divided into four main sections: An introduction outlining haematopoiesis, erythroid transcription factors, blood group systems and genetic variation; the specific aims and a brief summary of the obtained results, and a discussion and conclusion section that puts the studies into a broader scientific context. The last section is the four articles included at the end of the thesis.

Content

Studies included in this thesis

I. **Homozygosity for a null allele of SMIM1 defines the Vel-negative blood group phenotype**

Jill R. Storry*, Magnus Jöud*, **Mikael Kronborg Christophersen**, Britt Thuresson, Bo Åkerström, Birgitta Nilsson Sojka, Björn Nilsson*, Martin L. Olsson*.

Nature Genetics, 2013. 45(5): 537-541.

II. **SMIM1 variants rs1175550 and rs143702418 independently modulate Vel blood group antigen expression**

Mikael Kronborg Christophersen, Magnus Jöud, Ram Ajore, Sunitha Vege, Klara W. Ljungdahl, Connie M. Westhoff, Martin L. Olsson, Jill R. Storry*, Björn Nilsson*.

Scientific Reports, 2017. 7: 40451.

III. **Serologic And Molecular Studies Of The Vel– Phenotype In A Multi-ethnic Population**

Mikael Kronborg Christophersen, Jill R. Storry, Kim Hue-Roye, Randall W. Velliquette, Sunitha Vege, Martin L. Olsson, Björn Nilsson, Christine Lomas-Francis, Connie M. Westhoff.

Manuscript in preparation

IV. **Robust isolation of malignant plasma cells in multiple myeloma**

Ildikó Frigyesi, Jörgen Adolfsson, Mina Ali, **Mikael Kronborg Christophersen**, Ellinor Johnsson, Ingemar Turesson, Urban Gullberg, Markus Hansson, Björn Nilsson.

Blood, 2014. 123(9):1336-1340.

* denotes shared authorship.

Abbreviations

ACKR1	Atypical chemokine receptor 1
BFU-E	burst forming unit-erythroid
C/EBPα	CCAAT-enhancer binding protein α
CFU-E	colony forming unit-erythroid
CLP	common lymphoid progenitor
CMP	common myeloid progenitor
CSF	colony stimulating factor
eTF	erythroid transcription factor
Epo	Erythropoietin
EpoR	Epo receptor
eQTL	expression quantitative trait loci
FACS	fluorescence-activated cell sorting
FOG-1	Friend of GATA-1
Gfi-1[B]	Growth factor independent 1[B]
GMP	granulocyte/macrophage progenitor
GWAS	genome-wide association study
GPA/B	glycophorin A/B
GYPA/GYPB	glycophorin A/B genes
HBA	α-globin gene
HBB	β-globin gene
HBG1-2	γ-globin genes 1 and 2
HDAC	histone deacetylase
HSC	haematopoietic stem cell
HTR	haemolytic transfusion reaction
IL-3	interleukin-3
indel	insertion/deletion

ISBT	International Society of Blood Transfusion
KLF1	Krüppel-like factor 1
LD	linkage disequilibrium
LDB1	LIM domain binding 1
LMO2	LIM-only Domain 2
LSD1	Lysine-specific histone demethylase 1A
MCHC	mean corpuscular haemoglobin content
MeCP1	Methyl-CpG-binding protein-1
MEP	megakaryocyte-erythrocyte progenitor
MM	multiple myeloma
MPP	multipotent progenitor
NuRD	Nucleosome Remodelling Deacetylase
P-TEFb	Positive transcription elongation factor b
Pax5	Paired box 5
Pentameric	Pentameric eTF complex of GATA-1:TAL1:E2A:LMO2:LDB1
(5'/3') RACE	(5' or 3') Rapid amplification of cDNA ends
RBC	Red blood cell
RUNX1	Runt-related transcription factor 1
SCF	stem cell factor
SMIM1	small integral membrane protein 1
SNP	Single nucleotide polymorphism
SLC4A1	Solute carrier family 4 member 1
Sp1	Specificity protein 1
STAT5	signal transducer and activator of transcription 5
TAL1	T-cell acute lymphocytic leukemia protein 1
TFRC	transferrin receptor
TMP	transmembrane protein
ZBTB7A	Zinc Finger And BTB Domain Containing 7A

Introduction

Blood is the life-giving fluid that flows through us all. Coursing around inside an intricate network of blood vessels, it supplies our body with vital nutrients and oxygen. It is often the first line of detection and defence against foreign objects and microorganisms, able to act swiftly and efficiently clear invaders before they make it past the circulatory system.

The formation of blood is a delicate balance of self-renewal and lineage choices. From the early haematopoietic stem cells to the strikingly different end products, from the enucleated doughnut-shaped erythrocytes carrying vital oxygen over the giant megakaryocytes shedding pieces of itself as part of its normal function, to the granulocytes with nuclei of creative shapes and forms, the haematopoietic system is a truly remarkable tissue.

But blood can also, inadvertently, become dangerous. With the delicate regulatory system comes an inherent risk of errors that can prove detrimental, causing loss of control and involuntary abandonment of normal function.

And in case the errors prove non-fatal, they will still make the blood unique. A seemingly harmless mutation may change the blood cells just enough to set them apart from the rest of the population. Meaning that if ever the need for administration of foreign blood arises, if it is not carefully selected, the first line of defence could become the enemy as, in its eagerness to defend, it will end up destroying not only the intruders, but also itself in the process.

Blood is life-giving. It is complex and it is crucial. And it can become treacherous and fragile. It makes us unique and sets us apart. As such, extensive knowledge of blood, the factors that govern its creation and differentiation, and the systems that control donor-recipient compatibility, is an imperative necessity.

Haematopoiesis

The haematopoietic system is a paradigmatic, stem cell-maintained organ with enormous cell turnover. Up to a staggering one trillion blood cells are produced every day[1]. The haematopoietic system has been extensively studied, documenting the process of going from haematopoietic stem cells (HSCs) through several checkpoints to a mature, differentiated blood cell.

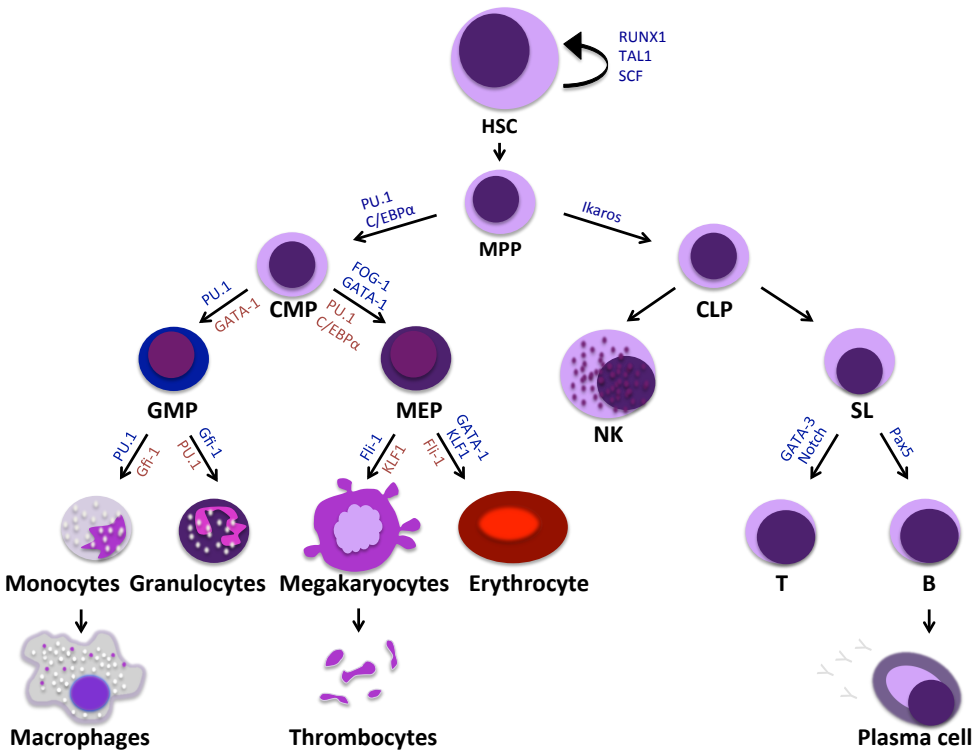


Figure 1 Haematopoiesis. An overview of haematopoiesis with the progenitor and precursor cells and the factors that guide them as described in the main text. For simplicity, not all cell intermediates nor factors are included. Factors that drive lineage differentiation are in blue font, those that inhibit are in red. Abbreviations: HSC = haematopoietic stem cell; MPP = multipotent progenitor; CMP = common myeloid progenitor; GMP = granulocyte-macrophage progenitor; MEP = megakaryocyte-erythrocyte progenitor; CLP = common lymphoid progenitor; NK = natural killer; SL = small lymphocyte; T = T cell; B = B cell. Factor names are described in the main text.

Definitive haematopoiesis in the bone marrow

Every blood cell is derived from haematopoietic stem cells (HSCs) residing in the bone marrow. An intricate system of regulators, enhancers and suppressors guide the differentiating HSCs through the many cell fate decisions, checkpoints and branch points in the haematopoietic lineage tree in order to differentiate into a mature blood cell. The main endpoint cells in the blood are the platelet-producing megakaryocytes, the oxygen-carrying erythrocytes, the microorganism-fighting granulocytes (neutrophilic, basophilic and eosinophilic) and the immune system's macrophages, T and B cells and plasma cells derived from the latter (**Figure 1**).

HSCs reside in the bone marrow, where they proliferate and are capable of self-renewal, maintained by T-cell acute lymphoblastic leukaemia protein (TAL1, also known as Stem cell leukaemia factor (SCL)), Runt-related transcription factor 1 (RUNX1) and stem cell factor (SCF) and others[2,3]. From here a HSC differentiates into a multipotent progenitor (MPP) cell and starts down the haematopoietic lineage tree[4]. The first major branch point is the lymphoid or myeloid lineage, with the former giving rise to the lymphocytes of the immune system and the latter the megakaryocytes and thrombocytes, granulocytes, erythrocytes and monocytes and macrophages. These cells are called common lymphoid/myeloid progenitors, respectively (CLP/CMP)[5,6]. The Ikaros transcription factor drives differentiation into the lymphoid lineage[7], and the lineage soon branches in two, one giving rise to natural killer cells and the other, the small lymphocyte, to the progenitors for the B and T cells. The T-cell lineage is promoted by Notch signalling, assisted by GATA-3[8] and finished T-cells are the active part of the adaptive immune system. Expression of transcription factor Paired box 5 (Pax5) is required for B cell formation[9,10] and B cells and the plasma cells derived from them produce the antibodies that form the helper and memory part of the adaptive immune system.

Differentiation towards the CMP and the myeloid lineage is driven by PU.1[11,12] and CCAAT-enhancer binding protein α (C/EBPα)[13] and has more branches than the lymphoid, leading to a more diverse set of cells. For simplicity, the lineage is perhaps best explained by describing the examples of tight competition going on in progenitor cells as two or more factors carry out dual roles of promoting their own lineage while antagonising another (**Figure 1**): GATA-1 drives CMPs to the megakaryocyte/erythrocyte progenitor (MEP) lineage, while PU.1 pushes cells toward the granulocyte/macrophage progenitor (GMP) lineage[14,15]. Mirroring this, a study by Mancini *et al.* found Friend of GATA-1 (FOG-1) to drive CMPs towards the megakaryocyte/erythroid lineage while blocking further myeloid differentiation by antagonising C/EBPα[16]. This study also implicated GATA-1 in the next step, driving MEPs toward the erythroid lineage, together with Krüppel-

like factor 1 (KLF1) which in turn also antagonises Fli-1's push for megakaryocytic development[17]. And, finally, PU.1 and Growth factor independent 1 (Gfi-1) antagonise each other while promoting the GMPs towards monocyte/macrophage or granulocyte differentiation, respectively[18].

Most of the haematopoietic differentiation is completely reversible (the enucleated mature red blood cells (RBCs) are an obvious exemption). Lineage commitment is strong and robust, but at the same time fragile due to the relatively few factors and the lack of redundancy in the haematopoietic system. Therefore the complete loss of a single factor can be catastrophic, and many of the haematopoietic malignancies (blood cancers like leukaemia and myeloma) arise from harmful genetic mutations disrupting a single or a few crucial factors' normal function[19-21]. But this dependency can also be exploited as a potential therapeutic strategy by reprogramming one terminally differentiated blood cell into another using only a few essential factors. This is possible due to a process called lineage priming; while cells differentiate into one lineage, they never lose the differentiation scheme of other lineages, it is only temporarily repressed[22]. As such, researchers have been able to reprogram terminally differentiated blood cells into cells of another lineage by reactivating the differentiation scheme of that cell[23,24]. And, taking it a step further, researchers have also successfully reprogrammed neural stem cells and differentiated fibroblasts into haematopoietic lineages[25-27].

Plasma cell differentiation and Multiple Myeloma

Plasma cells develop from mature, activated B cells and their main function is to produce large numbers of immunoglobulins (antibodies). Like all other aspects of haematopoiesis, plasma cell formation is a tightly regulated process of differentiation and lineage specification, characterised by an antagonistic relationship with the B cell lineage, not unlike those found in the myeloid lineage[28,29]. Expression of Pax5 and a number of other factors and cytokines drive B cell lineage specification from the small lymphocyte to the mature, naïve follicular B lymphocytes that populate the spleen and lymph nodes[30,31]. Here they monitor the lymphatic system and when a B cell encounters a foreign antigen it is quickly activated and undergo proliferation, somatic hypermutation and immunoglobulin class switching. From here it differentiate into two populations; B memory cells capable of eliciting a quick antibody response, and long-lasting plasma cells that migrate back to the bone marrow and continually produce antibodies against the antigen that activated the B cell[32]. Reactivation of Pax5-repressed genes is an early first step of plasma cell differentiation[33], followed by the upregulation of transcription factors X-box-binding protein 1 (XBP1), interferon-regulatory factor 4 (IRF4) and B lymphocyte-induced maturation protein 1 (BLIMP1), that are the main drivers of plasma cell maturation[32]. Distinct

surface proteins are expressed on the mature plasma cell, a fact that is frequently exploited by researchers to sort and enrich plasma cell populations, which normally only constitute ~1% of the bone marrow cells. As such, CD27, CD38 and CD138 (Syndecan 1, SDC1) are highly expressed in mature plasma cells, while the common B cell markers CD19, CD20 and CD45 (Protein tyrosine phosphatase, receptor type C, PTPRC) are absent[34].

Multiple myeloma (MM) is defined as an un-inhibited, clonal growth of plasma cells in the bone marrow, producing a monoclonal immunoglobulin (M protein) that can be detected in peripheral blood[35]. It is preceded by a condition called monoclonal gammopathy of undetermined significance (MGUS), characterised by the presence of the M-protein, but absence of any MM clinical signs or symptoms[36]. MGUS progresses slowly (~1% of cases per year) to MM, which is characterised by presence of the M component, clonal growth of plasma cells to comprise at least 10% of the cells in the bone marrow and evidence of organ damage, such as lytic bone lesions[37]. The incidence of MM varies globally from about 1 in 100,000 in Asia, to 5-6 in 100,000 in Sweden and among Caucasians**,** to about 10 in 100,000 among Africans and in African Americans[38-41]. In addition to geographic origin, the incidence is influenced by age (more common above 60 years old) and gender (more common among men)[42]. Several risk loci have been identified for MM and MGUS[43-45], and several studies have reported an aggregation of cases in families[46-48]. There is currently no cure for MM, but therapies to extend the survival have been developed (recently reviewed by Bianchi *et al.*[49]), including melphalan-prednisone treatment, autologous HSC transplantation, immunomodulatory drugs (*e.g.*, thalidomide and lenalidomide), protease inhibitors (*e.g.,* bortezomib) and monoclonal antibodies targeting plasma cell markers, such as CD38 (daratumumab).

MM differs from other haematological cancers in that the malignant cells comprise a small proportion of the bone marrow cell population at diagnosis. Plasma cell surface markers are therefore a major focus of MM research, as they can be used in fluorescence-activated cell sorting (FACS) to extract plasma cells from a bone marrow sample, to be used for classification, characterisation and research. Common markers used in MM cell sorting are CD19, CD45 and CD56, as multiple myeloma plasma cells (MMPCs) are normally negative for the two former, and positive for the latter, whereas the direct opposite is the case for normal plasma cells[50]. Other markers include CD38 and CD138, with the latter preferred due to higher specificity towards plasma cells[51,52]. However, CD138 is quickly degraded *ex vivo*[52,53], requiring immuno-phenotyping analyses to be done shortly after sampling, which is often not possible in a clinical setting. Additionally, the short half-life of CD138 limits its usability with biobanked samples. This prompted us to search for a more suitable plasma cell marker and we found CD319 and CD269[54]. This discovery forms the basis of **Study IV**.

Erythropoiesis

Erythropoiesis is the branch of haematopoiesis that produces the RBCs crucial for transporting oxygen to all tissues in the body. Oxygen is transported in haemoglobin molecules within the RBCs, taking up as much as 98% of the entire cytoplasm[55]. An estimated 2.5 billion RBCs are produced every day, with the system capable of production bursts when needed[56]. The lifespan of an RBC is 120 days, in which it continually travels through the bloodstream and circulation, racing through the largest aortic blood vessels and squeezing through the smallest capillaries, made possible by the flexible cytoskeleton and the absence of a nucleus[57]. However, before all this can be achieved, an intricate regulatory process has transformed the MEP cell into a mature RBC, where perhaps the most striking feature is the extrusion of the nucleus, thus irreversibly committing the resulting reticulocyte (as the cells are called at the penultimate stage) into the mature RBCs formed as the cells enter the blood stream[58].

The major drivers in differentiating HSCs into the erythroid lineage were described above, but several stages still remain before the RBC is fully matured **(Figure 2)**. Apart from transcription factors, several cytokines and colony stimulating factors (CSFs) also influence the erythroid lineage, some of which will be highlighted in the following. Differentiation is initiated by GATA-2 and KLF1[59] together with interleukin-3 (IL-3), SCF and granulocyte-macrophage CSF (GM-CSF)[60] that transform the MEP cell to a burst forming unit-erythroid (BFU-E). GATA-2 blocks further erythroid differentiation and instead facilitates proliferation and a surge in immature erythroblasts[61]. Eventually, GATA-1 displaces GATA-2 in what is known as the GATA-switch[62,63], that facilitates the differentiation to colony forming unit-erythroid (CFU-E) cells[61], assisted by TAL1, SCF, IL-3 and GM-CSF, among others[64]. At this point, expression of the prime erythroid cytokine erythropoietin (Epo) and its receptor (EpoR) increases and Epo replaces SCF as primary stimulatory cytokine[65-67]. Epo:EpoR interaction triggers the signal transducer and activator of transcription 5 (STAT5). This initiates a signal transduction cascade that activates transcription of *Bcl-x*, transferrin receptor (*TFRC*, or *CD71*) and other transcription factors[68,69]. Also in the nucleus the pentameric complex consisting of GATA-1, TAL1, E2A, LMO2 and LDB1 enhances transcription of erythroid genes (more on this later) to differentiate the CFU-E into a pro-erythroblast[64]. Subsequently, the cell goes through a number of morphological changes known as basophilic erythroblasts, polychromatic erythroblasts and orthochromatic erythroblasts where Epo dependency decrease and upregulation of TFRC facilitates increased iron uptake and haem production in the cell[70,71]. Haem in turn increases expression of globin genes which facilitate formation of the vital haemoglobin molecules[72,73]. Simultaneously, production of several important cell membrane and cytoskeletal

proteins, such as ankyrin-1, Band 3, Band 4.1, Band 4.2, glycophorin A (GPA), Duffy, RhD and RhCE, is accelerated[74]. In the final stages of erythrocyte maturation histone deacetylation and DNA methylation increase as chromatin in the nucleus condenses; the cytoskeleton is rearranged and the nucleus is extruded from the erythroblast[75,76]. At this stage the cell is called a reticulocyte and enters the bloodstream, the remaining cell organelles and nucleus debris is cleared and differentiation into a mature biconcave RBC is complete[58].

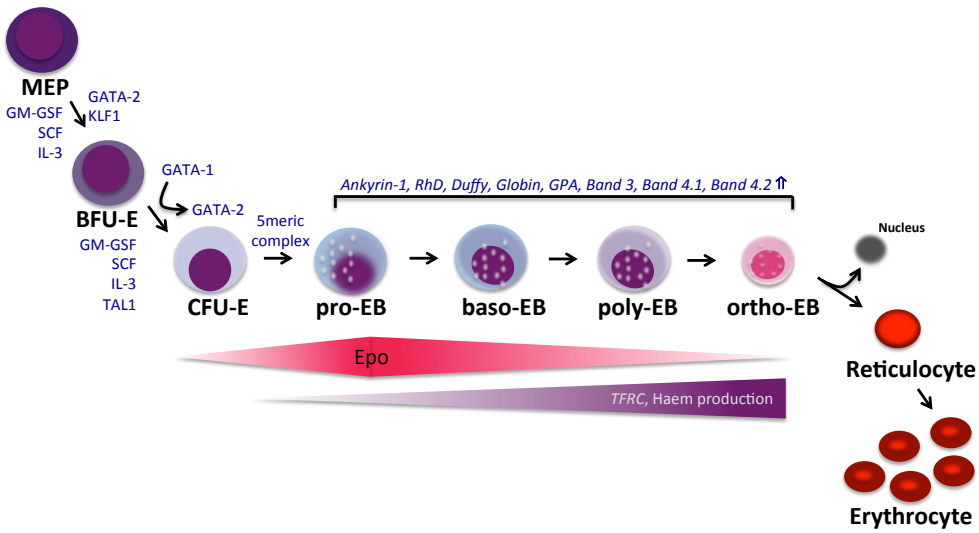


Figure 2 Erythropoiesis. Overview of erythropoiesis with the progenitor and precursor cells and the factors that guide them as described in the main text. For simplicity, not all cell intermediates nor factors are included. Abbreviations: MEP = megakaryocyte-erythrocyte progenitor; BFU-E = burst forming unit-erythroid; CFU-E = colony forming unit-erythroid; pro-EB = pro-erythroblast; baso-EB = basophilic erythroblast; poly-EB = polychromatic erythroblast; ortho-EB = orthochromatic erythroblast; 5meric complex = pentameric erythroid transcription factor complex consisting of GATA-1, TAL1, E2A, LMO2 and LDB1. Factor names are described in the main text.

