LUND UNIVERSITY

Natural genetic variation as a discovery platform for human immunology and hematopoiesis

Ali, Zain

2026

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

Link to publication

Citation for published version (APA):
Ali, Z. (2026). *Natural genetic variation as a discovery platform for human immunology and hematopoiesis*. [Doctoral Thesis (compilation), Department of Laboratory Medicine]. Lund University, Faculty of Medicine.

Total number of authors:
1

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

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

Download date: 08. okt.. 2026

Natural genetic variation as a discovery platform for human immunology and hematopoiesis

Natural genetic variation as a discovery platform for human immunology and hematopoiesis

Zain Ali

LUND
UNIVERSITY

DOCTORAL DISSERTATION

Doctoral dissertation for the degree of Doctor of Philosophy in Laboratory Medicine, specialisation Experimental Haematology, at the Faculty of Medicine at Lund University to be publicly defended on 30 October 2026 at 13.00 in Segerfalksalen, BMC A10, Sölvegatan 17, Lund

Faculty opponent
Professor Qiang Pan-Hammarström, Karolinska Institutet, Stockholm

Organization: LUND UNIVERSITY

Document name: DOCTORAL DISSERTATION **Date of issue** 2026-10-30

Author(s): Zain Ali **Sponsoring organization:**

Title and subtitle: Natural genetic variation as a discovery platform for human immunology and hematopoiesis

Abstract:

Blood and immune cell traits vary continuously between healthy individuals and are largely heritable, making them tractable to genome-wide association analysis. The variants they implicate are mostly non-coding, and their target genes, mechanisms and cell types of action are usually unknown. This thesis identifies the causal variants, genes and cell types behind such associations for two blood traits at opposite ends of the hematopoietic hierarchy: circulating $CD34^+$ cell levels and serum immunoglobulin levels. Signals were fine-mapped, interpreted against sorted-cell expression and chromatin data, and tested by reporter assay, factor knockdown and genome editing.

Blood CD34+ cell levels had not previously been subjected to genome-wide analysis; in 13,167 individuals, eight loci were implicated. The strongest resolved to an intronic variant that abolishes a binding site for the transcription factor MYB and lowers expression of PPM1H, a phosphatase with no prior role in hematopoiesis. Reducing *PPM1H* raises stem cell output, making it a candidate mobilization target. A second locus resolved to a variant removing a GATA2 site from a progenitor enhancer.

Nine immunoglobulin traits measured in 114,697 individuals yielded 504 associations, most of them new. Together they describe the B cell system stage by stage, from lineage commitment to antibody clearance, and implicate genes with no described role in immunity. The largest effects fall on a few control points where series of alleles tune antibody output continuously; at the receptor TNFRSF13B (TACI) and its ligand TNFSF13 (APRIL) these run from immunodeficiency-causing variants to one raising plasma cell cancer risk.

Quantitative traits measured in unselected populations therefore act as forward genetic screens of human tissue, recovering the regulators of normal physiology together with genes in which more severe variation causes disease.

Key words: GWAS, gene regulation, hematopoietic stem cell, immunoglobulins, B cell

Classification system and/or index terms (if any) Supplementary bibliographical information

Language English **Number of pages**: 90

ISSN and key title: 1652-8220, Lund University, Faculty of Medicine Doctoral Dissertation Series

ISBN: 978-91-8021-921-1

Recipient's notes Price Security classification

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

Signature Date 2026-09-16

Natural genetic variation as a discovery platform for human immunology and hematopoiesis

Zain Ali

LUND
UNIVERSITY

Copyright

Pages 1-90 © 2026 Zain Ali

Paper I © 2022 The American Society of Hematology. Published in Blood.

Paper II © 2026 The authors. Published by Springer Nature in Scientific Reports (licensed under CC BY 4.0).