Transcriptional regulation of erythropoiesis

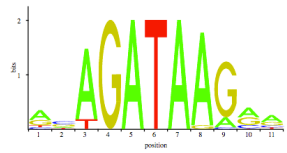
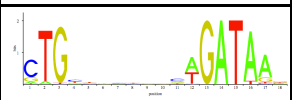
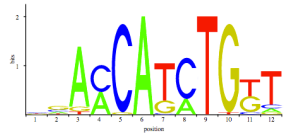
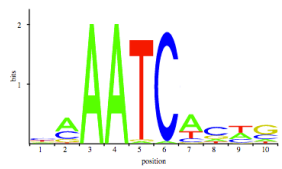
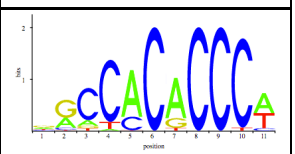
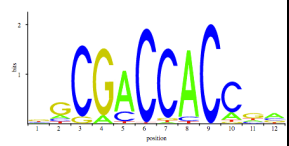
Erythropoiesis is a complicated process governed by a multitude of factors and regulators. Some factors have broad regulatory potential and are found in many different areas of haematopoiesis, while others are more specialised and only exert their influence in a single lineage or transition. Prime erythroid transcription factors (henceforth abbreviated eTFs for simplicity) such as GATA-1, KLF1 and TAL1 drive the haematopoietic cells towards the erythroid lineage, but also act to suppress those that guide the other lineages, such as PU.1 and Fli-1. The regulation by these crucial factors takes many forms and can act via direct DNA:protein or protein:protein interaction or attraction of additional factors, that facilitate chromatin remodelling or histone modifications. eTFs may influence transcription near their target or from many kilobases away, evident from a recent study that showed that GATA-1, TAL1 and KLF1 prefer to occupy sites in gene regions that are intronic (~47% of all sites) or distal (~26%, defined as between 50kb and 1kb from nearest gene), rather than intergenic (~18%, more than 50kb from nearest gene)[77].

Presence or absence of a given factor is not sufficient to determine its influence on transcription. Concentration is also an important determinant, as illustrated in the stoichiometric relationship between TAL1 and ETO2 described below and that erythroid maturation is impaired by lowering the concentration of GATA-1[78].

Although tightly regulated, erythropoiesis is carried out by relatively few main factors. As noted previously, researchers were recently able to reprogram human fibroblasts towards induced erythroid progenitors and established the essential transcription factors to be GATA-1, TAL1, cMyc and LMO2[27].

The following is a brief presentation of the most important regulators of erythropoiesis, followed by a description of how these come together in large regulatory complexes that are crucial for erythroid differentiation and synthesis of RBCs, summarised in **Table 1**.

Table 1: Transcription factors and complexes involved in erythropoiesis. Only those factors and target genes/effects described in the text are shown, each factor may have several other binding partners and related targets. To avoid overlap, complexes containing the same factors are only depicted for one of them, in no particular order.

Factor	Binding partners	Target genes / effects	DNA-binding motif[1]
GATA-1	STAT5	*FOG1* ⇑ *KLF1* ⇑ *EPOR* ⇑	
		ZBTB7A ⇑	
	SWI/SNF CBP/p300	*GATA2* ⇓ CR/HM[3]	
FOG-1	GATA-1	*GATA1* ⇑ *HBA* ⇑ *HBB* ⇑ *SLC4A1* ⇑	N/A
	GATA-1:MeCP1:NuRD	*GATA2* ⇓ CR/HM	
TAL1	Pentameric[2]	P-TEFb recruitment	
	LSD1:CoREST: HDAC1:HDAC2	*EPB42* ⇓ *GATA2* ⇓	
	Gfi-1B:E2A:ETO2	*HBA* ⇓ *HBB* ⇓ *ALAS2* ⇓	
Gfi-1B	GATA-1	*Bcl-X$_L$* ⇓ *GFI1B* ⇓	
	CoREST:LSD1: HDAC1:HDAC2	*GFI1B* ⇓ *c-Myb* ⇓ *MEIS1* ⇓ CR/HM	
KLF1		Fli-1 ⇓ γ- to β-globin switch	
	GATA-1 Pentameric (speculative)	Guidance/attraction to transcription factory	
LMO2	Pentameric	*LMO2* ⇑ *GYPA* ⇑ *HBA* ⇑ *KLF1* ⇑	N/A
ZBTB7A	MeCP1:NuRD	*Bim* ⇓ *HBG1/2* ⇓	

1: JASPAR database (http://jaspar.genereg.net/)
2: The pentameric complex consists of GATA-1, TAL1, E2A, LMO2 and LDB1.
3: CR = chromatin remodelling; HM = histone modification

GATA-1

GATA-1 is a member of a family of zinc-finger transcription factors and is involved in many stages of erythropoiesis. One of the first erythroid factors to be discovered[79,80] and arguably the most investigated, GATA-1 is implicated in several erythroid processes, such as the STAT5 signalling cascade which facilitates upregulation of several erythroid and haemoglobin genes, including *EPOR, KLF1* and *FOG1*[81,82]. GATA-1 also exhibits features that are common for the major regulators of haematopoiesis, namely the ability to activate erythroid-specific genes and repress or disrupt genes that drives other lineage. Examples of this include the antagonistic direct interaction between GATA-1 and the myeloid lineage-specific factor PU.1[15,83] or, even more strikingly, that overexpression of GATA-1 results in complete reprogramming of lymphoid and myelomonocytic cells into early erythroblasts[23,24]. GATA-1 is in the centre of the intricate erythropoietic network, by being involved in most of the complexes that govern it and loss of GATA-1 not surprisingly results in anaemia, thrombocytopenia and other malignancies[84,85].

The GATA family of transcription factors was named for their preferential binding to the DNA sequence motif (A/T)GATA(A/G), however it has more recently been discovered that GATA-1 does not always occupy these sites during erythropoiesis[86,87]. These genome-wide occupancy studies also found that GATA-1 mainly occupies sites within the target gene or in intergenic regions, with only 10-15% binding to the proximal promoter, which corresponds fairly well with the study by Su *et al.*[77] referenced above, which found nearly half of all occupancy sites in intronic regions. The core GATA DNA binding motif is crucial, perhaps with the exception of the fourth and final nucleotide, as altering the motif to GATT had little effect on murine GATA-1 protein binding efficacy[88].

FOG-1

The direct protein interaction between GATA-1 and FOG-1 is indispensable for erythroid differentiation[89]. FOG-1 is a co-regulator in the GATA-switch where GATA-1:FOG-1 blocks GATA-2 occupancy in its own *GATA2* promoter abrogating its positive autoregulation[62]. The switch also involves chromatin remodelling which shuts down *GATA2* expression, thus driving progenitor cells further towards erythroid differentiation[63]. Although no direct DNA-binding has been demonstrated for FOG-1, it is known to associate with GATA-1 and GATA-2 and the chromatin remodelling is not exclusive for the GATA-switch, leading researchers from Stuart Orkin's group to propose FOG-1 as a new class of molecules functioning as chromatin occupancy facilitators (COFs)[62].

TAL1

TAL1 is named after its discovery as a genetic translocation (t(1:14)(p32:q11)) causing T-cell acute lymphoblastic leukaemia[90-92]. TAL1 is expressed throughout haematopoiesis and is involved in many erythroid proliferation and maturation processes. TAL1 is important in haematopoiesis and erythropoiesis with specific DNA binding not necessarily required in the early stages of haematopoiesis[93] yet critically required throughout most of erythropoiesis[94]. DNA binding is carried out through TAL1's basic helix-loop-helix domain with the core DNA binding motif CANNTG, also called an E-box. Composite motifs with a GATA site 9 to 12 nucleotides from the E-box (designated CANNTG($n_{9\text{-}12}$)GATA) are often found in genes directly regulated by TAL1, and it is the preferred binding motif for the pentameric complex[95-97]. Additionally, a shortened motif for TAL1, CTG, 9 nucleotides upstream of the GATA-1 motif was recently identified[98]. In general, GATA-1 and TAL1 have very similar expression patterns in haematopoiesis (*i.e.* mainly expressed in erythroid, megakaryocytic and mast cells[99]), they are both crucial for correct erythropoiesis, and they often co-localise and affect gene transcription together. A TAL1 global occupancy network analysis determined its preference to intra- and intergenic over proximal promoter regions and that about half of the genes regulated by TAL1 are upregulated, with an added interesting notion that for those genes where TAL1 was found directly occupying the target DNA site, 75% were found to be upregulated, indicating that TAL1 is mainly a transcriptional activator[98].

Gfi-1B

Gfi-1 and its paralogue Gfi-1B are zinc-finger transcription factors with an N-terminal Snail/Gfi-1 (SNAG) repressor domain. They both bind a DNA sequence motif of TAAATCAC(A/T)GCA, core motif AATC[100,101]. Expression of both proteins occurs in early haematopoiesis, but diverges as the cells differentiate, with Gfi-1 mainly being expressed in the lymphoid and some myeloid lineages, and Gfi-1B mainly found in the erythroid and megakaryocytic lineages[102,103], where it is an indispensable factor in early erythropoiesis promoting erythroblast proliferation, but not differentiation[104]. GATA-2 is known to activate Gfi-1B and repress lymphopoiesis, whereas Gfi-1 was recently shown to inhibit GATA-2 function, presenting an interesting three-way regulatory network governing cell fate of early haematopoietic development[105]. Gfi-1B acts as a repressor of transcription by direct DNA binding to target sites and recruitment of histone modification enzymes[106], but also by direct interaction with GATA1 to repress *Bcl-x_L*[107,108] and *GFI1B* itself[109].

KLF1

KLF1 was first described in 1993, originally named erythroid Krüppel-like factor (EKLF), based on similarity to a gene found in *D. melanogaster*[110]. It is exclusively found in erythroid and mast cells and is a crucial component of the erythrocyte/megakaryocyte lineage switch[17,111]. KLF1 preferentially binds to the DNA sequence motif of CCNCNCCCN, with the core motif being CACCC[110].

Research in KLF1-deficient mice looking at various globin genes have established, that KLF1 is likely to participate in the switch from fetal γ-globin to adult β-globin[112,113].

Together with GATA-1, KLF1 is probably the most well-studied eTF. In two gene expression profiling studies of KLF1-null cell lines, two-thirds of the differentially expressed genes were downregulated, and most of these genes showed direct binding of KLF1, indicating that KLF1 mainly acts as a transcriptional activator[114,115], however examples of repressor function have also been demonstrated[116,117].

An interesting discovery has coupled KLF1 to an emerging gene transcription paradigm: That chromosomal organisation inside the nucleus is highly plastic and chromosomes intertwine and flow freely in the nucleus, and that transcription is localised to focal spots, or transcription factories where the chromosomes are transported to, rather than the transcription machinery moving around inside the nucleus[118-120]. KLF1 was found to be an important factor, at least for haemoglobin genes, as it was shown to be capable of "pulling" and guiding target genes towards the factory[121]. This model may also explain KLF1's long-range *cis* and *trans* regulation, and indeed, KLF1-enriched binding sites were shown to primarily be found more than 10kb away from the nearest known gene[115].

LMO2

The LIM-only Domain 2 gene was discovered by investigating the 11p13 T-cell translocation cluster, which is a hotspot for translocations in T-ALL[122]. It contains two LIM-only class zinc-fingers and an N-terminal domain that is capable of transcriptional transactivation, *i.e.* to interact with other factors and activate them[123]. LMO2 is a critical eTF as absence of LMO2 leads to abrogated yolk sac erythropoiesis and embryonic lethality[124].

To date, no direct DNA-binding has been described for LMO2, rather it is crucial as a bridging factor between TAL1 and GATA-1 in the pentameric complex also containing E2A (also known as transcription factor 3 (TCF3)) and LIM domain binding 1 (LDB1)[125,126]. This complex (or at least LMO2:GATA-1:TAL1) regulates expression of the *LMO2* gene itself[127]. Not surprisingly, the expression pattern of *LMO2* follows closely that of GATA-1 and TAL1 discussed above.

ZBTB7A

Zinc Finger And BTB Domain Containing 7A (ZBTB7A) is also known as Leukaemia/Lymphoma Related Factor (LRF), as well as a number of more or less colourful names (*e.g.* "Pokemon" and "FBI"). As the names suggest, ZBTB7A contains four DNA-binding zinc finger domains and a Broad Complex, Tramtrack, and Bric-a-brac (BTB) domain that facilitates protein:protein interaction and functions as a transcriptional repressor[128,129]. Upregulation of ZBTB7A is facilitated by GATA-1 binding directly to the *ZBTB7A* promoter and ZBT7A is essential in erythropoiesis to antagonise the proapoptotic factor *Bim* in late-stage erythroblasts[130]. ZBTB7A has also been shown to repress fetal γ-globin, and is thus an important factor in early haematopoiesis[131]. The DNA binding motif of ZBTB7A has been ambiguously reported as G(A/G)GGG(T/C)(C/T)(T/C)(C/T)[132], NGCGACCACCNN[133], (G/A)(C/A)GACCCCCCCCC[129] and GCGGGGGCGGGGGC[130], but with the former remarking that the consensus motif seems to be fluid and preferential ZBTB7A binding to DNA quite flexible. The latter two motifs found by Maeda and colleagues do have similarities (if we assume that one of them is in reverse complement orientation), and since only these two studies investigated ZBTB7A's role in erythropoiesis, this may indicate an erythroid-specific DNA binding motif.

Erythropoietic transcription factor complexes

Regulation of erythropoiesis by the above eTFs is never a one-factor task. Albeit a limited number of factors, they interplay and associate in various highly dynamic multimeric complexes (**Figure 3**) and may form, reform and dissolve throughout the differentiation from HSCs to mature erythrocytes. A given factor can thus either activate or repress transcription of its target genes, depending on which cofactors it associates with and where in the maturation process the cell is.

GATA-1 plays a central role in most of these complexes, and it was traditionally believed to primarily be a transcriptional activator. However, three genome-wide ChIP-seq studies found an estimated 41-60% of the genes, that GATA-1 directly influenced expression of, to be downregulated[86,87,134]. As expected, GATA-1's role as a transcriptional activator or repressor depends on which cofactors it associates with.

Transcriptional activation of target erythroid genes is mediated mainly through the pentameric complex. As described earlier, GATA-1 and TAL1 bind directly to DNA sequences (the E-box/GATA composite motifs), with LMO2 acting as a bridge or spacer molecule[126] and E2A functions as an obligate heterodimer of TAL1 to facilitate DNA-binding[98,135]. LDB1 is responsible for recruiting the kinase Positive transcription elongation factor b (P-TEFb) to the complex, which

in turn phosphorylates RNA Polymerase II, a crucial requirement for activation of the transcription machinery[136]. Another proposed function of LDB1 is to facilitate long-range interaction between two pentameric complexes through its self-association domain[137,138]. This is thought to be accomplished by the DNA bending or looping to bring the two distant complexes occupying promoter/intron and distal enhancer regions together and then tethered by direct LDB1:LDB1 interaction. This may explain why the pentameric complex is often found occupying sites several kilobases from the transcription start site of their target genes[86,87,98,115]. The pentameric complex has been shown to activate important erythroid genes such as glycophorin A (*GYPA*), the α-globin locus (*HBA*) and *KLF1*[95,107,139].

GATA-1 and KLF1 also form an activation complex, illustrated by Tallack *et al.*[115], who compared occupancy maps of GATA-1 and KLF1 and found that in nearly half of the sites they appeared within 1kb of each other with a peak distance of only 18 bp. Interestingly, they found a CTG sequence (TAL1 binding motif) often to be present in between the two sites, however while GATA-1:TAL1 and GATA-1:KLF1 complexes were abundant, no trimeric GATA-1:KLF1:TAL1 complex was found. Not many genes have so far been found as direct targets of GATA-1:KLF1 complexes, but given the proximity of the two factors, it is speculated that KLF1 facilitates transcription of pentameric-bound genes by translocating the gene region into the transcription factories described earlier[138].

Not surprisingly, GATA-1 and FOG-1 associate readily, but the complex represents dual roles depending on additional co-factors in the complex. GATA-1:FOG-1-mediated activation has been reported in many genes, for example α- and β-globin (*HBA* and *HBB*), Solute carrier family 4 member 1(*SLC4A1/*Band 3) and *GATA1* itself[62,140,141] but these same studies also presented numerous examples of transcriptional repression. This is facilitated by association with the multimeric MeCP1:NuRD complex, which contains multiple proteins capable of both chromatin remodelling and histone deacetylation, leading to tighter chromatin packing and thus decreased expression[142,143]. This complex was shown to bind directly to the distal promoter of *GATA2*, providing strong evidence that it is at least partly responsible for the decreased expression of *GATA2* in erythropoiesis[107]. While GATA-1 seems to be the one that tethers the complex to the DNA of the target gene, FOG-1 is responsible for attracting and binding the MeCP1:NuRD complex[144]. Recently it was also shown that ZBTB7A interact directly with the NuRD-complex and represses foetal γ-globin genes (*HBG1* and -*2*) keeping their expression low in adults[131].

Chromatin remodelling is the proposed function of the Gfi-1B:CoREST:LSD1:HDAC1:HDAC2 repressor complex as well. Gfi-1B is thought to bind target DNA and recruit the other proteins through its SNAG

domain. Lysine-specific histone demethylase 1A (LSD1) then demethylates lysine 4 on histone H3 (H3K4), which leads to tighter chromatin packing and limits access to DNA[145]. This repression affects the genes *GFI1B* itself (establishing a negative feedback loop), *c-Myb* and *MEIS1*[145,146].

Another Gfi-1B complex with cofactors TAL1, E2A and ETO2 associates in early erythroid progenitors (but not in megakaryocytes) and represses transcription of TAL1 target genes *HBA*, *HBB* and *ALAS2*[147]. This complex is absent in late erythropoiesis, signifying a role as an important erythroid differentiation switch[148]. The ETO-mediated repression is proposed as direct ETO2:E2A interaction which interferes with its association with transcriptional activator p300/CREB, while the de-repression and switch towards erythropoietic differentiation happens through an increase in the stoichiometric ratio of TAL1 to ETO2 present in the cells[149].

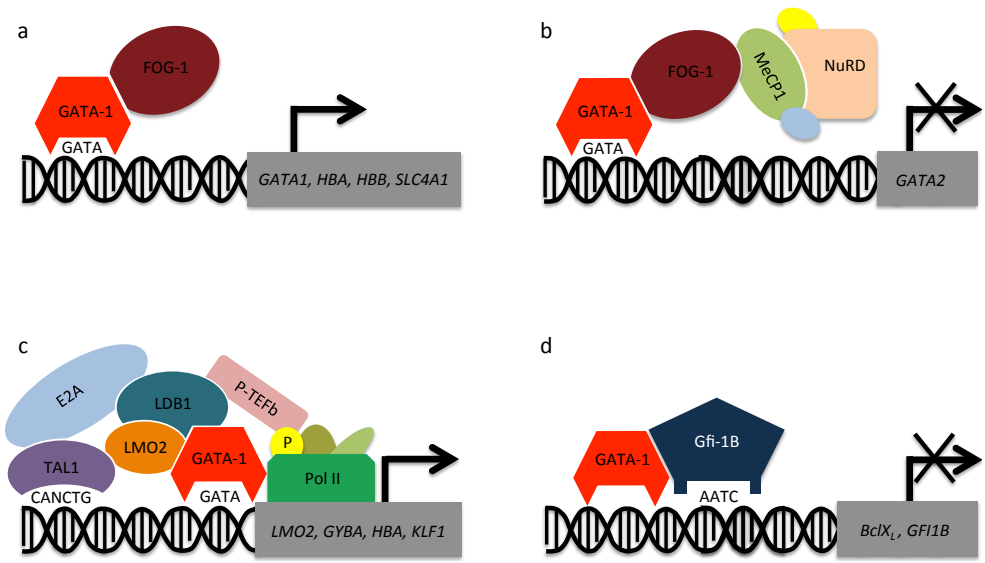


Figure 3 Erythroid transcription factor complexes. Selected complexes crucial for erythropoiesis. For simplicity, only the factors described in the main text are depicted, the complexes may contain several other co-factors. See also Table 1 for individual factors and binding motifs. a) GATA-1 binds to DNA and recruits FOG-1 to activate transcription of *GATA1*, *HBA, HBB and SLC4A1*. b) The GATA-switch. GATA-1:FOG-1 complex may also instigate transcriptional repression by attracting the MeCP1:NuRD complex that facilitate chromatin remodelling and histone modification, which limit transcription machinery access to *GATA2* and other genes. c) The pentameric complex (GATA-1:TAL1:E2A:LMO2:LDB1) facilitates transcriptional activation by recruiting the kinase P-TEFb that phosphorylates polymerase II, a crucial requirement for transcription initiation. d) GATA-1 and Gfi-1B interact at AATC sites to facilitate downregulation of *GFI1B* and *Bcl-XL*. The precise mechanism of downregulation is not clear, but Gfi-1B has been show to associate with the CoREST-complex (not shown) that facilitates chromatin remodelling and histone modification.

Additional layers of erythroid transcriptional regulation

It goes without saying that a system as complex as erythropoiesis needs thorough and robust regulation beyond the level of transcription factor complexes. Indeed, epigenetics play a major role, as already described above, but other regulatory mechanisms are also present.

Apart from the epigenetic examples already presented, GATA-1 also interacts with the SWI/SNF chromatin remodelling complex[150] and CBP/p300 histone acetyltransferases[151]; Grass and colleagues showed broad histone deacetylation dependent on FOG-1[63] and Hu and coworkers presented a repressor complex consisting of TAL1, LSD1, CoREST and histone deacetylase 1 and 2 (HDAC1/2) playing a crucial role in the onset of erythroid differentiation through repression of *GATA2* and *EPB42*[152].

Gene silencing, which is mediated by microRNAs and capable of degrading or sequestering newly transcribed mRNA, is a common posttranscriptional regulation process in cells[153,154]. Several thousand microRNAs have been discovered so far, and some of them have been associated with erythropoiesis. As such, miR-223 was shown to target LMO2 and downregulation of this microRNA is crucial for terminal differentiation[155], while miR-191 downregulates *Riok3* and *Mxi1* and facilitates mouse erythroblast enucleation[156]. It is beyond the scope of this thesis to go into more detail about this intriguing layer of regulation, additional knowledge can be gained from these excellent reviews[157,158].

Finally, the long-noncoding RNAs (lncRNAs) represent yet another layer of erythroid regulation. Knowledge of these abundant RNA elements, that play potentially crucial roles in gene regulation, is rapidly expanding and a few studies on erythroid-specific interaction have emerged. The LincRNA erythroid prosurvival (LincRNA-EPS) was found to promote erythropoiesis by halting apoptosis[159], while another study looked at genome-wide expression of lncRNAs in erythropoiesis and found major influences on RBC maturation and chromatin structure, and showed promoter occupancy by GATA-1, KLF1 and TAL1 at genes coding for lncRNAs, indicating erythroid-specific expression[160].

Blood group systems

Humans have been interested in blood transfusion since the Middle Ages, experimenting with more or less creative methods, including transfusing blood from dogs into live human beings, unsurprisingly with very limited success and very high mortality rate. Modern transfusion medicine owes much to the Austrian physician and scientist Karl Landsteiner, who at the turn of the 20th century conducted research into donor-recipient compatibility. By then it was known that mixing blood from two patients would cause the blood to clump, or agglutinate, but this was thought to be because of the sicknesses the patients had. Landsteiner was the first to mix blood from healthy individuals, and he quickly realised that the agglutination also happened here, yet not every time. Landsteiner also separated the RBCs from the plasma and was able to show, among other things, that RBCs from one individual could be mixed with plasma from another, but that same individual's RBCs would clump when mixed with the first individual's plasma[161]. His work resulted in the classification of the first blood group system, which he dubbed ABC (later renamed to the ABO we know today) and established the antibody-antigen relationship that forms the basis of blood group systems and transfusion medicine to this day.

Certain RBC membrane proteins and molecules exhibit minor variations which are recognised as blood group antigens. They play an important role in blood transfusions, where donor-recipient compatibility is paramount. A potentially fatal haemolytic transfusion reaction (HTR) may occur if foreign RBC antigens of the donor blood trigger the recipient's immune system to produce alloantibodies. In the most severe cases, these antibodies may bind complement and cause intravascular haemolysis. The alloantibodies reside in the plasma, and in the ABO blood group system, a type A individual will thus have the A antigen on the RBC surface and anti-B antibodies in the blood plasma, and vice versa for a type B individual. Type O has no A/B antigens on the RBC membrane, but has both anti-A and anti-B in the plasma, while type AB has both A/B antigens on the RBCs, and none of the antibodies in the serum. Apart from the obvious incompatibility between A and B, it follows that type O RBCs can be safely transfused to all other types, but can only receive RBCs from other type O individuals, while type AB can only be administered to other type AB individuals, yet a type AB individual can receive blood from any other type (since the recipient's plasma contains neither anti-A or anti-B). It should be noted that even naïve individuals (never having received a blood transfusion) express alloantibodies against the non-self A/B antigen(s). It is commonly believed that we are exposed to microorganisms exhibiting structures very similar to A and B antigens early in life and thus have

alloantibodies against these readily available in the blood[162,163]. This is in contrast to most other blood group antigens, where alloantibodies may be produced following transfusion with mismatched donor blood (or in some cases pregnancies of blood group negative mothers giving birth to positive babies), and as such will present a concern only on subsequent transfusions/pregnancies. The ABO system was the first blood group system to be described, and the antigen-antibody relationship resembles that of most other blood groups, yet the severity of mismatched transfusion reactions varies significantly.

Blood group terminology and nomenclature

Based on the breakthroughs made by Landsteiner and others, transfusion medicine became increasingly successful throughout the 20th century. Yet with increased volumes of performed transfusions, reports of HTR incidents, where the recipient reacted strongly against the donor blood, also surged. This resulted in an ever growing number of new antigens and blood groups reported by transfusion specialists from all around the world and to bring order to the different reporting and naming schemes, the International Society of Blood Transfusion (ISBT) instigated the Working Party on Red Cell Immunogenetics and Blood Group Terminology in 1980 to come up with a uniform nomenclature for all antigens and blood groups. This resulted in a naming and numbering scheme of blood group systems and the antigens defining them, with the remaining antigens not belonging to any system divided into two groups according to their population frequency, the high and low incidence antigens[164]. The working party has convened biannually ever since to characterise and catalogue the currently 36 blood group systems (**Table 2**), six collections (antigens with insufficient information to be assigned to a particular blood group), and two series (antigens waiting for their molecular background to be determined), all in all comprising 352 antigens.

The ISBT guidelines state that in order for antigens to be considered as a blood group system, the genetic background, *i.e.* the sequence variant and/or mutation creating the antigen, which raises the alloantibody in the recipient, must be known. The molecular background of the ABO system was resolved in 1990, when Yamamoto *et al.* found that the A and B antigens are determined by different sugar molecules attached to the RBC membrane by a specific glycosyltransferase enzyme, encoded by the *ABO* gene[165]. A mutation in this gene gives a slightly different version of the enzyme, which in turn affects which sugar molecule it attaches to the sugar backbone (also known as the H antigen) on the RBC membrane. People with the O blood type carry mutation(s) that completely abolishes the enzyme function, resulting in no sugar molecules being added to H, while type AB carry both variants of the enzymes and thus both A and B antigens can be found on the RBC membrane.

Table 2: Blood groups systems recognised by the ISBT, December 2016.

#	System name	System symbol	Gene name(s) (HUGO)	Number of antigens	Structure[1]
001	ABO	ABO	*ABO*	4	Carbohydrate
002	MNS	MNS	*GYPA, GYPB, (GYPE)*	48	Glycocalyx
003	P1PK	P1PK	*A4GALT*	3	Carbohydrate
004	Rh	RH	*RHD, RHCE*	54	Membrane transporter
005	Lutheran	LU	*LU*	21	Receptor/adhesion
006	Kell	KEL	*KEL*	35	Enzyme
007	Lewis	LE	*FUT3*	6	Carbohydrate
008	Duffy	FY	*ACKR1*	5	Receptor/adhesion
009	Kidd	JK	*SLC14A1*	3	Membrane transporter
010	Diego	DI	*SLC4A1*	22	Membrane transporter
011	Yt	YT	*ACHE*	2	Enzyme
012	Xg	XG	*XG, MIC2*	2	Receptor/adhesion
013	Scianna	SC	*ERMAP*	7	Receptor/adhesion
014	Dombrock	DO	*ART4*	8	Enzyme
015	Colton	CO	*AQP1*	4	Membrane transporter
016	Landsteiner-Wiener	LW	*ICAM4*	3	Receptor/adhesion
017	Chido/Rodgers	CH/RG	*C4A, C4B*	9	Complement
018	H	H	*FUT1*	1	Carbohydrate
019	Kx	XK	*XK*	1	Structural
020	Gerbich	GE	*GYPC*	11	Structural
021	Cromer	CROM	*CD55*	18	Complement
022	Knops	KN	*CR1*	9	Complement
023	Indian	IN	*CD44*	4	Receptor/adhesion
024	Ok	OK	*BSG*	3	Receptor/adhesion
025	Raph	RAPH	*CD151*	1	Receptor/adhesion
026	John Milton Hagen	JMH	*SEMA7A*	6	Receptor/adhesion
027	I	I	*GCNT2*	1	Carbohydrate
028	Globoside	GLOB	*B3GALT3*	2	Carbohydrate
029	Gill	GIL	*AQP3*	1	Membrane transporter
030	Rh-associated glycoprotein	RHAG	*RHAG*	4	Membrane transporter
031	Forssman	FORS	*GBGT1*	1	Carbohydrate
032	Junior	JR	*ABCG2*	1	Membrane transporter
033	Lan	LAN	*ABCB6*	1	Membrane transporter
034	Vel	VEL	*SMIM1*	1	Unknown
035	CD59	CD59	*CD59*	1	Complement
036	Augustine	AUG	*SLC29A1*	2	Membrane transporter

1) From Daniels 2011[166]. For CD59[167,168] and for Augustine[169].

Structure and function of antigens

Blood group antigens are found on well-defined structures on the surface of RBCs. These structures are often classified into six groups according to proposed or known function[170]: the glycocalyx, membrane transporters and channels, receptors and adhesion molecules, complement regulatory proteins, enzymes and cytoskeletal proteins (**Table 2**). The ABO antigens belong to the carbohydrates of the glycocalyx, sugar molecules that make up an outer protective layer of the RBC membrane. The structure of SMIM1, the protein responsible for the Vel blood group antigen, has been predicted to be a single-pass transmembrane protein (TMP)[171], yet a function has not yet been described.

In contrast to their crucial role in transfusion medicine, most of the molecules that carry blood group antigens seem to be non-essential for the individual, as evidenced by the null individuals of various blood group systems (*i.e.* individuals homozygous for mutated and dysfunctional copies of the blood group gene, so that no antigen is expressed on RBCs), who show little to no signs of compromised health[172]. In most cases, other proteins take over and problems directly related to the lack of the antigen/protein only arise when the individual is exposed to stressful conditions[173]. This raises the question why we even have blood group systems, which is perhaps best answered by Geoff Daniels in his book *Human Blood Groups*, as he muses over a quote from Charles Darwin's *On The Origin of Species*: That (antigen) polymorphisms may have risen through many centuries or possibly millennia of evolutionary selection pressure, which, in our modern world today, no longer exists, making blood group antigens largely vestigial as remnants of a long forgotten past (paraphrased from[166]).

Pathological implications of blood group systems

There are several examples of blood antigens being remnants from selection pressures in early human history, and while modern medicine has definitely lessened or even removed some of this pressure, blood group antigens and certain pathological conditions remain closely linked to this day.

The majority of examples revolves around the malaria parasite, whose life cycle is intricately linked to human blood; the parasite infects RBCs by directly interacting with membrane receptors and multiply within before bursting the cells and finding new to infect[174]. Malaria is a crippling disease killing an estimated 429,000 people annually (World Health Organisation estimate for year 2015), and as such have been, and continues to be in some areas, the biggest evolutionary selection pressure in the world today[175].

The ABO system seems intricately linked to the malaria species *Plasmodium falciparum*. Cserti and Dzik performed a thorough analysis of the numerous

published studies linking ABO to malaria (some of which deemed to be flawed, some well-designed) and showed a convincingly high prevalence of the O blood group accompanied by a lack of A individuals in many areas around the equator where *P. falciparum* is endemic[176]. A functional explanation for this was reported by Carlson and colleagues[177] showing that rosette-formation, *i.e.* clumping of infected RBCs to uninfected, is highest in blood group A and AB-individuals, intermediate in B and weakest in group O. The molecular biology behind rosette-formation is the parasite-derived *P. falciparum* erythrocyte membrane protein-1 (PfEMP-1) expressed on infected RBCs, and this associates strongly with the sugar moieties of blood group A and B[178]. Subsequently, PfEMP-1 has also been shown to bind complement receptor 1 (*CR1*, also known as glycoprotein CD35), which is the gene associated with the Knops blood group system[179].

Several other blood group genes have been proven as receptors for *P. falciparum* infection: glycophorin A[180] and B[181] (blood group system MNS), glycophorin C[182] (Gerbich) and basigin[183] (Ok), all summarised in a recent review[184].

The Duffy blood group system, first described in 1950[185] and later found to be determined by the Atypical chemokine receptor 1(*ACKR1*)[186,187], has been linked to two other malaria species, *P. vivax* and *P. knowlesi*. A mutation in this gene results in the Duffy-null phenotype Fy(a–b–)[188], which is highly prevalent in West Africa[189]. As it happens, ACKR1 is the primary receptor used by *P. vivax* and *P. knowlesi*[190,191], which undoubtedly have represented a solid selection pressure and have been determined to be the cause for the extremely high prevalence of Fy(a–b–) in West Africa, and as a result, the very low occurrence of *P. vivax* and *P. knowlesi* in these regions[190,192].

Other malignancies with ties to blood groups include evidence that having a specific ABO allele increases the risk of developing certain cancers, where carrying any non-O allele makes an individual susceptible to develop gastric (60 years of research summarized in a meta-analysis by Wang *et al.*[193]) and pancreatic cancer (multiple studies, summarized in[194]), the latter presumably through increased glycosyltransferase activity of the A^1 allele compared to A^2 and B[195,196].

Furthermore, the ABO and Lewis blood group systems both contain histo-blood group antigens (A, B, H and Le^b) meaning that the antigens are present on other tissues than blood, and these systems must be considered with regards to organ transplantation and host-graft rejection. H and Le^b antigens have also been shown to be receptors for *Helicobacter pylori* bacterial infection[197].

Last but not least, and an exception to the rule that individuals with blood group null phenotypes are generally in good health, Kx-null individuals, carrying a mutation in the XK gene, suffer from McLeod syndrome presenting with acanthocytosis, neurological disorders and cardiomyopathy[198].

The Vel blood group system

The Vel antigen was first described in 1952 in New York, when a female patient, Mrs. Vel, presented with an acute HTR after receiving a second transfusion of crossmatch-compatible group O RhD-negative blood[199]. Her serum was tested against 10,000 blood donors in New York and only four were compatible, establishing a Vel-negative (Vel–) frequency of 1 in 2,500 individuals and the antigen was thus placed in the high incidence antigen collection by the ISBT. Other screening studies in the UK, Australia, Norway, Finland and France found a total of 43 Vel– people out of 159,565, giving an estimated global incidence of 1 in 3,711[200-205], summarised by Issitt and Anstee[206]. Interestingly though, a study performed by Cedergren *et al.*[207] found a higher prevalence of 1 in 1,761 individuals in Northern Sweden, pointing towards a potential founder effect for this historically isolated and sparsely populated area.

The Vel antigen is a major concern in transfusion due to the severity of HTRs in some transfused individuals. Vel– individuals usually present without anti-Vel in their blood, but develop alloantibodies upon contact with the Vel antigen. The antibody present as both IgG and IgM, but rarely causes haemolytic disease of the foetus and newborn (HDFN)[206]. Antigen expression on the RBCs of Vel-positive (Vel+) individuals vary greatly, with some having such low values in serological tests that they are at risk of being typed as Vel– (henceforth referred to as Vel+^{weak})[171]. Furthermore, due to the low incidence of Vel– individuals, a sufficient supply of anti-Vel for screening has not been available, which has made phenotyping, research and classification all the more demanding. In 1968, Issitt *et al.* reported a second Vel antibody, Vel 2, and presented cases characterised as Vel:1,–2 and Vel:–1,–2[208], however the authors later disputed their own findings deeming the difference between the two antibodies to be purely quantitative[206]. The Vel antigen remained a mystery until 2013 when we and two other groups independently, and more or less simultaneously, discovered that the Vel– phenotype is caused by homozygosity for a 17-bp deletion in the coding sequence of the erythroid gene small integral membrane protein 1 (*SMIM1*)[171,209,210]. This formally established Vel as a blood group system, recognised as system 034 by the ISBT in 2014[211]. **Study I** describes this discovery.

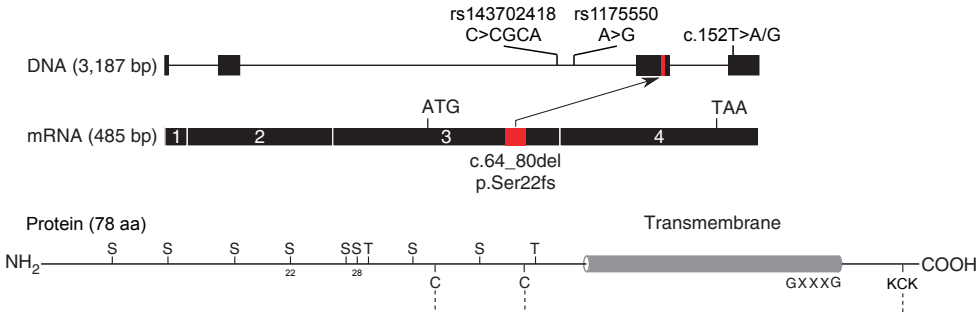


Figure 4. SMIM1 gene, transcript and protein. Four exons make up the 78 amino acid protein. The genomic sequence contains at least four sequence variants influencing the expression of SMIM1. The red patch is the Vel-defining 17-bp deletion in exon 3. The SMIM1 protein is a single-pass transmembrane protein with several serines and threonines available for phophorylation, three cysteine residues to facitate protein:protein interaction and a GXXXG homodimer motif. The final three C-terminal amino acids Lys76-Cys77-Lys78 is hypothesised to have a role in Vel antigen epitope presentation. Adapted from Study I.

The SMIM1 gene responsible for the Vel blood group system

SMIM1 is a small gene of 3,187 bp located on chromosome 1p36.32 (**Figure 4)**. It is highly conserved with homologs in many different species[171] and has four exons with the open reading frame spanning exon 3 and 4. As noted above, a 17-bp deletion in the open reading frame of exon 3 (c.64_80delAGCCTAGGGGCTGTGTC) results in a frame-shift mutation and stop codon skip, creating a dysfunctional protein, and thus a SMIM1-null phenotype. Other sequence variants are found throughout the gene, and in intron 2 in particular, in a region which is also rich in transcription factor binding sites, eTF occupancy (GATA-1, TAL1, ZBTB7A and others) and chromatin remodelling favouring active DNA transcription (as found by the Encyclopedia of DNA Elements (ENCODE)[212]). Previous studies showed the sequence variant rs1175550 present in this region to directly influence the expression of the Vel antigen[210,213], making it a strong candidate to explain the large variation in Vel antigen expression routinely observed with Vel+ individuals. We confirmed this finding and also found an additional sequence variant, rs143702418 located only 96 bp upstream of rs1175550, which independently influence *SMIM1* and Vel antigen expression[214] (**Study II** of this thesis). An additional sequence variant in the coding region c.152T>G/A causes missense mutations p.Met51Arg and p.Met51Lys, respectively, and both mutations have been associated with lower Vel antigen expression[210].

The function of SMIM1 remains unknown; however, structure analysis predicted it to be a single-pass TMP with a GXXXG motif in the transmembrane domain[171], which is commonly associated with homodimer formation[215]. An *in silico* gene network model showed *SMIM1*'s closest neighbours to be typical erythroid genes

(*RHD*, *XK*, *KEL*, *KLF1 etc.*)[171]. We originally proposed it to be a type I TMP, but a recent study reported findings to support a type II TMP with a very short C-terminal extracellular domain of only 3-12 amino acids[216]. This study also pinpointed the final three amino acids (KCK, or Lys76-Cys77-Lys78) as crucial for the expression of the Vel epitope recognised by human anti-Vel. Finally, large GWA studies have linked rs1175550 with mean corpuscular haemoglobin content (MCHC)[217] and blood copper concentration[218], suggesting that SMIM1 could play a role in copper transport, either as a main transporter or a cofactor of known copper transporters.

Large-scale analyses of genetic variation

Mutations in the genome occur naturally in the cell as the DNA is replicated, but these errors are quickly repaired or are often located in places that do not take part in active replication, transcription or regulation and thus are, apparently, silent. Only rarely do mutation have a detrimental effect on an individual's health. Still, silent mutations contribute to changing the individual's genome and since these mutations are passed on to the next generation, over time a certain population will contain a genomic profile that varies from other populations having been exposed to different mutation events and selection pressures[219]. This gives rise to designations such as single nucleotide polymorphisms (SNPs), insertion/deletions (indels) and copy number variations (CNVs), collectively known as genetic variation, which has been a major research focus over the last decades. Starting with the multinational research project The Human Genome Project, that successfully sequenced the entire human genome[220] several other studies have since followed; the UK10K aimed to map the genetic variation in 10,000 individuals from the United Kingdom with special focus on disease-causing variants[221]; the African Genome Variation Project aimed at mapping the genetic diversity of 1,481 individuals in sub-Saharan Africa to gather information on this population group that has played an important role in ancient human history and continues to be under selection pressure[222]; and the 1000 Genomes Project, a comprehensive study of global population variation in the full genomes of 2,504 individuals from five global population (African, Ad-mixed American, European, East Asian and South Asian), divided into 26 subpopulations (such as African American, Finns, Japanese, Peruvians, Southern Han Chinese *etc.*)[223]. Additional noteworthy databases are the Encyclopedia for DNA Elements[212] focused on transcription factors binding sites in the genome as well as various types of chromatin modelling affecting transcription; the Exome Aggregation Consortium focused on analysing whole exome data from 60,706 individuals[224], and dbSNP which collects and curates the continuous stream of new sequence variants being reported (currently contains more than 89 million validated human SNPs)[225]. The last few years have seen the scope of genomics projects significantly expanded with several planned projects aiming at sequencing individuals in the millions; the Qatar Genome, AstraZeneca's multicentre collaboration (genomes of up to two million individuals to be analysed in the United Kingdom, Finland and the United States), the 100,000 Genomes Project in the United Kingdom, and the Precision Medicine Initiative (one million individuals) in the U.S.