Paper III © 2026 The authors (manuscript unpublished)

Cover image by the author

Published by:

Department of Laboratory Medicine

Faculty of Medicine

Lund University

Lund 2026

ISBN 978-91-8021-921-1 (print)

Lund University, Faculty of Medicine Doctoral Dissertation Series 2026:123

ISSN 1652-8220

Printed in Sweden by Media-Tryck, Lund University, Lund, 2026

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

MADE IN SWEDEN

Dedicated to Amani and Arani, who will conquer the world

Table of Contents

List of Papers 10
Abbreviations 11
Popular scientific summary 12
Declaration on the use of generative AI 13

Chapter 1 – Conceptual background 14
The infinitesimal model and the foundations of modern genetics 14
The molecular basis of biological complexity 15
The Human Genome Project and GWAS 17
The future of human genetics 19
The hematopoietic system 20
The B cell and the antibody response 23

Chapter 2 – Aims 26

Chapter 3 – Methodological considerations 27
3.1 Genome-wide association and statistical fine-mapping 27
3.2 Expression quantitative trait analysis 29
3.3 Colocalization 30
3.4 Trans associations and inverse-variance weighted regression 31
3.5 Chromatin accessibility and ATAC-sequencing 33
3.6 Heritability partitioning and variant-set enrichment 35
3.7 Single-cell resolution of cell state 35
3.8 Prediction of transcription factor binding disruption 36
3.9 Reporter assays and the dissection of regulatory dependence 37
3.10 Genome editing 38
3.11 Allelic imbalance 39
3.12 Ethical considerations 39

Chapter 4 – Results 41
4.1 Paper I: The genetic architecture of the blood CD34+ cell compartment 41

4.2 Paper II: Functional dissection of the 1p36.23/ENO1-RERE locus .. 43
4.3 Paper III: The genetic architecture of the human B-cell system 46
Chapter 5 – Discussion 55
5.1 Quantitative variation as a discovery platform for clinical genetics ... 55
5.2 Discovery of novel B-cell genes and pathways 57
5.3 Trait engineering as a general principle 58
5.4 Systems-level properties: buffering and compensation 59
5.5 Forward genetics with natural variation as the mutagen 60
5.6 Quantitative traits and disease gene discovery 62
5.7 Therapeutic leverage 63
5.8 Limitations 64
5.9 Contribution 66
Chapter 6 – Conclusions and future perspectives 68
6.1 Conclusions 68
6.2 Completing the mechanistic chain for novel genes 68
6.3 Testing variants and elements in parallel 69
6.4 Further dissection of regulatory organization 69
6.5 Clinical follow-up 70
6.6 Closing remarks 70
References 72
Acknowledgements 89

List of Papers

Paper I. Lopez de Lapuente Portilla A, Ekdahl L, Cafaro C, **Ali Z**, Miharada N, Thorleifsson G, Zemaitis K, Lamarca Arrizabalaga A, Thodberg M, Pertesi M, Dhapola P, Bao E, Niroula A, Bali D, Norddahl G, Ugidos Damboriena N, Sankaran VG, Karlsson G, Thorsteinsdottir U, Larsson J, Stefansson K, Nilsson B. Genome-wide association study on 13 167 individuals identifies regulators of blood CD34+ cell levels. *Blood*. 2022;139(11):1659-1669.

Paper II. **Ali Z**, Cafaro C, Ekdahl L, Lamarca A, Lopez de Lapuente Portilla A, Nilsson B. Functional dissection of the gene regulatory mechanism underlying variation in blood CD34+ cell levels at 1p36.23/ENO1. *Scientific Reports*. 2026;16:558.

Paper III. **Ali Z**, Lopez de Lapuente Portilla A, Thorleifsson G, Lamarca-Arrizzabalaga A, Cafaro C, Halldorsson GH, Ekdahl L, Ota M, Melsted P, Stefansdottir L, Jonsdottir A, Sigurdsson A, Ivarsdottir E, Fujio K, Harding SJ, Ludviksson BR, Love TJ, Kristinsson SY, Sulem P, Gudbjartsson DF, Stefansson K, Thorsteinsdottir U, Jonsdottir I, Olafsdottir T, Nilsson B. Population-scale immunoglobulin genetics resolves the human B-cell system. Manuscript, under review. Preprint: bioRxiv 2026.03.30.715216.