Genome-wide association studies and expression quantitative trait loci

Sequence variants such as SNPs and indels can only be described in relation to a given population. It is a locus (a single-base pair substitution for the former, and an insertion or deletion of multiple base pairs for the latter) in the genetic code, where a group of individuals varies from the rest of the population. These variants may confer a selective advantage in some cases, but are also one of the prime causes for diseases and disorders. With the completed Human Genome Project came the prospect of precisely pinpointing a disease to a gene region. However, the found sequence variants were in the millions, too great a number to test in large groups of patients. This was overcome by the discovery of haplotype blocks: During homologous recombination the DNA is broken at non-random sites leading to larger chunks of DNA ("haplotype blocks") remaining intact through several generations[226]. These haplotype blocks will thus contain several sequence variants that are said to be in linkage disequilibrium (LD) and inherited together. Mapping the entire genome's haplotype blocks was done by the HapMap project[227]. Combined with sequence variants collected in databases such as dbSNP[225] this facilitated the identification of index or tag SNPs representing only a single sequence variant but by proxy including the other variants in the haplotype block as well, significantly reducing the number of SNPs to investigate. From there the genome-wide association study (GWAS) was quickly developed, which combined index SNP analysis with case-control population cohorts, thus determining potential risk alleles, *i.e.* specific variants commonly associated with various diseases. The largest of the initial GWA studies investigated seven diseases in 14,000 cases and 3,000 controls and found several SNPs and risk loci[228]; many have since followed, most relevant for this thesis is a study that found 75 SNPs (one of them rs1175550) that influence various red blood cell parameters[217]. An approach similar to GWAS, but using gene expression profiles instead of case-control studies, yields expression quantitative trait loci (eQTLs)[229] and combining the two to find overlapping reported SNPs greatly enhances the ability to characterise true association[230]. This is achieved as GWAS looks at the phenotypic results of the variation on the organism as a whole, while eQTLs identifies the affected gene and thus gives a genetic background for the observed phenotype.

Trans-ancestry association studies

GWAS and eQTL studies were major breakthroughs and facilitated the discovery of disease-associated risk loci (*i.e.* causal variants) in numerous diseases including haematological disorders such as marginal zone lymphoma[231], acute lymphoblastic leukaemia[232] and multiple myeloma[44]. Yet, the techniques are inherently confined to discovering haplotype blocks, which makes it difficult to determine a specific causal variant locus if the LD in a region is strong. Likewise, due to the LD, a particular discovered index SNP may not be the disease-causing variant at all,

rather a proxy SNP for one of the other variants in the block that actually drives the trait. This is where trans-ancestry (or -ethnic) association studies come in, which can be seen as a way to mix GWAS/eQTLs with population genetics. Through generations of genetic code rearrangements and human migrations the unique genetic code of different populations spread across the globe and they continue to diverge (and occasionally converge) to this day. This means that haplotype blocks (and thus strength of LD) between sequence variants differ among population groups, and this can be exploited by combining GWAS or eQTL studies with population-based genetic variation, such as that reported by the 1000 Genomes Project. The last few years have seen an increase in high-throughput trans-ancestry association studies[233-236], and a very recent study identified seven loci for erythrocyte traits (including MCHC) using indviduals from the African, European and South Asian superpopulations[237].

Genetic variation in blood group systems

A blood group antigen may arise when a sequence variation in the genetic code creates an immunogenic epitope that sets one individual apart from the general population. Such variation can be as small as a single or a few SNPs, as is the case for the *ABO* gene[165], or it can be indels of longer stretches of DNA in the genetic code, *e.g.* the 17-bp deletion in the coding region of *SMIM1*[171,209,210]. The ISBT keeps track of and categorises all blood groups systems and antigens as they are discovered, and they are thoroughly described in the *Blood Group Antigen Factsbook*[238]. Lastly, blood group systems and population genetics are obviously connected, and a recently published online database annotates sequence variants in blood group genes with population allele frequencies extracted from the 1000 Genomes catalogue[239].

The present investigation

Aims

This thesis investigated the gene *SMIM1,* which at the start of the thesis had only just been identified as having a potential role in the expression of the high incidence antigen Vel.

The overall aims were to:

- Confirm the putative erythroid gene *SMIM1* as the gene underlying the Vel blood group system.
- Identify and characterise sequence variants in *SMIM1* that modulate the expression of *SMIM1* and the Vel blood group antigen.
- Investigate genotype-phenotype association in archived Vel-negative samples collected before the genetic background for Vel was discovered.

Summary of studies

Study I

Homozygosity for a null allele of *SMIM1* defines the Vel-negative blood group phenotype

This study, together with two other studies published almost at the same time, determined the molecular background of the Vel antigen. The three studies used different approaches, but all reached the same conclusion: that a mutation in the gene *SMIM1* was the prime cause of the lack of expression of the Vel antigen.

Our approach consisted of microarray-based SNP profiling of the genomes of 20 Vel– individuals from two Swedish families and comparing them to their Vel+ relatives. Comparing 5 Vel– individuals to their Vel+ siblings we narrowed the potentially causal SNPs from more than two million to 8,780 SNPs that segregated with the Vel– phenotype. After a second filtering combining the remaining 15 Vel– individuals with minor alleles frequencies in individuals of European descent in the 1000 Genomes catalogue[223], we found a region on chromosome 1p36, which was unique for all Vel– tested. This region contained five genes, where one in particular caught our attention, *SMIM1*, or, as it was known then, *LOC388588*. Protein structure and function algorithms predicted it to be a type I TMP and an *in silico* analysis of the gene network neighbourhood surrounding *SMIM1* revealed the nearest genes to be largely erythroid, *e.g. KLF1*, *RHD* and *KEL*. Furthermore, *SMIM1* was found highly expressed in erythroleukaemia cell lines (Cancer Cell Line Encyclopedia[240]) as well as upregulated in human CD34+ cells cultured towards erythroid differentiation[241]. Sequencing of *SMIM1* revealed that all 20 Vel– individuals were homozygous for a 17-bp deletion whereas all Vel+ had the consensus sequence. This deletion is located within the coding sequence and introduces a frame-shift that skips the stop codon. We took the following steps in our efforts to determine if SMIM1 was indeed responsible for the long sought after Vel antigen:

- 5’ and 3’ Rapid Amplification of cDNA Ends (RACE) experiments found the consensus SMIM1 mRNA sequence when using total RNA from Vel+ individuals, and no detectable sequence with RNA from Vel– individuals.
- Western blot with human anti-Vel and rabbit polyclonal antibodies raised against the presumed extracellular domain of SMIM1 detected bands of ~20kDa on Vel+ RBC membranes, but nothing on Vel– RBC membranes.

- Flow cytometry with a human anti-Vel gave a strong reaction with samples homozygous for wild type *SMIM1*, less so for heterozygous samples and virtually no reaction for samples homozygous for the *SMIM1* 17-bp deletion.
- Overexpression of SMIM1 wild type in the erythroleukaemia cell line K562 greatly increased human anti-Vel reactivity assayed by flow cytometry, whereas cells expressing the deletion mutant did not. This was also seen in Western blots of the cell membranes using the rabbit anti-SMIM1 antibody.

Conclusions

On top of the conclusions already made, the above steps determined that:

- *SMIM1* encodes an erythroid single-pass transmembrane protein with a GXXXG homodimer domain and a 17-bp deletion causing a frameshift mutation exclusively present in Vel– individuals.
- No translatable *SMIM1* mRNA is created in Vel– individuals.
- Human anti-Vel and rabbit anti-SMIM1 both recognise the same protein on Vel+ RBC membranes; they do not detect anything on Vel– RBCs.
- *SMIM1* deletion zygosity affects the expression of the Vel antigen on the RBC surface (*i.e.* evidence of a gene dosage effect).
- The *SMIM1* open reading frame encodes a protein that reacts strongly with human anti-Vel antibody and the deletion abolishes this reactivity.

Based on this, *SMIM1* was established to be the gene responsible for the elusive Vel blood group antigen, and that the Vel– phenotype is caused by a 17-bp deletion resulting in abolished *SMIM1* expression. This determined the molecular background for the Vel antigen, prompting it to be recognised as the Vel blood group system, and be given number 034 by the ISBT.

Study II

SMIM1 variants rs1175550 and rs143702418 independently modulate Vel blood group antigen expression

Having found *SMIM1* to be the gene responsible for the Vel blood group system, we then set out to examine it in detail. A genome-wide search for sequence variants affecting binding of eTFs located the SNP rs1175550 (A>G) in intron 2 of *SMIM1*. This SNP changes a non-canonical binding site of GATA-1, GATT, to GGTT, and could potentially influence binding of this prime erythroid transcription factor to *SMIM1*. rs1175550 was reported to influence mean corpuscular haemoglobin content[217] and was later shown to have a role in Vel expression[210,213], however no functional explanation for this was given.

Examining 150 samples from Swedish Vel+ donors we:

- Confirmed that rs1175550G is associated with high expression as determined by mRNA expression levels, human anti-Vel reactivity and SMIM1 protein expression on RBC membranes.
- Discovered a new sequence variant, rs143702418 – a trinucleotide insertion (C>CGCA), located 97 bp upstream of rs1175550.

The two variants were in near-perfect linkage disequilibrium (LD) making it impossible to pinpoint the causal variant. By consulting the 1000 Genomes catalogue, we realised that the two SNPs were frequently unlinked in individuals of African descent. Thus, repeating the above experiments in a collection of 202 African American samples, we found:

- rs1175550G still correlated with high expression.
- rs143702418 was more mixed and no direct correlation was found.
- Regression analysis conditioned on rs1175550 showed an independent effect with rs143702418C being the high expressing allele.

We expanded our analysis with functional studies of the two variants. We created four luciferase constructs corresponding to the four possible combination alleles existing between the two bi-allelic variants and confirmed that the rs1175550G and rs143702418C are the most transcriptionally active variant alleles. In gel shift assays with antibody supershifts and with probes spanning either allele of the variants we found that GATA-1 and Gfi-1B bind near the rs1175550 variant in both alleles, whereas a complex containing TAL1 preferentially binds only to the rs1175550G allele. Assays for rs143702418 showed differential binding patterns, but no antibody supershifts.

Conclusions:

Analysing 150 Swedish and 202 African American blood samples, we found that:

- rs1175550 and rs143702418 independently influence *SMIM1* and Vel antigen expression, partly explaining the large variation in anti-Vel reactivity routinely observed for Vel+ individuals.
- TAL1 is likely to mediate the increased expression as it binds exclusively to the high-expressing rs1175550G allele.
- GATA-1 and Gfi-1B are mainly associated with the low-expressing rs1175550A allele.
- Using population genetics to deconvolve tightly linked variants we demonstrated a small-scale trans-ancestry association analysis, which is a promising new tool for molecular geneticists.

Study III

Serologic And Molecular Studies Of The Vel– Phenotype In A Multi-ethnic Population

With the molecular background determined, we examined 40 historic Vel– blood samples characterised as Vel:1,–2 and Vel:–1,–2[208] for the 17-bp deletion. We found:

- 22 individuals originally classified as Vel:–1,–2 were homozygous for the deletion.
- 14 individuals originally classified as Vel:1,–2 were heterozygous for the deletion.
- One sample was heterozygous for the c.152T>G mutation in the coding sequence.
- The remaining three samples did not carry the deletion or the point mutation.

We sequenced the entire *SMIM1* gene region in the three outlying samples but found no apparent sequence variants to explain the Vel– phenotype. We then performed whole exome sequencing and preliminary results indicate potentially interesting variants in *KLF1*, where mutations have previously been linked to the rare In(Lu) phenotype of the Lutheran blood group system[242].

Conclusions

Investigating the archived Vel– samples with the current knowledge of the Vel blood group system, we found:

- The historic classification Vel:–1,–2 corresponded 100% to samples homozygous for the 17-bp deletion (*i.e.* true Vel– as we know now).
- The historic classification Vel:1,–2 is indicative of heterozygous carriers of the mutation, which corresponds to very weak Vel antigen expression.
- Likewise, the one sample carrying the point mutation causing an amino acid change (p.Met51Arg) also had very weakly expressed Vel antigen.
- No apparent reason for the three outlying samples has been found; whole exome sequencing data is currently being analysed.
- The 17-bp deletion is present in individuals of African, Hispanic and European descent indicating the mutation is globally distributed and happened early in human history.

Study IV

Robust isolation of malignant plasma cells in multiple myeloma

This study focused on biomarker shedding in multiple myeloma plasma cells. Membrane protein CD138 is a well-established marker[51,52], yet it has one major disadvantage: It quickly disappears from plasma cells *ex vivo*[52,53]. This makes it unsuitable for delayed analysis or when working with biobank or frozen samples. We therefore sat out to discover a more stable MMPC marker.

We retrieved gene expression profiles of 1,285 MM samples from microarray data in the NCBI Gene Expression Omnibus and compared expression of individual genes to that observed in 3,164 non-MM haematological malignancies to find differentially expressed genes specific for MM. This *in silico* search was further narrowed by gene ontology terms, literature and database crossmatches, potential cell surface expression, and availability of FACS-suitable antibodies. This yielded seven potential targets, CD269 (NFRSF17/BCMA), CD319 (SLAMF7/CS1), GPRC5D, FKBP11, CD208 (LAMP3), ITGA8, and CD307e (FCRL5).

We then did FACS analyses of bone marrow cells either freshly sampled, stored at 8°C or frozen, using antibodies against these seven targets plus CD138 and found:

- CD319, CD269 and CD307e consistently sorted distinct cell populations in all three experimental conditions. The remaining four did not.
- These populations were confirmed to be MMPCs by morphology, clonal excess, and DNA copy number microarray (chromosomal abnormalities).
- CD138 was confirmed to be suitable for fresh samples, yet ineffective in samples that had been stored or frozen.

Conclusions

Narrowing a large list of potential MMPC-specific proteins to seven putative MM-specific surface markers, we found that:

- CD269 and CD319 are robust markers suitable for isolating MMPCs from bone marrow cells, regardless of storage conditions.
- These are solid replacements for CD138 as MMPC markers in FACS analysis.
- Robust isolation of MMPCs with handling methods suitable in a clinical setting enables thorough analysis and characterisation of MMPCs.

Methodology

A list of methods utilised in this thesis is presented in **Table 3**. No new methodology was developed, but a gene-specific PCR capable of quickly detecting the *SMIM1* 17-bp deletion (and thus Vel status) in a single PCR run was designed, as well as a new marker (CD319) for immunophenotyping MMPCs in FACS analysis was characterised.

Table 3. Methods employed in this thesis.

Name	Study	Brief description of usage
5' and 3' RACE	I	Determine the start and end point of *SMIM1* mRNA from Vel+ and Vel– individuals.
Cell culture	I, II	Maintain and passage K562 and HEL cells.
Cloning and transfection	I, II	Insert the SMIM1 coding sequence or a DNA fragment into GFP/luciferase vectors, then transfect into K562 or HEL cells before flow cytometry analysis/luciferase assays.
Flow cytometry	I, II	Analysis of transfected K562 cells or human RBCs coated with a human anti-Vel antibody.
PCR	I, II, III	Wide usage to amplify various *SMIM1* regions.
Gene-specific PCR	I	Primer design and parameter optimisation of a PCR able to determine *SMIM1* 17-bp deletion zygosity.
Sanger sequencing	I, II, III	Identify genetic variants in *SMIM1* intron 2 or full gene; *in silico* electropherogram analysis.
Total RNA extraction	II	Phenol-chloroform extraction from whole blood samples of Swedish and African-American origin.
RT-qPCR	II	cDNA synthesis from total RNA samples, then quantitative PCR to examine *SMIM1* expression.
Gel shift assay with antibody supershift	II	Nuclear extracts from K562 and HEL cells, mixed with biotin-labeled oligos corresponding to the found *SMIM1* genetic variants and antibodies recognising select eTFs.
Luciferase assays	II	Luciferase constructs transfected into K562 or HEL cells, then harvested, lysed and analysed by chemiluminescence.
Whole genome amplification	III	Unspecific, unbiased amplification of the full genome of the outlying Vel– samples in preparation for whole exome sequencing.
Bioinformatic analyses	I, II, III	UCSC genome browser, 1000 Genomes database and dbSNP search and filtering, sequencing data, JASPAR TF motif scan, whole exome sequencing data.
FACS	IV	Sort bone marrow cells coated with fluorophor-conjugated antibodies.

Discussion

The Vel blood group system has now been solidly linked to the *SMIM1* gene. With the 17-bp deletion known, we designed a gene-specific PCR that amplify a region of 178 bp spanning the deletion, allowing for quick identification of deletion carriers, and while several studies have reported on various aspects of SMIM1, such as **Study II** of this thesis and the study of the Vel antigen epitope on SMIM1 by Arnaud and co-workers[216], a specific function remains to be found.

Population genetics of Vel/SMIM1

As noted in the introduction, several studies were performed in the decades after the Vel antigen was reported in 1952, most of them on Caucasian population groups, but also two small studies suggesting a much higher incidence of Vel– individuals (1 in ~80) in both Thais[243] and Chilcotin Indians in Canada[244]. The discovery of the 17-bp deletion in *SMIM1* opened up for genetic screens for the Vel antigen and we used our gene-specific PCR on a collection of 520 random blood donors to establish a Vel– incidence in Southern Sweden of 1 in ~1,200[171], which is significantly higher than the global incidence, but this might be explained by the close relation to the hypothesised potential founder population in Northern Sweden described by Cedergren *et al.* in 1976. Similar gene-specific PCRs were employed to establish a Vel– incidence of 1 in ~2,100 individuals in southern Germany[245] and 1 in ~4,700 in a Brazil[246].

The apparent high incidence of Vel– individuals in Thais and Chilcotin Indians could be the result of a founder effect but it should also be noted that Vel status in these early studies were determined by serology with human anti-Vel. There is considerable variation in anti-Vel reactivity even among Vel+ individuals and it has been shown that individuals heterozygous for the *SMIM1* 17-bp deletion always show very low reactivity, even if the Vel antigen is present on their RBCs. Thus, the unusually high frequency of Vel– in Thais and Indians can possibly be explained by a high number of Vel+^{weak} individuals wrongfully typed as true Vel–. In support of this hypothesis, at least in a Southeast Asian population, no 17-bp deletion carriers were found in a study of 325 individuals of Malaysian and Indian ancestry (personal communication; Veera Sekaran Nadarajan). Allele frequency for the low *SMIM1*-expressing rs1175550A allele was 97,5% in these samples, which is in line with the allele frequencies of 99,8% in the East Asian

superpopulation in the 1000 Genomes catalogue. This could mean that there is a high frequency of Vel+^{weak} individuals in the Asian population, leading to a potentially higher amount of individuals falsely identified as Vel– in serotyping screens. There are unfortunately no Native American populations in the 1000 Genomes catalogue, nor have any *SMIM1* genetic screens been performed, so we cannot know if the high incidence here is due to true Vel– or Vel+^{weak} individuals.

Unfortunately, the 17-bp deletion is not called in the 1000 Genomes catalogue presumably due low sequencing coverage in the region and that indels are inherently hard to impute from SNP microarray data. This precludes a thorough analysis of the global distribution of the Vel-defining sequence variant, yet we were able to find the deletion in a few samples of Hispanic, Caucasian and African descent in **Study III**, and combined with the studies of Asian and Brazilian population groups, there is ample evidence to assume that the mutation is distributed worldwide, and that it thus presumably arose early in human history. Evidence for an ancestral allele was also presented by Ballif and coworkers who found the SNP rs71634364, located 240 bp upstream of the 17-bp deletion, to be in LD with the deletion in 67 out of 68 Vel– cases[209]. And we and the third group that discovered *SMIM1* have also observed that rs1175550A and the 17-bp deletion almost always exist on the same allele, pointing at potential LD between the two (unpublished data and personal communication with Ellen van der Schoot).

Vel antigen presentation and SMIM1 function

As described earlier, SMIM1 was recently characterised as a type II TMP with a short extracellular C-terminal and the final three amino acid residues crucial for proper Vel antigen presentation[216], yet questions remain regarding how the antigen is presented on the RBC surface and what the specific function of SMIM1 is. The find directly opposes our conclusion in **Study I** that SMIM1 is a type I TMP and in the following, I will try to reconcile these two seemingly contradictory characterisations, and argue that they are not necessarily mutually exclusive. But first, a bit about antibody specificity and antigen presentation.

A recent study reported the production of a monoclonal anti-Vel antibody (SpG213Dc) raised from plasma anti-Vel found in a Vel– individual and propagated in a human myeloma cell line[247]. This detects a 32kDa protein in Western blots under non-reducing conditions and a 20kDa when reduced. This is in line with results presented by Ballif *et al.*[209] and in our own **Study I** with anti-Vel from immunised Vel– individuals. Since the predicted molecular weight of SMIM1 is 8.75kDa it can be speculated that the 20kDa band found under reducing conditions might be a homodimer probably held together by the reported transmembrane GXXXG motif[171], and protected from reducing agents by the RBC

membrane, as is the case for GPA, which forms detergent-resistant transmembrane dimers through its GXXXG motif[215,248]. The size difference may also be due to posttranslational modifications, yet we found SMIM1 to not be O-glycosylated or sialylated. The SMIM1 32kDa band observed in native conditions can be thought to be a multimeric complex consisting of homodimeric SMIM1 and one or more unknown SMIM1 partner(s), interacting through domains located outside the transmembrane region. Indeed, SMIM1 contains two cysteine residues N-terminal of this region (located in the cytoplasm if SMIM1 is a type II TMP) capable of forming disulphide bonds, however these cysteines were shown to be expendable for proper human anti-Vel reactivity by Arnaud *et al.*[216]. Another cysteine residue (Cys77) is present as the penultimate residue in the C-terminal, and this might very well be crucial for direct interaction with one or more proteins to form the Vel antigen epitope, as absence of this and the two lysine amino acids flanking it (referred to as the KCK motif in the following) abolished anti-Vel reactivity entirely[216]. In **Study I**, we raised an anti-SMIM1 antibody by immunising a rabbit with a polypeptide corresponding to a region in the N-terminal of SMIM1[171]. In contrast to the bands detected above, we observed two bands of ~9-10 and ~20kDa when examining solubilised RBC membranes under reducing conditions. And when we treated whole RBCs with the protease chymotrypsin, the bands disappeared, indicating that the anti-SMIM1 epitope presumably sits on the RBC surface. This find seemingly contradicts SMIM1 being a type II TMP as anti-SMIM1 would then have been raised against the intracellular part.

So the questions remain: What do the different antibodies recognise, and what is the correct orientation of SMIM1 in the RBC membrane. Regarding the first question; our anti-SMIM1 recognises a linear sequence of amino acids in SMIM1, but not necessarily the Vel antigen epitope (indeed, despite multiple attempts, we were unable to haemagglutinate RBCs using anti-SMIM1[171]), while the human anti-Vel and the SpG213Dc antibodies obviously recognise the Vel antigen epitope (since it is pulled directly from plasma from immunised Vel– individuals). But since antigen epitopes can be dependent on conformation and protein fold as much as amino acid sequence, it follows that anti-Vel does not necessarily bind specifically to a linear stretch of amino acids in the SMIM1 protein, but rather to whatever native conformation and/or complex SMIM1 exists in in the RBC membrane. The anti-Vel can even be thought to specifically recognise another protein entirely, but a protein (complex) that is critically dependent on a SMIM1 anchor on the RBC membrane for proper antigen presentation. Such membrane multicomplexes are not a new concept for blood group systems, where the proteins of several systems come together in one massive macrocomplex. For example, GPA and GPB (MNS blood group system), RhAG (RHAG), RHD and RHCE (Rh), LW (Landsteiner-Wiener) and CD47 all form complexes that are anchored to the cell membrane through Band 3 (Duffy), Protein 4.2 and Ankyrin. Absence

of Band 3 results in significant reduction of MNS, Rh and RhAG antigens on the RBC surface[249]. If SMIM1 is indeed a scaffold/anchor protein, it follows that absence of SMIM1 would mean no antigen epitope formation and display, leading to a Vel– phenotype.

The question of SMIM1 membrane orientation in the RBC membrane is not straightforward to answer, based on the few studies published so far. Either antibody (anti-SMIM1 and anti-Vel) supposedly recognise epitopes on either side of the transmembrane region and both antibodies presumably recognise something on the extracellular surface of RBCs. And both detect proteins that can be deduced to be SMIM1 or complexes that SMIM1 is a part of. Interestingly, papain treatment resulted in a reduction of ~4kDa with SpG213Dc anti-Vel[247], and decreased band intensity with anti-SMIM1[171], meaning that either conformation complex is sensitive to protease degradation of some of its components.

In summary, there is evidence that SMIM1 exists as both a type I and a type II TMP in the RBC membrane. As a type I monomer and (homo-)dimer, but not necessarily presenting the Vel antigen and as a type II with a short extracellular C-terminal and the final three amino acids important for antigen presentation. Future research should focus on finding SMIM1 binding partners, which would ascertain if either of these orientations can be excluded and shed light on the actual nature of the Vel antigen presentation.

On a final note, even various online protein sequence analysis tools predict the membrane topology differently. All tools predicted a single-pass TMP without signal peptide, but eight out of ten (*e.g.* CCTOP[250] and Philius[251]) predicted SMIM1 to have the N-terminal on the cytoplasmic side (*i.e.* Type II TMP), while Phobius[252] and TMHMM[253] predicted the opposite orientation. The lack of a predicted signal peptide conflicts with classifying SMIM1 as a type I TMP, but there are examples of extracellular N-terminal proteins with reverse signal anchors that function as a signal peptide, but are not cleaved when the proteins are inserted in the ER membrane (these are also referred to as type III TMPs)[254].

SMIM1 as an invasion receptor for malaria?

In **Study I** we hypothesise that SMIM1 may be a receptor for *P. falciparum* and there is some evidence to support this. First, SMIM1's structure as a single-pass TMP with a GXXXG motif is also characteristic for GPA, which, as noted in the introduction, has been shown to be a receptor for *P. falciparum* infection. Second, we found SMIM1 to not be sialylated, yet resistant to trypsin and very sensitive to chymotrypsin treatment, three characteristics of an unknown erythroid receptor (receptor Z[255]) for the *P. falciparum* ligand *P. falciparum* reticulocyte binding homologue 2b (PfRH2b) which has been described[256], but not yet linked to a certain protein in the RBC membrane. And third, SMIM1 was found to be

phosphorylated on Ser22 and Ser28 (see **Figure 4** on page 37) in a study of the global phospho-proteome in RBCs infected with *P. falciparum*, along with 25 other human proteins, several of which are well-known erythroid proteins, such as Band 3, Glycophorin C, Protein 4.1, Kell, CD44 and Ankyrin-1[257]. It seems likely that SMIM1 plays some form of role in malaria pathogenesis, yet at this point, more research is needed.

Sequence variation in SMIM1

In **Study II,** we found that both rs1175550 and rs143702418 in *SMIM1* intron 2 influence *SMIM1* and Vel expression, partially explaining the large variation in anti-Vel reactivity observed among Vel+ individuals. We initially performed a scan of sequence variants disrupting eTF binding sites, thus potentially altering expression of the target gene and found rs1175550 due to its position in a non-canonical binding site for GATA-1; GATT, with the minor allele changing this to motif GGTT. rs1175550G is the effector allele responsible for high Vel expression[210,213,214]. But how is this regulation achieved? Our luciferase assays in **Study II** showed that the region surrounding the variant is certainly capable of attracting factors that affect transcription, and reports have shown that GATA-1 bound to the GATT motif can lead to transcriptional activation in testicular cells[258]. However when GATA-1 is associated with Gfi-1B (whose binding motif is AATC, the reverse-complement of GATT), it leads to repression in erythroid cells of the genes *Gfi-1B*[109] (negative feedback loop) and *Bcl-x_L*[108]. So maybe GATA-1/Gfi-1B are responsible for repressing SMIM1 transcription when bound to rs1175550A? Our gel shift assays were inconclusive, with GATA-1 not binding directly to GATT, however Gfi-1B was shown to not bind a complex specific for rs1175550G. However, and more intriguingly, we found TAL1 to specifically bind to this complex, consistent with a very recent report showing that TAL1 binds near rs1175550 and preferentially to the G allele[259]. We thus assumed that TAL1 is a major factor governing the high expression of rs1175550G. Again, how is this achieved? Obviously the A>G shift disrupts Gfi-1B binding and potentially inhibits its repressor function. But what about TAL1 binding sites? Shortly upstream of rs1175550 is a variant of the TAL1/GATA-1 composite motif described earlier[98] – (CTG(n_8)GATT) – yet this would suggest that the rs1175550A should be the high-expressing allele, so presumably this motif is not used. Another TAL1-like motif is also present, CAGCCT, (degenerate from the canonical CAGNTG motif) and might be what TAL1 binds to in our gel shift assays. However the most compelling motif in the region is actually found on the antisense strand where an exact match for the composite motif can be found, starting six basepairs upstream from rs1175550 (motif on the sense strand is TTATCA(n_7)CAG, **Figure 5**). Although GATA-1 binding to GATA/GATT motifs seems fairly redundant for transcriptional regulation[86,87], TAL1 and the

composite motif is a good predictor for transcriptional activation[98], so it is very plausible this motif is used *in vivo*. This theory can easily be tested by running gel shift assays with probes spanning the composite motif and antibody supershifts with anti-GATA-1, -Gfi-1B and -TAL1. Binding specificity can then be examined by using probes with the specific binding sites mutated.

The question remains why rs1175550 is an eQTL in this regard, if TAL1 binding is not directly affected by it. I would argue that since the CAG and GATT motifs are only separated by five basepairs, the occupancy could very well be governed by steric hindrance; for rs1175550A the GATT (AATC) motif is occupied by GATA-1 and/or Gfi-1B while the substitution to rs1175550G disrupts this binding and allows TAL1 to bind immediately upstream, displacing or blocking Gfi-1B (**Figure 5**).

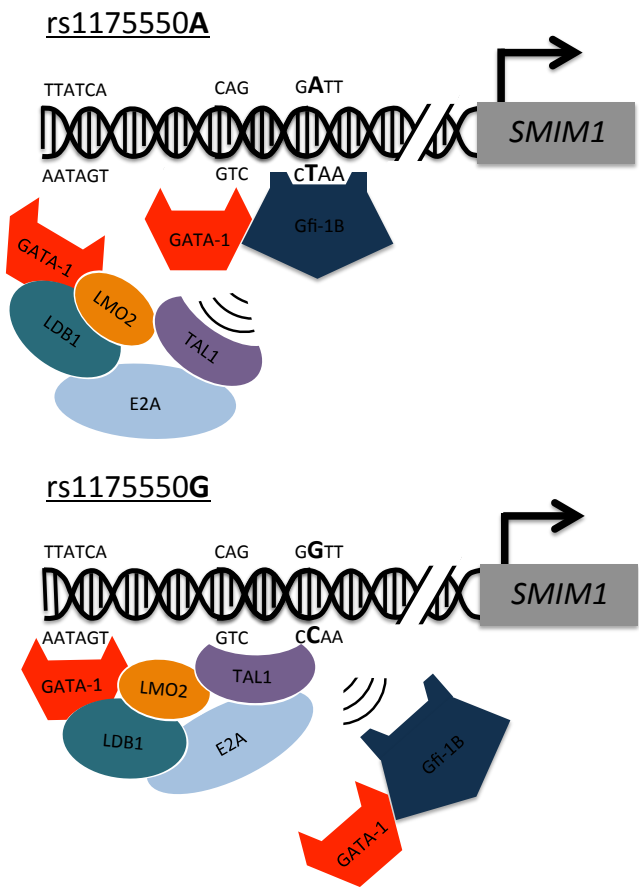


Figure 5 rs1175550 alleles and proposed bound erythroid transcription factor complexes. Expression variation is hypothesised to be governed by physical blocking of one eTF complex by another. The rs1175550A allele allows binding of the transcriptional repressor complex consisting of Gfi-1B and GATA-1 (and possibly the CoREST multicomplex, not shown). rs1175550G disrupts this site, allowing TAL1 and the transcriptional activation pentameric complex to bind to the composite TAL1-GATA motif only six basepairs upstream. Transcriptional regulation resulting in high or low expression of *SMIM1* and the Vel antigen thus becomes a matter of steric hindrance by two opposite-acting complexes.

We suspected the other variant in **Study II**, rs143702418 (C>CGCA) to also be a potential eQTL, since it is located within a binding motif for one of the other major eTFs, KLF1. The variant changes the core CACCC sequence to CACGCACC and even though this resembles the extended motif CCNCNCCCN (see **Table 1**, page 23), we found that the tri-nucleotide insertion correlates with lower *SMIM1* expression and transcriptional activity, so presumably the motif is altered sufficiently to disrupt KLF1 binding. Unfortunately, we were unable to show a specific binding of KLF1 to the probes spanning the variant in our gel shift assays, but this may not necessarily rule out KLF1-binding entirely. Supershifts in gel shift assays are technically challenging and often require an antibody specifically tested for this method, which was not the case for the antibody we used. We did see an altered binding profile for the two probes (one for each variant allele) without antibody supershifts, indicating that different proteins bind to either probe. We did not find solid evidence of other erythroid proteins (*i.e.* GATA-1, TAL1, Gfi-1B or ZBTB7A, these antibodies worked well in the gel shift assays for rs1175550) binding at or near the rs143702418. This is not surprising as the probes do not contain any binding motifs for the antibodies being tested (**Table 1**), perhaps with the exception of ZBTB7A's alternative GC-rich motifs reported by Maeda and coworkers[129,130]. If KLF1 does not bind to rs143702418, which protein could then be responsible for the transcriptional regulation? The KLF-related ubiquitous transcription factor Specificity protein 1 (Sp1) binds to GC-rich regions and particularly CCCCNCCCCC (JASPAR motif) and this may very well also play a role in *SMIM1* expression, as Sp1 has been shown to interact and recruit GATA-1 to DNA in the absence of GATA sites[260]. In summary, while we were unable to detect a supershift with KLF1, this does not necessarily rule out binding in this region. This could be investigated by optimising antibody binding conditions in the gel shift assay or by performing chromatin immunoprecipitation (ChIP) of erythroleukaemia cells.

KLF1 and TAL1/GATA-1 binding sites are often found within 1kb of each other[115] and the 96 bp between rs143702418 and rs1175550 fits well with this prediction. This distance could be sufficient for rs143702418-bound KLF1 to loop the DNA and interact with the pentameric complex bound to rs1175550 to bring the *SMIM1* gene into the transcription factory, as described in the introduction[121,138]. This could be tested by mixing the gel shift assay probes with nuclear extracts and then co-immunoprecipitate with KLF1 and one of the components of the pentameric complex.

Linkage disequilibrium and trans-ancestry association analyses

We were surprised to find the two *SMIM1* variants tightly linked in our Swedish samples and quickly realised that with the strong LD, separating their effects would be statistically impossible. We originally planned to separate the two variants in a cell line using CRISPR-Cas9 genome editing with homology-directed repair, but soon realised that the two variants had significantly different allele frequencies in the African superpopulation in the 1000 Genomes catalogue. In other words, the LD was most likely not as strong as in Swedes. Therefore, we analysed a set of African American samples, and were able to separate the effects of the two variants using nothing but the naturally occurring population variation[214].

Genetic variation profile for SMIM1 full gene

In **Study III,** we investigated three outlying Vel samples not carrying the 17-bp deletion, but originally typed to be Vel–. Knowledge of the three samples is somewhat limited (one of them was archived in 1976), but we know that they were all Caucasian and that at least one of them presented with Vel antibodies in the plasma (*i.e.* probably a true Vel– individual).

We sequenced the entire SMIM1 gene plus up- and downstream regions, 6,000 bp, in all three outlying samples and a Vel+ control, and even with a small sample size and DNA region we got an interesting glimpse of the substantial interindividual variation that exist in a population. The region contains 37 sequence variants with a minor allele frequency above 1% in dbSNP version 147[225] and in 13 (~35%) of these, at least one sample had a different genotype than the others. Unfortunately, no variant was found to be shared for all three Vel– samples and not the control, yet we can not necessarily assume that their Vel status is caused by the same sequence variant.

There are a few known eQTLs that we already know causes low to no Vel antigen expression. The first is obviously the 17-bp deletion, but all three samples were homozygous wild type. Then there is the rs1175550 and rs143702418 described in **Study II**. All three samples did not contain the trinucleotide insertion (rs143702418) and were homozygous rs1175550A. The latter would confer low antigen expression, which could explain a potential mistyping, however at least one of the individuals is supposed to be true Vel–, and not Vel+^{weak}. And finally, all three samples were also homozygous wild type for the triallelic variant c.152T>G/A, where the two minor alleles results in lowered *SMIM1* expression[210].

We did find a 97-bp deletion compared to the curated RefSeq sequence, only 200 bp upstream of the transcription start site, which was present in all four samples. Interestingly, the sequence variant rs367810010 (minor allele is a deletion of 16 bp GCCCCGCCCCCCTCTC, minor allele frequency in Europeans: 23%) is located

at the very start of this deletion. The region is quite GC-rich and contains several repeated sequences, so we cannot rule out either sequencing or annotation errors in the RefSeq or our sequencing reaction, nor can we say if the rs367810010 is actually longer than the 16 bp reported by dbSNP. Unfortunately, low sequence coverage in the 1000 Genomes catalogue prevented us from examining if this is a variant calling artefact and if there are samples in this database that also contains the 97-bp deletion instead of only 16 bp. Some form of correlation to the variant can be assumed though, it seems too coincidental to find two deletion variants starting at the same basepair, but of varying length. Interestingly, the deleted stretch contains several binding motifs for Sp1 (and possibly KLF1), and given it's proximity to the transcription start site, this could potentially have functional relevance for *SMIM1* expression (as discussed above). It would not explain the lack of Vel antigen expression in the three outlying samples, since the Vel+ control was also homozygous for the deletion, but it is an intriguing observation nevertheless.

The question remains of what caused these samples to type as Vel–. We found no cis-eQTLs, so the answer must lie in the remaining genome. The samples come from healthy individuals, so we do not expect any detrimental mutation(s) in common TFs or genes, rather a (heterozygous) mutation in any of the eTFs or other erythroid genes. We are currently analysing data from whole exome sequencing performed on the three outlying samples and a Vel– control. We have found potential variants in erythroid genes such as *GFI1B*, *RB1*, *YPEL4*, *ZBTB7A* and *KLF1*. The *KLF1* mutations are particularly interesting as mutations in this gene are known to be implicated in the rare blood group phenotype In(Lu)[242] and cause reduced expression of *CD44* (Indian blood group)[261] and Kell, Duffy, RhD and other blood group antigens[114,262,263].

Conclusions and Future Perspectives

In the work comprising this PhD thesis, we:

- established *SMIM1* as the gene behind the Vel blood group antigen and devised a gene-specific PCR to detect the Vel– defining 17-bp deletion.
- characterised two sequence variants as determinants of *SMIM1* and Vel antigen expression and found indications that the transcriptional activator TAL1 is involved in transcriptional regulation.
- determined a Vel incidence in southern Sweden and found that the 17-bp deletion is globally distributed and probably arose early in human history.

But several questions still remain: What is SMIM1's functional role in the RBC membrane? Is SMIM1 the Vel antigen epitope, a carrier molecule or a scaffold protein for a membrane complex? Which other factors bind to the two sequence variants and what is the mechanism of regulation? Is SMIM1 the missing chymotrypsin-sensitive *P. falciparum* receptor? And why did the outlying samples type as Vel–?

Future research should focus on characterising the function of SMIM1. It is likely that SMIM1 interacts with proteins in the RBC membrane and determining these binding partners, perhaps by mass spectrometry, will give valuable insight and likely a major hint of SMIM1's function.

Determining the factors binding to the two sequence variants rs143702418 and rs1175550 is also an important task. Their eQTL effects are substantial and it is very possible that the factors would not only shed light on *SMIM1* expression, but on erythroid transcriptional regulation in general.

This prospect is also present for the study of the outlying Vel– samples. Since we found nothing in the *SMIM1* gene to explain the Vel– phenotype, it will be interesting see if our whole exome sequencing data will uncover potential trans-acting eTFs.

This thesis assisted in characterising the molecular background of the Vel antigen, which is a clinically significant antigen, that can cause severe haemolytic transfusion reactions. This enabled the development of molecular Vel blood group

typing in the clinic as well as the prospect of creating a synthetic monoclonal anti-Vel antibody to be used in diagnostics, blood group typing and research. Furthermore this PhD found TAL1 to be potentially involved in the expression of *SMIM1* and the Vel antigen. Lastly, this thesis demonstrated how pre-existing population genetics can be exploited to deconvolve potentially causal sequence variants locked in tight LD, a powerful approach to solve the inherent limitations of GWAS and eQTL studies.

As always in scientific endeavours, solving one long-standing mystery only lead us to dozens more intriguing new questions, puzzles and conundrums, which, in essence, is what science is all about – the most exciting thing is not the destination, but the road you take to get there.

Populærvidenskabelig sammenfatning

Denne Ph.d. afhandling beskriver resultater opnået gennem fire års laboratoriearbejde på Afdelingen for Hæmatologi og Transfusionsmedicin, Medicinsk Fakultet på Lunds Universitet. Den består af tre studier med fokus på den sjældne blodtype Vel. To af studierne er publiceret i de videnskabelige tidsskrifter *Nature Genetics* og *Scientific Reports.*

Mens de fleste kender til ABO og Rh blodtypesystemerne, er det nok de færreste, der ved, at der faktisk findes yderligere 34 blodtypesystemer. ABO-systemet blev beskrevet af den østrigske biolog og læge Karl Landsteiner i starten af 1900-tallet, og med det blev grundstenene til den moderne transfusionsmedicin lagt. Paradigmet, som er gældende for de fleste blodtypesystemer, er kort fortalt, at alle mennesker har visse strukturer, antigener, på overfladen af de røde blodlegemer, som vores immunsystem genkender. Men for nogen individer kan disse strukturer have et lidt andet udseende, eller helt være fraværende. Hvis blodlegemer fra én person bliver blandet med blod fra én, som har et andet antigen, så kan immunsystemet ikke genkende strukturen på de indkommende blodlegemer og vil reagere som om, det var en fremmed mikroorganisme og danne antistoffer imod det fremmede antigen. Dette kan i værste tilfælde forårsage kraftige og dødelige hæmolytiske transfusionsreaktioner.

Blodtypen Vel er ét af de mere sjældne systemer. Hvis man er Vel-positiv, sidder Vel antigenet på overfladen af de røde blodlegemer, hvis ikke, er man Vel-negativ, og kroppens immunforsvar danner antistoffer imod Vel antigenet – disse kaldes anti-Vel. På verdensplan siger man, at 1 ud af 4.000 er Vel-negative, alle andre er Vel-positive. Der er evidens for en højere frekvens af Vel-negative personer i Norden, hvor vi i denne afhandling bl.a. fandt 1 Vel-negativ person ud af 1.200 i Region Skåne.

Vel antigenet blev første gang beskrevet hos en 66-årig kvinde, Mrs. Vel, i 1952 i New York, der fik en hæmolytisk transfusionsreaktion efter en blodtransfusion. I de efterfølgende årtier blev Vel antigenet fundet hos patienter over hele verden; i Sverige, Australien, Thailand, blandt indianere i Canada og afroamerikanere i USA, men den molekylære årsag, det vil sige det gen eller den mutation, der er skyld i at nogen er Vel-negative, forblev i det uvisse.

Da jeg startede på min Ph.d., var forskningsgruppen allerede godt på vej til at have løst det 60 år gamle mysterium om Vels molekylære baggrund. Man havde bl.a. sammenlignet mutationer i hele genomet hos 20 Vel-negative personer med deres Vel-positive familiemedlemmer som kontrolpersoner og derved fundet frem til et område i genomet, hvor der var en ophobning af mutationer kun fundet hos de Vel-negative. I dette område var der fem potentielle gener, som kunne danne baggrund for Vel. Vi undersøgte nu disse gener for diverse typiske karakteristika for gener udtrykt i blodceller og fandt frem til, at genet Small Integral Membrane Protein 1 (*SMIM1*) var den mest oplagte kandidat af de fem. Ved computersimulation blev SMIM1's nærmeste naboer påvist at være andre gener, der koder for proteiner i røde blodlegemer, og ved at sekventere hele *SMIM1's* DNA sekvens fandt vi en mutation, som var tilstede hos alle de Vel-negative og ingen af de Vel-positive. Denne mutation er en deletion af 17 basepar (byggesten) i DNA'et, som bevirker, at SMIM1 proteinet ikke kommer til udtryk i røde blodlegemer hos Vel-negative personer.

Vi gik derfra videre for at forsøge at bevise sammenhængen mellem SMIM1 og Vel antigenet. Vi undersøgte tilstedeværelse af mRNA (et mellemstadie imellem DNA og protein) for SMIM1 og fandt det kun hos Vel-positive individer. Vi fandt ligeledes kun Vel antigenet samt SMIM1 proteinet udtrykt på overfladen af røde blodceller fra Vel-positive personer. Derudover kiggede vi på sammenhængen imellem personer, som havde nul, én eller to kopier af *SMIM1* mutationen (på de såkaldte alleller, som vi har to af for alle gener i genomet), og hvor kraftigt anti-Vel antistoffet reagerede med deres blodceller. Her fandt vi, at dem som ikke havde *SMIM1* deletionen på nogen alleller reagerede kraftigt med anti-Vel; dem, der havde deletionen på den ene allel, reagerede middel, og dem, der havde deletionen på begge alleller, reagerede svagt – altså et tegn på at SMIM1 og Vel antigenet har noget med hinanden at gøre. Sidst, men ikke mindst reagerede anti-Vel antistoffet kraftigt mod et kunstigt fremstillet SMIM1 protein, mens det nærmest ikke reagerede med det samme protein, men som indeholdt deletionen. Med andre ord, når *SMIM1* genet blev udtrykt, opførte anti-Vel sig, som det ville, hvis det blev blandet med blod fra Vel-positive, mens når *SMIM1*-genet blev muteret, opførte anti-Vel sig, som det ville gøre med blod fra Vel-negative.

Disse forsøg fremførte kraftige argumenter for, at *SMIM1* er baggrund for Vel blodtype antigenet, og at mutationen i *SMIM1* bevirker, at man ikke udtrykker Vel antigenet på sine røde blodlegemer, og dermed at man siges at være Vel-negativ.

Samtidig med os havde to andre forskningsgrupper også fundet frem til mutationen i *SMIM1* og påvist sammenhængen mellem SMIM1 og Vel, og dermed kunne Vel officielt godkendes som et blodtypesystem af International Society of Blood Transfusion i 2013. Projekt I i denne afhandling omhandler de omtalte forsøg.

Med Vel blodtypesystemet på plads kiggede vi herefter på et velkendt fænomen i blodlaboratorier: At selv om man er Vel-positiv, så er der meget stor variation i, hvor kraftigt ens røde blodlegemer reagerer med anti-Vel antistoffet. Det kan komme så vidt, at en svagt-reagerende Vel-positiv person fejlagtet bliver bestemt til at være Vel-negativ, hvilket kan få store konsekvenser ved en blodtransfusion. Vi ville derfor undersøge, hvad der kunne være årsagen til denne store variation og mistænkte en region i *SMIM1* genet, hvor der foregår meget såkaldt gen-regulering. I denne region fandt vi en anden mutation, en såkaldt enkeltnukleotidpolymorfi, forkortet SNP fra det engelske single nucleotide polymorphism, hvor DNA-nukleotidet A (et nukleotid er et af DNA's byggesten – to nukleotider danner sammen ét basepar) er skiftet ud med et G i en vis procentdel af befolkningen. Denne SNP, kaldet rs11755550, var tidligere vist at have indflydelse på Vel udtryk, og vi satte os for at undersøge den nærmere. I blodprøver fra 150 svenskere fandt vi, at de individer, der havde G-nukleotidet i deres DNA-sekvens, viste klart højest reaktivitet med anti-Vel, dvs. de udtrykker meget Vel antigen på celleoverfladen. Og vice versa, personer med A-nukleotidet havde lav forekomst af Vel antigen og lav anti-Vel reaktivitet. Vi undersøgte nu DNA regionen omkring rs1175550 og fandt flere bindingssteder, hvor såkaldte transkriptionsfaktorer kan binde. Det er proteiner, som regulerer hvor meget af et givent protein, der bliver udtrykt i cellen. Gennem flere forsøg med binding af transkriptionsfaktorer til sekvensen omkring rs1175550 fandt vi, at særligt én faktor, kaldet TAL1, så ud til at binde specifikt til DNA-regionen, kun når G-nukleotidet og ikke A var til stede. Eftersom TAL1 oftest aktiverer de gener den binder til og derved fremmer produktion af det resulterende protein, så er det vores hypotese, at TAL1 er en af grundene til, at personer med G-nukleotidet i deres DNA sekvens har højere ekspression af SMIM1 og dermed Vel antigen end personer med A-nukleotiet. Disse forsøg udgør Projekt II.

Det sidste studie om Vel blodtypen i denne afhandling omhandlede bl.a. en tidligere metode at klassificere Vel på: I 1968 blev det påvist, at der fandtes to Vel antigener, Vel 1 og Vel 2, og man klassificerede personer, der var svagt reaktive med anti-Vel antistof som Vel;1,–2 (altså antog man at personen havde Vel 1, men manglede Vel 2), mens personer med meget svag til ingen reaktion med anti-Vel var Vel;–1,–2. Med den molekylære baggrund for Vel antigenet kendt, var vores mål for projekt III at klassificere disse gamle prøver ved hjælp af den nye viden. Til det formål fandt vi 40 gamle prøver fra Vel-negative personer i New York Blood Center's blodbank, som alle var blevet Vel blodtypebestemt med det gamle to-antigen system. Vi undersøgte DNA fra prøverne for 17-bp deletionen i SMIM1, hvor tilstedeværelse af deletionen som bekendt er ensbetydende med lavere eller helt fraværende Vel antigen udtryk. Vi fandt frem til, at samtlige prøver, som oprindeligt var blevet bestemt til at være Vel;1,–2, havde én kopi af deletionen i deres DNA. Og for Vel;–1,–2 prøverne, som var vist at have lav til

ingen Vel antigen ekspression, fandtes deletionen på begge alleller, og altså kunne SMIM1 proteinet ikke udtrykkes hos disse personer. Vi fandt med andre ord komplet overensstemmelse imellem mængden af udtrykt Vel antigen og tilstedeværelsen af 17-bp deletionen i SMIM1 på én eller begge alleller. Herudover fandt vi tre prøver, som oprindeligt blev klassificeret som Vel negative, men som ikke indeholdt 17-bp deletionen på nogen af allellerne. Dette er højest usædvanligt, og noget der kun er rapporteret få gange tidligere, og vi er derfor nu i gang med at analysere disse prøver nærmere for at finde en eller flere yderligere mutationer, som kan forklare det lave Vel antigen udtryk.

Vel antigenet var én af de få resterende klinisk vigtige antigener på grund af dens potentielt kraftige transfusionsreaktioner, og opdagelsen af den molekylære baggrund har åbnet op for hurtigt at kunne blodtypebestemme ved hjælp af standard laboratorieteknikker. Anti-Vel antistof egnet til klinisk forskning har gennem historien været svært at få adgang til, men med den molekylære baggrund kendt åbnes nu op for at kunne fremstille kunstigt anti-Vel antistof, som kan bruges til diagnosticering, blodtypebestemmelse og forskning.

Acknowledgements

First of all, to my main supervisor Björn Nilsson; thank you for taking me in and believing in a young hopeful graduate from *grannlandet*. I am grateful you supported my sometimes crazy globetrotter ideas, be it a conference in Seoul, a lab visit on Manhattan or manuscript and thesis writing in Melbourne! I admire your ability to always have valuable feedback even on complex wet lab techniques, far away from your *in silico* and clinical experience!

Secondly, to my other supervisor Jill Storry; thank you for always believing in and caring for me; for your optimistic and enthusiastic approach to science or whenever I had new results to present (I will never forget your excited "*Titta!*" outbursts when I presented something particularly interesting!). But also for reminding me and understanding that there is a life outside the lab and that it is just as important, if not more.

A special thanks to Anna Rignell-Hydblom, for caring about the well-being of all PhD students at MedFak. Your support over the last two years has been invaluable.

Thank you to Connie Westhoff for valuable project feedback and for, on a very short notice, pulling strings and finding loopholes in the strict American legislation to allow me to visit the New York Blood Center.

Speaking of pulling strings, I owe a big thank you to Roger Pocock, who without hesitation welcomed me into his group and provided a much needed lab space in Melbourne.

To Ildiko, my PhD twin and mutual support in countless FACS Aria frustrations, thank you for all our thoughtful conversations about life, *nabolands*-bickering and of course, the Scandihulu.

To the original members of the hematogenomics group; Magnus, Jörgen, Ellinor and Marcus, thank you for bearing with me and my bad Danish/Swedish language crossover in the first half year.

To the rest of the group, Bhairavi, Ram, Britt-Marie, Linnea, Mina, Evelina, Aitzkoa, Angelica, Maroulitsa, Anna-Karin, Ludvig, Jenny and Abhishek; I already miss the lunches, *fika* and everyday lab life.

To Martin Olsson and the Gengrupp, thank you for wholeheartedly welcoming me to all your meetings, lunches, seminars, dinners and parties.

Thank you to Urban Gullberg's group and everyone else at BMC B13 who contributed to make my four years all the more enjoyable.

And a heartfelt thank you to Jesper, my Danish partner-in-crime who more than anyone else made everyday life in the lab (plus the daily cross-border commute) so much more enjoyable! I miss our conversations; nerdy, scientific, silly, serious and everything in between!

Thank you to my awesome friends, Mads, Alexander, Peter, Ronni and Ask, who kept me sane through the years and helped take my mind of science and focus on what matters in life.

To FC Logen, my football/cycling team of like-minded molecular biomedicine graduates, for showing me that well-educated and successful academics, can be incredibly immature and obscene, and have an amazing time together entirely outside the world of science – the bromance is strong in this group!

A big thank you to my family, who have always supported me in everything I've done. Mor og Far, Marianne, Luke (and Oskar), who have always taken a great interest in my scientific endeavours and spurred me on with talks about one day winning the Nobel Prize. And to Luke, thank you for all the craft beer appreciation nights we've shared!

But most of all – to Camilla, my wonderful wife-to-be, my best friend, my globetrotter soulmate, My Love!

References

1 Emerson, S. in *Hematology : basic principles and practice* (eds R. Hoffman *et al.*) 72-81 (Churchill Livingstone, 1991).

2 Shivdasani, R. A., Mayer, E. L. & Orkin, S. H. Absence of blood formation in mice lacking the T-cell leukaemia oncoprotein tal-1/SCL. *Nature* **373**, 432-434, doi:10.1038/373432a0 (1995).

3 Okuda, T., van Deursen, J., Hiebert, S. W., Grosveld, G. & Downing, J. R. AML1, the target of multiple chromosomal translocations in human leukemia, is essential for normal fetal liver hematopoiesis. *Cell* **84**, 321-330 (1996).

4 Lai, A. Y. & Kondo, M. Asymmetrical lymphoid and myeloid lineage commitment in multipotent hematopoietic progenitors. *J Exp Med* **203**, 1867-1873, doi:10.1084/jem.20060697 (2006).

5 Kondo, M., Weissman, I. L. & Akashi, K. Identification of clonogenic common lymphoid progenitors in mouse bone marrow. *Cell* **91**, 661-672 (1997).

6 Akashi, K., Traver, D., Miyamoto, T. & Weissman, I. L. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature* **404**, 193-197, doi:10.1038/35004599 (2000).

7 Georgopoulos, K. *et al.* The Ikaros gene is required for the development of all lymphoid lineages. *Cell* **79**, 143-156 (1994).

8 Rothenberg, E. V. Negotiation of the T lineage fate decision by transcription-factor interplay and microenvironmental signals. *Immunity* **26**, 690-702, doi:10.1016/j.immuni.2007.06.005 (2007).

9 Adams, B. *et al.* Pax-5 encodes the transcription factor BSAP and is expressed in B lymphocytes, the developing CNS, and adult testis. *Genes Dev* **6**, 1589-1607 (1992).

10 Urbanek, P., Wang, Z. Q., Fetka, I., Wagner, E. F. & Busslinger, M. Complete block of early B cell differentiation and altered patterning of the posterior midbrain in mice lacking Pax5/BSAP. *Cell* **79**, 901-912 (1994).

11 Scott, E. W. *et al.* PU.1 functions in a cell-autonomous manner to control the differentiation of multipotential lymphoid-myeloid progenitors. *Immunity* **6**, 437-447 (1997).

12 Iwasaki, H. *et al.* Distinctive and indispensable roles of PU.1 in maintenance of hematopoietic stem cells and their differentiation. *Blood* **106**, 1590-1600, doi:10.1182/blood-2005-03-0860 (2005).

13 Wolfler, A. *et al.* Lineage-instructive function of C/EBPalpha in multipotent hematopoietic cells and early thymic progenitors. *Blood* **116**, 4116-4125, doi:10.1182/blood-2010-03-275404 (2010).
14 Arinobu, Y. *et al.* Reciprocal activation of GATA-1 and PU.1 marks initial specification of hematopoietic stem cells into myeloerythroid and myelolymphoid lineages. *Cell Stem Cell* **1**, 416-427, doi:10.1016/j.stem.2007.07.004 (2007).
15 Rekhtman, N., Radparvar, F., Evans, T. & Skoultchi, A. I. Direct interaction of hematopoietic transcription factors PU.1 and GATA-1: functional antagonism in erythroid cells. *Genes Dev* **13**, 1398-1411 (1999).
16 Mancini, E. *et al.* FOG-1 and GATA-1 act sequentially to specify definitive megakaryocytic and erythroid progenitors. *EMBO J* **31**, 351-365, doi:10.1038/emboj.2011.390 (2012).
17 Starck, J. *et al.* Functional cross-antagonism between transcription factors FLI-1 and EKLF. *Mol Cell Biol* **23**, 1390-1402 (2003).
18 Dahl, R., Iyer, S. R., Owens, K. S., Cuylear, D. D. & Simon, M. C. The transcriptional repressor GFI-1 antagonizes PU.1 activity through protein-protein interaction. *J Biol Chem* **282**, 6473-6483, doi:10.1074/jbc.M607613200 (2007).
19 Schnittger, S. *et al.* RUNX1 mutations are frequent in de novo AML with noncomplex karyotype and confer an unfavorable prognosis. *Blood* **117**, 2348-2357, doi:10.1182/blood-2009-11-255976 (2011).
20 Patel, B. *et al.* Aberrant TAL1 activation is mediated by an interchromosomal interaction in human T-cell acute lymphoblastic leukemia. *Leukemia* **28**, 349-361, doi:10.1038/leu.2013.158 (2014).
21 O'Neil, J., Shank, J., Cusson, N., Murre, C. & Kelliher, M. TAL1/SCL induces leukemia by inhibiting the transcriptional activity of E47/HEB. *Cancer Cell* **5**, 587-596, doi:10.1016/j.ccr.2004.05.023 (2004).
22 Orkin, S. H. Priming the hematopoietic pump. *Immunity* **19**, 633-634 (2003).
23 Kulessa, H., Frampton, J. & Graf, T. GATA-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboblasts, and erythroblasts. *Genes Dev* **9**, 1250-1262 (1995).
24 Iwasaki, H. *et al.* GATA-1 converts lymphoid and myelomonocytic progenitors into the megakaryocyte/erythrocyte lineages. *Immunity* **19**, 451-462 (2003).
25 Bjornson, C. R., Rietze, R. L., Reynolds, B. A., Magli, M. C. & Vescovi, A. L. Turning brain into blood: a hematopoietic fate adopted by adult neural stem cells in vivo. *Science* **283**, 534-537 (1999).
26 Pulecio, J. *et al.* Direct Conversion of Fibroblasts to Megakaryocyte Progenitors. *Cell Rep* **17**, 671-683, doi:10.1016/j.celrep.2016.09.036 (2016).
27 Capellera-Garcia, S. *et al.* Defining the Minimal Factors Required for Erythropoiesis through Direct Lineage Conversion. *Cell Rep* **15**, 2550-2562, doi:10.1016/j.celrep.2016.05.027 (2016).
28 Nutt, S. L., Taubenheim, N., Hasbold, J., Corcoran, L. M. & Hodgkin, P. D. The genetic network controlling plasma cell differentiation. *Semin Immunol* **23**, 341-349, doi:10.1016/j.smim.2011.08.010 (2011).

29 Shaffer, A. L. *et al.* Blimp-1 orchestrates plasma cell differentiation by extinguishing the mature B cell gene expression program. *Immunity* **17**, 51-62 (2002).
30 Nutt, S. L. & Kee, B. L. The transcriptional regulation of B cell lineage commitment. *Immunity* **26**, 715-725, doi:10.1016/j.immuni.2007.05.010 (2007).
31 LeBien, T. W. Fates of human B-cell precursors. *Blood* **96**, 9-23 (2000).
32 Nutt, S. L., Hodgkin, P. D., Tarlinton, D. M. & Corcoran, L. M. The generation of antibody-secreting plasma cells. *Nat Rev Immunol* **15**, 160-171, doi:10.1038/nri3795 (2015).
33 Kallies, A. *et al.* Initiation of plasma-cell differentiation is independent of the transcription factor Blimp-1. *Immunity* **26**, 555-566, doi:10.1016/j.immuni.2007.04.007 (2007).
34 Rawstron, A. C. Immunophenotyping of plasma cells. *Curr Protoc Cytom* **Chapter 6**, Unit6 23, doi:10.1002/0471142956.cy0623s36 (2006).
35 Rajkumar, S. V. *et al.* International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma. *Lancet Oncol* **15**, e538-548, doi:10.1016/S1470-2045(14)70442-5 (2014).
36 Landgren, O. *et al.* Monoclonal gammopathy of undetermined significance (MGUS) consistently precedes multiple myeloma: a prospective study. *Blood* **113**, 5412-5417, doi:10.1182/blood-2008-12-194241 (2009).
37 International Myeloma Working, G. Criteria for the classification of monoclonal gammopathies, multiple myeloma and related disorders: a report of the International Myeloma Working Group. *Br J Haematol* **121**, 749-757 (2003).
38 Landgren, O. *et al.* Risk of monoclonal gammopathy of undetermined significance (MGUS) and subsequent multiple myeloma among African American and white veterans in the United States. *Blood* **107**, 904-906, doi:10.1182/blood-2005-08-3449 (2006).
39 Kyle, R. A. *et al.* Review of 1027 patients with newly diagnosed multiple myeloma. *Mayo Clin Proc* **78**, 21-33, doi:10.4065/78.1.21 (2003).
40 Bergsagel, D. The incidence and epidemiology of plasma cell neoplasms. *Stem Cells* **13 Suppl 2**, 1-9 (1995).
41 Baker, A. *et al.* Uncovering the biology of multiple myeloma among African Americans: a comprehensive genomics approach. *Blood* **121**, 3147-3152, doi:10.1182/blood-2012-07-443606 (2013).
42 Becker, N. Epidemiology of multiple myeloma. *Recent Results Cancer Res* **183**, 25-35, doi:10.1007/978-3-540-85772-3_2 (2011).
43 Broderick, P. *et al.* Common variation at 3p22.1 and 7p15.3 influences multiple myeloma risk. *Nat Genet* **44**, 58-61, doi:10.1038/ng.993 (2011).
44 Swaminathan, B. *et al.* Variants in ELL2 influencing immunoglobulin levels associate with multiple myeloma. *Nat Commun* **6**, 7213, doi:10.1038/ncomms8213 (2015).
45 Weinhold, N. *et al.* The CCND1 c.870G>A polymorphism is a risk factor for t(11;14)(q13;q32) multiple myeloma. *Nat Genet* **45**, 522-525, doi:10.1038/ng.2583 (2013).

46 Koura, D. T. & Langston, A. A. Inherited predisposition to multiple myeloma. *Ther Adv Hematol* **4**, 291-297, doi:10.1177/2040620713485375 (2013).
47 Landgren, O. *et al.* Risk of plasma cell and lymphoproliferative disorders among 14621 first-degree relatives of 4458 patients with monoclonal gammopathy of undetermined significance in Sweden. *Blood* **114**, 791-795, doi:10.1182/blood-2008-12-191676 (2009).
48 Morgan, G. J. *et al.* Inherited genetic susceptibility to multiple myeloma. *Leukemia* **28**, 518-524, doi:10.1038/leu.2013.344 (2014).
49 Bianchi, G., Richardson, P. G. & Anderson, K. C. Promising therapies in multiple myeloma. *Blood* **126**, 300-310, doi:10.1182/blood-2015-03-575365 (2015).
50 San Miguel, J. F. *et al.* Immunophenotypic evaluation of the plasma cell compartment in multiple myeloma: a tool for comparing the efficacy of different treatment strategies and predicting outcome. *Blood* **99**, 1853-1856 (2002).
51 Lin, P., Owens, R., Tricot, G. & Wilson, C. S. Flow cytometric immunophenotypic analysis of 306 cases of multiple myeloma. *Am J Clin Pathol* **121**, 482-488, doi:10.1309/74R4-TB90-BUWH-27JX (2004).
52 Sun, R. X. *et al.* Large scale and clinical grade purification of syndecan-1+ malignant plasma cells. *J Immunol Methods* **205**, 73-79 (1997).
53 Kumar, S., Kimlinger, T. & Morice, W. Immunophenotyping in multiple myeloma and related plasma cell disorders. *Best Pract Res Clin Haematol* **23**, 433-451, doi:10.1016/j.beha.2010.09.002 (2010).
54 Frigyesi, I. *et al.* Robust isolation of malignant plasma cells in multiple myeloma. *Blood* **123**, 1336-1340, doi:10.1182/blood-2013-09-529800 (2014).
55 D'Alessandro, A., Righetti, P. G. & Zolla, L. The red blood cell proteome and interactome: an update. *J Proteome Res* **9**, 144-163, doi:10.1021/pr900831f (2010).
56 Elliott, S., Pham, E. & Macdougall, I. C. Erythropoietins: a common mechanism of action. *Exp Hematol* **36**, 1573-1584, doi:10.1016/j.exphem.2008.08.003 (2008).
57 Koury, M. J., Sawyer, S. T. & Brandt, S. J. New insights into erythropoiesis. *Curr Opin Hematol* **9**, 93-100 (2002).
58 Schifferli, J. A. & Taylor, R. P. Physiological and pathological aspects of circulating immune complexes. *Kidney Int* **35**, 993-1003 (1989).
59 Siatecka, M. & Bieker, J. J. The multifunctional role of EKLF/KLF1 during erythropoiesis. *Blood* **118**, 2044-2054, doi:10.1182/blood-2011-03-331371 (2011).
60 Birkmann, J. *et al.* Effects of recombinant human thrombopoietin alone and in combination with erythropoietin and early-acting cytokines on human mobilized purified CD34+ progenitor cells cultured in serum-depleted medium. *Stem Cells* **15**, 18-32, doi:10.1002/stem.150018 (1997).
61 Gregory, C. J. & Eaves, A. C. Human marrow cells capable of erythropoietic differentiation in vitro: definition of three erythroid colony responses. *Blood* **49**, 855-864 (1977).

62 Pal, S. *et al.* Coregulator-dependent facilitation of chromatin occupancy by GATA-1. *Proc Natl Acad Sci U S A* **101**, 980-985, doi:10.1073/pnas.0307612100 (2004).
63 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**, 8811-8816, doi:10.1073/pnas.1432147100 (2003).
64 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**, 6258-6268, doi:10.1182/blood-2011-07-356006 (2011).
65 Wu, H., Liu, X., Jaenisch, R. & Lodish, H. F. Generation of committed erythroid BFU-E and CFU-E progenitors does not require erythropoietin or the erythropoietin receptor. *Cell* **83**, 59-67 (1995).
66 Lin, C. S., Lim, S. K., D'Agati, V. & Costantini, F. Differential effects of an erythropoietin receptor gene disruption on primitive and definitive erythropoiesis. *Genes Dev* **10**, 154-164 (1996).
67 Koury, M. J. & Bondurant, M. C. Erythropoietin retards DNA breakdown and prevents programmed death in erythroid progenitor cells. *Science* **248**, 378-381 (1990).
68 Zhu, B. M. *et al.* Hematopoietic-specific Stat5-null mice display microcytic hypochromic anemia associated with reduced transferrin receptor gene expression. *Blood* **112**, 2071-2080, doi:10.1182/blood-2007-12-127480 (2008).
69 Socolovsky, M., Fallon, A. E., Wang, S., Brugnara, C. & Lodish, H. F. Fetal anemia and apoptosis of red cell progenitors in Stat5a-/-5b-/- mice: a direct role for Stat5 in Bcl-X(L) induction. *Cell* **98**, 181-191 (1999).
70 Zhang, A. S., Sheftel, A. D. & Ponka, P. Intracellular kinetics of iron in reticulocytes: evidence for endosome involvement in iron targeting to mitochondria. *Blood* **105**, 368-375, doi:10.1182/blood-2004-06-2226 (2005).
71 Chen, W., Dailey, H. A. & Paw, B. H. Ferrochelatase forms an oligomeric complex with mitoferrin-1 and Abcb10 for erythroid heme biosynthesis. *Blood* **116**, 628-630, doi:10.1182/blood-2009-12-259614 (2010).
72 Levy, J. E., Jin, O., Fujiwara, Y., Kuo, F. & Andrews, N. C. Transferrin receptor is necessary for development of erythrocytes and the nervous system. *Nat Genet* **21**, 396-399, doi:10.1038/7727 (1999).
73 Sun, J. *et al.* Heme regulates the dynamic exchange of Bach1 and NF-E2-related factors in the Maf transcription factor network. *Proc Natl Acad Sci U S A* **101**, 1461-1466, doi:10.1073/pnas.0308083100 (2004).
74 Chen, K. *et al.* Resolving the distinct stages in erythroid differentiation based on dynamic changes in membrane protein expression during erythropoiesis. *Proc Natl Acad Sci U S A* **106**, 17413-17418, doi:10.1073/pnas.0909296106 (2009).
75 Koury, S. T., Koury, M. J. & Bondurant, M. C. Cytoskeletal distribution and function during the maturation and enucleation of mammalian erythroblasts. *J Cell Biol* **109**, 3005-3013 (1989).

76 Ji, P., Murata-Hori, M. & Lodish, H. F. Formation of mammalian erythrocytes: chromatin condensation and enucleation. *Trends Cell Biol* **21**, 409-415, doi:10.1016/j.tcb.2011.04.003 (2011).
77 Su, M. Y. *et al.* Identification of biologically relevant enhancers in human erythroid cells. *J Biol Chem* **288**, 8433-8444, doi:10.1074/jbc.M112.413260 (2013).
78 McDevitt, M. A., Shivdasani, R. A., Fujiwara, Y., Yang, H. & Orkin, S. H. A "knockdown" mutation created by cis-element gene targeting reveals the dependence of erythroid cell maturation on the level of transcription factor GATA-1. *Proc Natl Acad Sci U S A* **94**, 6781-6785 (1997).
79 Wall, L., deBoer, E. & Grosveld, F. The human beta-globin gene 3' enhancer contains multiple binding sites for an erythroid-specific protein. *Genes Dev* **2**, 1089-1100 (1988).
80 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**, 5976-5980 (1988).
81 Wierenga, A. T., Vellenga, E. & Schuringa, J. J. Down-regulation of GATA1 uncouples STAT5-induced erythroid differentiation from stem/progenitor cell proliferation. *Blood* **115**, 4367-4376, doi:10.1182/blood-2009-10-250894 (2010).
82 Welch, J. J. *et al.* Global regulation of erythroid gene expression by transcription factor GATA-1. *Blood* **104**, 3136-3147, doi:10.1182/blood-2004-04-1603 (2004).
83 Nerlov, C., Querfurth, E., Kulessa, H. & Graf, T. GATA-1 interacts with the myeloid PU.1 transcription factor and represses PU.1-dependent transcription. *Blood* **95**, 2543-2551 (2000).
84 Crispino, J. D. & Weiss, M. J. Erythro-megakaryocytic transcription factors associated with hereditary anemia. *Blood* **123**, 3080-3088, doi:10.1182/blood-2014-01-453167 (2014).
85 Nichols, K. E. *et al.* Familial dyserythropoietic anaemia and thrombocytopenia due to an inherited mutation in GATA1. *Nat Genet* **24**, 266-270, doi:10.1038/73480 (2000).
86 Fujiwara, T. *et al.* Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. *Mol Cell* **36**, 667-681, doi:10.1016/j.molcel.2009.11.001 (2009).
87 Yu, M. *et al.* Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. *Mol Cell* **36**, 682-695, doi:10.1016/j.molcel.2009.11.002 (2009).
88 Merika, M. & Orkin, S. H. DNA-binding specificity of GATA family transcription factors. *Mol Cell Biol* **13**, 3999-4010 (1993).
89 Tsang, A. P. *et al.* FOG, a multitype zinc finger protein, acts as a cofactor for transcription factor GATA-1 in erythroid and megakaryocytic differentiation. *Cell* **90**, 109-119 (1997).
90 Begley, C. G. *et al.* Chromosomal translocation in a human leukemic stem-cell line disrupts the T-cell antigen receptor delta-chain diversity region and results in a previously unreported fusion transcript. *Proc Natl Acad Sci U S A* **86**, 2031-2035 (1989).

91 Begley, C. G. *et al.* The gene SCL is expressed during early hematopoiesis and encodes a differentiation-related DNA-binding motif. *Proc Natl Acad Sci U S A* **86**, 10128-10132 (1989).
92 Chen, Q. *et al.* The tal gene undergoes chromosome translocation in T cell leukemia and potentially encodes a helix-loop-helix protein. *EMBO J* **9**, 415-424 (1990).
93 Porcher, C., Liao, E. C., Fujiwara, Y., Zon, L. I. & Orkin, S. H. Specification of hematopoietic and vascular development by the bHLH transcription factor SCL without direct DNA binding. *Development* **126**, 4603-4615 (1999).
94 Kassouf, M. T., Chagraoui, H., Vyas, P. & Porcher, C. Differential use of SCL/TAL-1 DNA-binding domain in developmental hematopoiesis. *Blood* **112**, 1056-1067, doi:10.1182/blood-2007-12-128900 (2008).
95 Lahlil, R., Lecuyer, E., Herblot, S. & Hoang, T. SCL assembles a multifactorial complex that determines glycophorin A expression. *Mol Cell Biol* **24**, 1439-1452 (2004).
96 Vyas, P. *et al.* Different sequence requirements for expression in erythroid and megakaryocytic cells within a regulatory element upstream of the GATA-1 gene. *Development* **126**, 2799-2811 (1999).
97 Anderson, K. P., Crable, S. C. & Lingrel, J. B. The GATA-E box-GATA motif in the EKLF promoter is required for in vivo expression. *Blood* **95**, 1652-1655 (2000).
98 Kassouf, M. T. *et al.* Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. *Genome Res* **20**, 1064-1083, doi:10.1101/gr.104935.110 (2010).
99 Mouthon, M. A. *et al.* Expression of tal-1 and GATA-binding proteins during human hematopoiesis. *Blood* **81**, 647-655 (1993).
100 Zweidler-Mckay, P. A., Grimes, H. L., Flubacher, M. M. & Tsichlis, P. N. Gfi-1 encodes a nuclear zinc finger protein that binds DNA and functions as a transcriptional repressor. *Mol Cell Biol* **16**, 4024-4034 (1996).
101 Grimes, H. L., Chan, T. O., Zweidler-McKay, P. A., Tong, B. & Tsichlis, P. N. The Gfi-1 proto-oncoprotein contains a novel transcriptional repressor domain, SNAG, and inhibits G1 arrest induced by interleukin-2 withdrawal. *Mol Cell Biol* **16**, 6263-6272 (1996).
102 Vassen, L., Okayama, T. & Moroy, T. Gfi1b:green fluorescent protein knock-in mice reveal a dynamic expression pattern of Gfi1b during hematopoiesis that is largely complementary to Gfi1. *Blood* **109**, 2356-2364, doi:10.1182/blood-2006-06-030031 (2007).
103 Yucel, R., Kosan, C., Heyd, F. & Moroy, T. Gfi1:green fluorescent protein knock-in mutant reveals differential expression and autoregulation of the growth factor independence 1 (Gfi1) gene during lymphocyte development. *J Biol Chem* **279**, 40906-40917, doi:10.1074/jbc.M400808200 (2004).
104 Osawa, M. *et al.* Erythroid expansion mediated by the Gfi-1B zinc finger protein: role in normal hematopoiesis. *Blood* **100**, 2769-2777, doi:10.1182/blood-2002-01-0182 (2002).

105 Moignard, V. *et al.* Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nat Cell Biol* **15**, 363-372, doi:10.1038/ncb2709 (2013).
106 Vassen, L., Fiolka, K. & Moroy, T. Gfi1b alters histone methylation at target gene promoters and sites of gamma-satellite containing heterochromatin. *EMBO J* **25**, 2409-2419, doi:10.1038/sj.emboj.7601124 (2006).
107 Rodriguez, P. *et al.* GATA-1 forms distinct activating and repressive complexes in erythroid cells. *EMBO J* **24**, 2354-2366, doi:10.1038/sj.emboj.7600702 (2005).
108 Kuo, Y. Y. & Chang, Z. F. GATA-1 and Gfi-1B interplay to regulate Bcl-xL transcription. *Mol Cell Biol* **27**, 4261-4272, doi:10.1128/MCB.02212-06 (2007).
109 Huang, D. Y., Kuo, Y. Y. & Chang, Z. F. GATA-1 mediates auto-regulation of Gfi-1B transcription in K562 cells. *Nucleic Acids Res* **33**, 5331-5342, doi:10.1093/nar/gki838 (2005).
110 Miller, I. J. & Bieker, J. J. A novel, erythroid cell-specific murine transcription factor that binds to the CACCC element and is related to the Kruppel family of nuclear proteins. *Mol Cell Biol* **13**, 2776-2786 (1993).
111 Tallack, M. R. & Perkins, A. C. Megakaryocyte-erythroid lineage promiscuity in EKLF null mouse blood. *Haematologica* **95**, 144-147, doi:10.3324/haematol.2009.010017 (2010).
112 Wijgerde, M. *et al.* The role of EKLF in human beta-globin gene competition. *Genes Dev* **10**, 2894-2902 (1996).
113 Perkins, A. C., Gaensler, K. M. & Orkin, S. H. Silencing of human fetal globin expression is impaired in the absence of the adult beta-globin gene activator protein EKLF. *Proc Natl Acad Sci U S A* **93**, 12267-12271 (1996).
114 Hodge, D. *et al.* A global role for EKLF in definitive and primitive erythropoiesis. *Blood* **107**, 3359-3370, doi:10.1182/blood-2005-07-2888 (2006).
115 Tallack, M. R. *et al.* A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. *Genome Res* **20**, 1052-1063, doi:10.1101/gr.106575.110 (2010).
116 Chen, X. & Bieker, J. J. Stage-specific repression by the EKLF transcriptional activator. *Mol Cell Biol* **24**, 10416-10424, doi:10.1128/MCB.24.23.10416-10424.2004 (2004).
117 Siatecka, M., Xue, L. & Bieker, J. J. Sumoylation of EKLF promotes transcriptional repression and is involved in inhibition of megakaryopoiesis. *Mol Cell Biol* **27**, 8547-8560, doi:10.1128/MCB.00589-07 (2007).
118 Mitchell, J. A. & Fraser, P. Transcription factories are nuclear subcompartments that remain in the absence of transcription. *Genes Dev* **22**, 20-25, doi:10.1101/gad.454008 (2008).
119 Osborne, C. S. *et al.* Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat Genet* **36**, 1065-1071, doi:10.1038/ng1423 (2004).
120 Wansink, D. G. *et al.* Fluorescent labeling of nascent RNA reveals transcription by RNA polymerase II in domains scattered throughout the nucleus. *J Cell Biol* **122**, 283-293 (1993).

121 Schoenfelder, S. *et al.* Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet* **42**, 53-61, doi:10.1038/ng.496 (2010).
122 Royer-Pokora, B., Loos, U. & Ludwig, W. D. TTG-2, a new gene encoding a cysteine-rich protein with the LIM motif, is overexpressed in acute T-cell leukaemia with the t(11;14)(p13;q11). *Oncogene* **6**, 1887-1893 (1991).
123 Sanchez-Garcia, I., Axelson, H. & Rabbitts, T. H. Functional diversity of LIM proteins: amino-terminal activation domains in the oncogenic proteins RBTN1 and RBTN2. *Oncogene* **10**, 1301-1306 (1995).
124 Warren, A. J. *et al.* The oncogenic cysteine-rich LIM domain protein rbtn2 is essential for erythroid development. *Cell* **78**, 45-57 (1994).
125 Osada, H., Grutz, G., Axelson, H., Forster, A. & Rabbitts, T. H. Association of erythroid transcription factors: complexes involving the LIM protein RBTN2 and the zinc-finger protein GATA1. *Proc Natl Acad Sci U S A* **92**, 9585-9589 (1995).
126 Wadman, I. A. *et al.* The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. *EMBO J* **16**, 3145-3157, doi:10.1093/emboj/16.11.3145 (1997).
127 Landry, J. R. *et al.* Expression of the leukemia oncogene Lmo2 is controlled by an array of tissue-specific elements dispersed over 100 kb and bound by Tal1/Lmo2, Ets, and Gata factors. *Blood* **113**, 5783-5792, doi:10.1182/blood-2008-11-187757 (2009).
128 Davies, J. M. *et al.* Novel BTB/POZ domain zinc-finger protein, LRF, is a potential target of the LAZ-3/BCL-6 oncogene. *Oncogene* **18**, 365-375, doi:10.1038/sj.onc.1202332 (1999).
129 Maeda, T. *et al.* Role of the proto-oncogene Pokemon in cellular transformation and ARF repression. *Nature* **433**, 278-285, doi:10.1038/nature03203 (2005).
130 Maeda, T. *et al.* LRF is an essential downstream target of GATA1 in erythroid development and regulates BIM-dependent apoptosis. *Dev Cell* **17**, 527-540, doi:10.1016/j.devcel.2009.09.005 (2009).
131 Masuda, T. *et al.* Transcription factors LRF and BCL11A independently repress expression of fetal hemoglobin. *Science* **351**, 285-289, doi:10.1126/science.aad3312 (2016).
132 Pessler, F. & Hernandez, N. Flexible DNA binding of the BTB/POZ-domain protein FBI-1. *J Biol Chem* **278**, 29327-29335, doi:10.1074/jbc.M302980200 (2003).
133 Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327-339, doi:10.1016/j.cell.2012.12.009 (2013).
134 Cheng, Y. *et al.* Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Res* **19**, 2172-2184, doi:10.1101/gr.098921.109 (2009).
135 Hsu, H. L., Wadman, I. & Baer, R. Formation of in vivo complexes between the TAL1 and E2A polypeptides of leukemic T cells. *Proc Natl Acad Sci U S A* **91**, 3181-3185 (1994).

136 Song, S. H. *et al.* Multiple functions of Ldb1 required for beta-globin activation during erythroid differentiation. *Blood* **116**, 2356-2364, doi:10.1182/blood-2010-03-272252 (2010).
137 Cross, A. J., Jeffries, C. M., Trewhella, J. & Matthews, J. M. LIM domain binding proteins 1 and 2 have different oligomeric states. *J Mol Biol* **399**, 133-144, doi:10.1016/j.jmb.2010.04.006 (2010).
138 Love, P. E., Warzecha, C. & Li, L. Ldb1 complexes: the new master regulators of erythroid gene transcription. *Trends Genet* **30**, 1-9, doi:10.1016/j.tig.2013.10.001 (2014).
139 Anguita, E. *et al.* Globin gene activation during haemopoiesis is driven by protein complexes nucleated by GATA-1 and GATA-2. *EMBO J* **23**, 2841-2852, doi:10.1038/sj.emboj.7600274 (2004).
140 Crispino, J. D., Lodish, M. B., MacKay, J. P. & Orkin, S. H. Use of altered specificity mutants to probe a specific protein-protein interaction in differentiation: the GATA-1:FOG complex. *Mol Cell* **3**, 219-228 (1999).
141 Letting, D. L., Chen, Y. Y., Rakowski, C., Reedy, S. & Blobel, G. A. Context-dependent regulation of GATA-1 by friend of GATA-1. *Proc Natl Acad Sci U S A* **101**, 476-481, doi:10.1073/pnas.0306315101 (2004).
142 Feng, Q. & Zhang, Y. The MeCP1 complex represses transcription through preferential binding, remodeling, and deacetylating methylated nucleosomes. *Genes Dev* **15**, 827-832, doi:10.1101/gad.876201 (2001).
143 Xue, Y. *et al.* NURD, a novel complex with both ATP-dependent chromatin-remodeling and histone deacetylase activities. *Mol Cell* **2**, 851-861 (1998).
144 Hong, W. *et al.* FOG-1 recruits the NuRD repressor complex to mediate transcriptional repression by GATA-1. *EMBO J* **24**, 2367-2378, doi:10.1038/sj.emboj.7600703 (2005).
145 Saleque, S., Kim, J., Rooke, H. M. & Orkin, S. H. Epigenetic regulation of hematopoietic differentiation by Gfi-1 and Gfi-1b is mediated by the cofactors CoREST and LSD1. *Mol Cell* **27**, 562-572, doi:10.1016/j.molcel.2007.06.039 (2007).
146 Chowdhury, A. H. *et al.* Differential transcriptional regulation of meis1 by Gfi1b and its co-factors LSD1 and CoREST. *PLoS One* **8**, e53666, doi:10.1371/journal.pone.0053666 (2013).
147 Fujiwara, T. *et al.* Role of transcriptional corepressor ETO2 in erythroid cells. *Exp Hematol* **41**, 303-315 e301, doi:10.1016/j.exphem.2012.10.015 (2013).
148 Schuh, A. H. *et al.* ETO-2 associates with SCL in erythroid cells and megakaryocytes and provides repressor functions in erythropoiesis. *Mol Cell Biol* **25**, 10235-10250, doi:10.1128/MCB.25.23.10235-10250.2005 (2005).
149 Goardon, N. *et al.* ETO2 coordinates cellular proliferation and differentiation during erythropoiesis. *EMBO J* **25**, 357-366, doi:10.1038/sj.emboj.7600934 (2006).
150 Kadam, S. & Emerson, B. M. Transcriptional specificity of human SWI/SNF BRG1 and BRM chromatin remodeling complexes. *Mol Cell* **11**, 377-389 (2003).

151 Blobel, G. A., Nakajima, T., Eckner, R., Montminy, M. & Orkin, S. H. CREB-binding protein cooperates with transcription factor GATA-1 and is required for erythroid differentiation. *Proc Natl Acad Sci U S A* **95**, 2061-2066 (1998).
152 Hu, X. *et al.* LSD1-mediated epigenetic modification is required for TAL1 function and hematopoiesis. *Proc Natl Acad Sci U S A* **106**, 10141-10146, doi:10.1073/pnas.0900437106 (2009).
153 Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215-233, doi:10.1016/j.cell.2009.01.002 (2009).
154 Guo, H., Ingolia, N. T., Weissman, J. S. & Bartel, D. P. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**, 835-840, doi:10.1038/nature09267 (2010).
155 Felli, N. *et al.* MicroRNA 223-dependent expression of LMO2 regulates normal erythropoiesis. *Haematologica* **94**, 479-486, doi:10.3324/haematol.2008.002345 (2009).
156 Zhang, L., Flygare, J., Wong, P., Lim, B. & Lodish, H. F. miR-191 regulates mouse erythroblast enucleation by down-regulating Riok3 and Mxi1. *Genes Dev* **25**, 119-124, doi:10.1101/gad.1998711 (2011).
157 Byon, J. C. & Papayannopoulou, T. MicroRNAs: Allies or foes in erythropoiesis? *J Cell Physiol* **227**, 7-13, doi:10.1002/jcp.22729 (2012).
158 Listowski, M. A. *et al.* microRNAs: fine tuning of erythropoiesis. *Cell Mol Biol Lett* **18**, 34-46, doi:10.2478/s11658-012-0038-z (2013).
159 Hu, W., Yuan, B., Flygare, J. & Lodish, H. F. Long noncoding RNA-mediated anti-apoptotic activity in murine erythroid terminal differentiation. *Genes Dev* **25**, 2573-2578, doi:10.1101/gad.178780.111 (2011).
160 Alvarez-Dominguez, J. R. *et al.* Global discovery of erythroid long noncoding RNAs reveals novel regulators of red cell maturation. *Blood* **123**, 570-581, doi:10.1182/blood-2013-10-530683 (2014).
161 Schwarz, H. P. & Dorner, F. Karl Landsteiner and his major contributions to haematology. *Br J Haematol* **121**, 556-565 (2003).
162 Pettenkofer, H. J., Maassen, W. & Bickerich, R. [Antigen relations between human blood groups and enterobacteriacea]. *Z Immun exp ther* **119**, 415-429 (1960).
163 Springer, G. F. [On the origin of normal antibodies]. *Klin Wochenschr* **38**, 513-514 (1960).
164 Allen, F. H. *et al.* ISBT working party on terminology for red cell surface antigens. Preliminary report. *Vox Sang* **42**, 164-165 (1982).
165 Yamamoto, F., Clausen, H., White, T., Marken, J. & Hakomori, S. Molecular genetic basis of the histo-blood group ABO system. *Nature* **345**, 229-233 (1990).
166 Daniels, G. in *Human Blood Groups* 1-10 (John Wiley & Sons,, Chichester, West Sussex, 2013).
167 Anliker, M. *et al.* A new blood group antigen is defined by anti-CD59, detected in a CD59-deficient patient. *Transfusion* **54**, 1817-1822, doi:10.1111/trf.12531 (2014).

168 Hochsmann, B., Dohna-Schwake, C., Kyrieleis, H. A., Pannicke, U. & Schrezenmeier, H. Targeted therapy with eculizumab for inherited CD59 deficiency. *N Engl J Med* **370**, 90-92, doi:10.1056/NEJMc1308104 (2014).
169 Daniels, G. *et al.* Lack of the nucleoside transporter ENT1 results in the Augustine-null blood type and ectopic mineralization. *Blood* **125**, 3651-3654, doi:10.1182/blood-2015-03-631598 (2015).
170 Daniels, G. Functions of red cell surface proteins. *Vox Sang* **93**, 331-340, doi:10.1111/j.1423-0410.2007.00970.x (2007).
171 Storry, J. R. *et al.* Homozygosity for a null allele of SMIM1 defines the Vel-negative blood group phenotype. *Nat Genet* **45**, 537-541, doi:10.1038/ng.2600 (2013).
172 Anstee, D. J. The functional importance of blood group-active molecules in human red blood cells. *Vox Sang* **100**, 140-149, doi:10.1111/j.1423-0410.2010.01388.x (2011).
173 King, L. S., Choi, M., Fernandez, P. C., Cartron, J. P. & Agre, P. Defective urinary-concentrating ability due to a complete deficiency of aquaporin-1. *N Engl J Med* **345**, 175-179, doi:10.1056/NEJM200107193450304 (2001).
174 Gazzinelli, R. T., Kalantari, P., Fitzgerald, K. A. & Golenbock, D. T. Innate sensing of malaria parasites. *Nat Rev Immunol* **14**, 744-757, doi:10.1038/nri3742 (2014).
175 Kwiatkowski, D. P. How malaria has affected the human genome and what human genetics can teach us about malaria. *Am J Hum Genet* **77**, 171-192, doi:10.1086/432519 (2005).
176 Cserti, C. M. & Dzik, W. H. The ABO blood group system and Plasmodium falciparum malaria. *Blood* **110**, 2250-2258, doi:10.1182/blood-2007-03-077602 (2007).
177 Carlson, J. & Wahlgren, M. Plasmodium falciparum erythrocyte rosetting is mediated by promiscuous lectin-like interactions. *J Exp Med* **176**, 1311-1317 (1992).
178 Chen, Q. *et al.* The semiconserved head structure of Plasmodium falciparum erythrocyte membrane protein 1 mediates binding to multiple independent host receptors. *J Exp Med* **192**, 1-10 (2000).
179 Krych-Goldberg, M., Moulds, J. M. & Atkinson, J. P. Human complement receptor type 1 (CR1) binds to a major malarial adhesin. *Trends Mol Med* **8**, 531-537 (2002).
180 Orlandi, P. A., Klotz, F. W. & Haynes, J. D. A malaria invasion receptor, the 175-kilodalton erythrocyte binding antigen of Plasmodium falciparum recognizes the terminal Neu5Ac(alpha 2-3)Gal- sequences of glycophorin A. *J Cell Biol* **116**, 901-909 (1992).
181 Mayer, D. C. *et al.* Glycophorin B is the erythrocyte receptor of Plasmodium falciparum erythrocyte-binding ligand, EBL-1. *Proc Natl Acad Sci U S A* **106**, 5348-5352, doi:10.1073/pnas.0900878106 (2009).
182 Maier, A. G. *et al.* Plasmodium falciparum erythrocyte invasion through glycophorin C and selection for Gerbich negativity in human populations. *Nat Med* **9**, 87-92, doi:10.1038/nm807 (2003).

183 Crosnier, C. *et al.* Basigin is a receptor essential for erythrocyte invasion by Plasmodium falciparum. *Nature* **480**, 534-537, doi:10.1038/nature10606 (2011).
184 Satchwell, T. J. Erythrocyte invasion receptors for Plasmodium falciparum: new and old. *Transfus Med* **26**, 77-88, doi:10.1111/tme.12280 (2016).
185 Cutbush, M. & Mollison, P. L. The Duffy blood group system. *Heredity (Edinb)* **4**, 383-389 (1950).
186 Chaudhuri, A. *et al.* Expression of the Duffy antigen in K562 cells. Evidence that it is the human erythrocyte chemokine receptor. *J Biol Chem* **269**, 7835-7838 (1994).
187 Neote, K., Mak, J. Y., Kolakowski, L. F., Jr. & Schall, T. J. Functional and biochemical analysis of the cloned Duffy antigen: identity with the red blood cell chemokine receptor. *Blood* **84**, 44-52 (1994).
188 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**, 224-228, doi:10.1038/ng0695-224 (1995).
189 Sanger, R., Race, R. R. & Jack, J. The Duffy blood groups of New York negroes: the phenotype Fy (a-b-). *Br J Haematol* **1**, 370-374 (1955).
190 Miller, L. H., Mason, S. J., Clyde, D. F. & McGinniss, M. H. The resistance factor to Plasmodium vivax in blacks. The Duffy-blood-group genotype, FyFy. *N Engl J Med* **295**, 302-304, doi:10.1056/NEJM197608052950602 (1976).
191 Miller, L. H., Mason, S. J., Dvorak, J. A., McGinniss, M. H. & Rothman, I. K. Erythrocyte receptors for (Plasmodium knowlesi) malaria: Duffy blood group determinants. *Science* **189**, 561-563 (1975).
192 Mason, S. J., Miller, L. H., Shiroishi, T., Dvorak, J. A. & McGinniss, M. H. The Duffy blood group determinants: their role in the susceptibility of human and animal erythrocytes to Plasmodium knowlesi malaria. *Br J Haematol* **36**, 327-335 (1977).
193 Wang, Z. *et al.* ABO blood group system and gastric cancer: a case-control study and meta-analysis. *Int J Mol Sci* **13**, 13308-13321, doi:10.3390/ijms131013308 (2012).
194 Liumbruno, G. M. & Franchini, M. Beyond immunohaematology: the role of the ABO blood group in human diseases. *Blood transfusion = Trasfusione del sangue* **11**, 491-499, doi:10.2450/2013.0152-13 (2013).
195 Rizzato, C. *et al.* ABO blood groups and pancreatic cancer risk and survival: results from the PANcreatic Disease ReseArch (PANDoRA) consortium. *Oncol Rep* **29**, 1637-1644, doi:10.3892/or.2013.2285 (2013).
196 Wolpin, B. M. *et al.* Variant ABO blood group alleles, secretor status, and risk of pancreatic cancer: results from the pancreatic cancer cohort consortium. *Cancer Epidemiol Biomarkers Prev* **19**, 3140-3149, doi:10.1158/1055-9965.EPI-10-0751 (2010).
197 Boren, T., Falk, P., Roth, K. A., Larson, G. & Normark, S. Attachment of Helicobacter pylori to human gastric epithelium mediated by blood group antigens. *Science* **262**, 1892-1895 (1993).

198 Jung, H. H., Danek, A. & Frey, B. M. McLeod syndrome: a neurohaematological disorder. *Vox Sang* **93**, 112-121, doi:10.1111/j.1423-0410.2007.00949.x (2007).
199 Sussman, L. N. & Miller, E. B. [New blood factor: Vel.]. *Rev.Hematol.* **7**, 368-371 (1952).
200 Albrey, J. A., McCulloch, W. J. & Simmons, R. T. Inheritance of the Vel blood group in three families. *Med J Aust* **2**, 662-665 (1965).
201 Battaglini, P. F., Ranque, J., Bridonneau, C., Salmon, C. & Nicoli, R. M. [Study of the VEL factor in the Marseilles population apropos of a case of anti-VEL immunization]. *Bibl Haematol* **23**, 309-311 (1965).
202 Sussman, L. N. Current status of the Vel blood group system. *Transfusion* **2**, 163-171 (1962).
203 Levine, P., Robinson, E. A., Herrington, L. B. & Sussman, L. N. Second example of the antibody for the high-incidence blood factor Vel. *Am J Clin Pathol* **25**, 751-754 (1955).
204 Herron, R., Hyde, R. D. & Hillier, S. J. The second example of an anti-vel auto-antibody. *Vox Sang* **36**, 179-181 (1979).
205 Gale, S. A., Rowe, G. P. & Northfield, F. E. Application of a microtitre plate antiglobulin technique to determine the incidence of donors lacking high frequency antigens. *Vox Sang* **54**, 172-173 (1988).
206 Issitt, P. D. & Anstee, D. J. in *Applied blood group serology* Ch. 31, 801-804 (Montgomery Scientific, 1998).
207 Cedergren, B., Giles, C. M. & Ikin, E. W. The Vel blood group in northern Sweden. *Vox Sanguinis* **31**, 344-355 (1976).
208 Issitt, P. D. *et al.* Anti-Vel 2, a new antibody showing heterogeneity of Vel system antibodies. *Vox Sang* **15**, 125-132 (1968).
209 Ballif, B. A. *et al.* Disruption of SMIM1 causes the Vel- blood type. *EMBO Mol Med* **5**, 751-761, doi:10.1002/emmm.201302466 (2013).
210 Cvejic, A. *et al.* SMIM1 underlies the Vel blood group and influences red blood cell traits. *Nat Genet* **45**, 542-545, doi:10.1038/ng.2603 (2013).
211 Storry, J. R. *et al.* International Society of Blood Transfusion Working Party on red cell immunogenetics and terminology: report of the Seoul and London meetings. *ISBT Science Series* **11**, 118-122, doi:10.1111/voxs.12280 (2016).
212 ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).
213 Haer-Wigman, L. *et al.* Impact of genetic variation in the SMIM1 gene on Vel expression levels. *Transfusion* **55**, 1457-1466, doi:10.1111/trf.13014 (2015).
214 Christophersen, M. K. *et al.* SMIM1 variants rs1175550 and rs143702418 independently modulate Vel blood group antigen expression. *Scientific reports* **7**, 40451, doi:10.1038/srep40451 (2017).
215 MacKenzie, K. R. & Engelman, D. M. Structure-based prediction of the stability of transmembrane helix-helix interactions: the sequence dependence of glycophorin A dimerization. *Proc Natl Acad Sci U S A* **95**, 3583-3590 (1998).
216 Arnaud, L., Kelley, L. P., Helias, V., Cartron, J. P. & Ballif, B. A. SMIM1 is a type II transmembrane phosphoprotein and displays the Vel blood group antigen

at its carboxyl-terminus. *FEBS Lett* **589**, 3624-3630, doi:10.1016/j.febslet.2015.09.029 (2015).
217 van der Harst, P. *et al.* Seventy-five genetic loci influencing the human red blood cell. *Nature* **492**, 369-375, doi:10.1038/nature11677 (2012).
218 Evans, D. M. *et al.* Genome-wide association study identifies loci affecting blood copper, selenium and zinc. *Hum Mol Genet* **22**, 3998-4006, doi:10.1093/hmg/ddt239 (2013).
219 Gravel, S. Population genetics models of local ancestry. *Genetics* **191**, 607-619, doi:10.1534/genetics.112.139808 (2012).
220 International Human Genome Sequencing, C. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945, doi:10.1038/nature03001 (2004).
221 Consortium, U. K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90, doi:10.1038/nature14962 (2015).
222 Gurdasani, D. *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**, 327-332, doi:10.1038/nature13997 (2015).
223 The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).
224 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).
225 Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-311 (2001).
226 Wall, J. D. & Pritchard, J. K. Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet* **4**, 587-597, doi:10.1038/nrg1123 (2003).
227 International HapMap, C. A haplotype map of the human genome. *Nature* **437**, 1299-1320, doi:10.1038/nature04226 (2005).
228 Wellcome Trust Case Control, C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678, doi:10.1038/nature05911 (2007).
229 Schadt, E. E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297-302, doi:10.1038/nature01434 (2003).
230 Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* **6**, e1000888, doi:10.1371/journal.pgen.1000888 (2010).
231 Vijai, J. *et al.* A genome-wide association study of marginal zone lymphoma shows association to the HLA region. *Nat Commun* **6**, 5751, doi:10.1038/ncomms6751 (2015).
232 Vijayakrishnan, J. *et al.* A genome-wide association study identifies risk loci for childhood acute lymphoblastic leukemia at 10q26.13 and 12q23.1. *Leukemia*, doi:10.1038/leu.2016.271 (2016).
233 Iyengar, S. K. *et al.* Genome-Wide Association and Trans-ethnic Meta-Analysis for Advanced Diabetic Kidney Disease: Family Investigation of Nephropathy and Diabetes (FIND). *PLoS Genet* **11**, e1005352, doi:10.1371/journal.pgen.1005352 (2015).

234 Liu, C. T. *et al.* Trans-ethnic Meta-analysis and Functional Annotation Illuminates the Genetic Architecture of Fasting Glucose and Insulin. *Am J Hum Genet* **99**, 56-75, doi:10.1016/j.ajhg.2016.05.006 (2016).
235 Mahajan, A. *et al.* Trans-ethnic Fine Mapping Highlights Kidney-Function Genes Linked to Salt Sensitivity. *Am J Hum Genet* **99**, 636-646, doi:10.1016/j.ajhg.2016.07.012 (2016).
236 Replication, D. I. G. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* **46**, 234-244, doi:10.1038/ng.2897 (2014).
237 van Rooij, F. J. *et al.* Genome-wide Trans-ethnic Meta-analysis Identifies Seven Genetic Loci Influencing Erythrocyte Traits and a Role for RBPMS in Erythropoiesis. *Am J Hum Genet* **100**, 51-63, doi:10.1016/j.ajhg.2016.11.016 (2017).
238 Reid, M. E., Lomas-Francis, C., Olsson, M. L. & ebrary Inc. xiii, 561 p. (Elsevier/Academic Press,, Amsterdam ; Boston, 2012).
239 Möller, M., Jöud, M., Storry, J. R. & Olsson, M. L. Erythrogene: a database for in-depth analysis of the extensive variation in 36 blood group systems in the 1000 Genomes Project. *Blood Advances* **1**, 240-249, doi:10.1182/bloodadvances.2016001867 (2016).
240 Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607, doi:10.1038/nature11003 (2012).
241 Keller, M. A. *et al.* Transcriptional regulatory network analysis of developing human erythroid progenitors reveals patterns of coregulation and potential transcriptional regulators. *Physiol Genomics* **28**, 114-128, doi:10.1152/physiolgenomics.00055.2006 (2006).
242 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**, 2081-2088, doi:10.1182/blood-2008-03-145672 (2008).
243 Chandanayingyong, D., Sasaki, T. T. & Greenwalt, T. J. Blood groups of the Thais. *Transfusion* **7**, 269-276 (1967).
244 Alfred, B. M., Stout, T. D., Lee, M., Birkbeck, J. & Petrakis, N. L. Blood groups, phosphoglucomutase, and cerumen types of the Anaham (Chilcotin) Indians. *Am J Phys Anthropol* **32**, 329-337, doi:10.1002/ajpa.1330320303 (1970).
245 Wieckhusen, C. *et al.* Molecular Screening for Vel- Blood Donors in Southwestern Germany. *Transfus Med Hemother* **42**, 356-360, doi:10.1159/000440791 (2015).
246 Dezan, M. R. *et al.* High-throughput strategy for molecular identification of Vel-blood donors employing nucleic acids extracted from plasma pools used for viral nucleic acid test screening. *Transfusion* **56**, 1430-1434, doi:10.1111/trf.13572 (2016).
247 Danger, Y. *et al.* Characterization of a new human monoclonal antibody directed against the Vel antigen. *Vox Sang* **110**, 172-178, doi:10.1111/vox.12321 (2016).

248 Lemmon, M. A., Flanagan, J. M., Treutlein, H. R., Zhang, J. & Engelman, D. M. Sequence specificity in the dimerization of transmembrane alpha-helices. *Biochemistry* **31**, 12719-12725 (1992).
249 Bruce, L. J. *et al.* A band 3-based macrocomplex of integral and peripheral proteins in the RBC membrane. *Blood* **101**, 4180-4188, doi:10.1182/blood-2002-09-2824 (2003).
250 Dobson, L., Remenyi, I. & Tusnady, G. E. CCTOP: a Consensus Constrained TOPology prediction web server. *Nucleic Acids Res* **43**, W408-412, doi:10.1093/nar/gkv451 (2015).
251 Reynolds, S. M., Kall, L., Riffle, M. E., Bilmes, J. A. & Noble, W. S. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS Comput Biol* **4**, e1000213, doi:10.1371/journal.pcbi.1000213 (2008).
252 Kall, L., Krogh, A. & Sonnhammer, E. L. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* **338**, 1027-1036, doi:10.1016/j.jmb.2004.03.016 (2004).
253 Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**, 567-580, doi:10.1006/jmbi.2000.4315 (2001).
254 Goder, V. & Spiess, M. Topogenesis of membrane proteins: determinants and dynamics. *FEBS Lett* **504**, 87-93 (2001).
255 Duraisingh, M. T. *et al.* Phenotypic variation of Plasmodium falciparum merozoite proteins directs receptor targeting for invasion of human erythrocytes. *EMBO J* **22**, 1047-1057, doi:10.1093/emboj/cdg096 (2003).
256 Sahar, T. *et al.* Plasmodium falciparum reticulocyte binding-like homologue protein 2 (PfRH2) is a key adhesive molecule involved in erythrocyte invasion. *PLoS One* **6**, e17102, doi:10.1371/journal.pone.0017102 (2011).
257 Solyakov, L. *et al.* Global kinomic and phospho-proteomic analyses of the human malaria parasite Plasmodium falciparum. *Nat Commun* **2**, 565, doi:10.1038/ncomms1558 (2011).
258 Feng, Z. M., Wu, A. Z., Zhang, Z. & Chen, C. L. GATA-1 and GATA-4 transactivate inhibin/activin beta-B-subunit gene transcription in testicular cells. *Mol Endocrinol* **14**, 1820-1835, doi:10.1210/mend.14.11.0549 (2000).
259 Ulirsch, J. C. *et al.* Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* **165**, 1530-1545, doi:10.1016/j.cell.2016.04.048 (2016).
260 Merika, M. & Orkin, S. H. Functional synergy and physical interactions of the erythroid transcription factor GATA-1 with the Kruppel family proteins Sp1 and EKLF. *Mol Cell Biol* **15**, 2437-2447 (1995).
261 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**, 221-228, doi:10.1002/humu.22218 (2013).
262 Tallack, M. R. *et al.* Novel roles for KLF1 in erythropoiesis revealed by mRNA-seq. *Genome Res* **22**, 2385-2398, doi:10.1101/gr.135707.111 (2012).

263 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**, 721-727, doi:10.1016/j.ajhg.2010.10.010 (2010).