Abbreviations

A/G, A/M, G/M	Ratios of serum IgA to IgG, IgA to IgM, and IgG to IgM
AG, AGM	Products of IgA and IgG, and of IgA, IgG and IgM
AICDA	Activation-induced cytidine deaminase
ATAC-sequencing	Assay for transposase-accessible chromatin with sequencing
CD34	Cluster of differentiation 34
ChIP-sequencing	Chromatin immunoprecipitation with sequencing
ClinVar	NCBI archive of variant-phenotype relationships
CRISPR	Clustered regularly interspaced short palindromic repeats
CVID	Common variable immunodeficiency
DLBCL	Diffuse large B-cell lymphoma
DNA	Deoxyribonucleic acid
DNase-sequencing	Deoxyribonuclease I hypersensitive site sequencing
ELISA	Enzyme-linked immunosorbent assay
ENCODE	Encyclopedia of DNA Elements
eQTL	Expression quantitative trait locus
GTEx	Genotype-Tissue Expression project
GWAS	Genome-wide association study
HLA	Human leukocyte antigen
HSPC	Hematopoietic stem and progenitor cell
Ig	Immunoglobulin
IgA, IgD, IgE, IgG, IgM	Immunoglobulin isotypes A, D, E, G and M
IgG1 to IgG4	Immunoglobulin G subclasses 1 to 4
IGH	Immunoglobulin heavy-chain locus
ImmuNexUT	Immune Cell Gene Expression Atlas from University of Tokyo
IVW	Inverse-variance weighted
LD	Linkage disequilibrium
M/AG	Ratio of serum IgM to the product of IgA and IgG
MPRA	Massively parallel reporter assay
PBMC	Peripheral blood mononuclear cell
pQTL	Protein quantitative trait locus
RNA	Ribonucleic acid
RNA-sequencing	Ribonucleic acid sequencing
siRNA	Short interfering RNA
SNP	Single nucleotide polymorphism
SuSiE	Sum of single effects regression
WHIM	Warts, Hypogammaglobulinemia, Immunodeficiency and Myelo-kathexis syndrome
3'RR1	3' regulatory region 1 of the immunoglobulin heavy-chain locus

Popular scientific summary

Everyone differs a little in the measurements a doctor takes from a blood sample. Some people have more of a given blood cell type than others, or more of a given circulating protein, and much of that difference is inherited from their parents. These differences are not usually large enough to make anyone ill. They are simply the normal range of human variation.

Experimental biologists have a standard way of finding out which genes build a biological system. They damage genes at random in an animal such as a fruit fly, look for animals in which the system has gone wrong, and then work out which gene was damaged in each case. This approach identified the genes that lay out the body plan of an embryo. It works because the experimenter controls both the damage and the breeding, and neither is possible in humans.

Human populations, however, have already performed a version of this experiment over thousands of years. Every person inherits several million small differences in their DNA sequence, and across a large enough group of people almost every gene in the genome has been altered in somebody. These alterations are often milder than the ones a laboratory would introduce, because natural selection has removed the most damaging over many generations, but they are real perturbations acting in human tissue. If a clinical measurement (such as the blood tests you routinely receive in the hospital) is available for enough people, and their DNA sequence has been analyzed in sufficient detail, the result of that experiment can be recovered.

This thesis applies that idea to two types of measurements taken from blood: the number of stem cells circulating in the bloodstream, and the amount of antibodies in their serum. The first was analyzed in 13,167 people and the second in 114,697. In each case, inherited differences in the measurement could be traced back to the particular genes that set it, and many of those genes had no previously known role in these processes, adding to our biological understanding of stem cells and antibodies.

Both studies find similar results and uncover new genes controlling stem cell and antibody production. In each study, the genes that explain why one healthy person's blood test reads differently from another's turned out to be genes that cause real illness when they are more badly disrupted. The stem cell work pointed to a gene called *PPM1H* that could be worth turning into a drug. The antibody work laid out the whole set of genes that decides how much antibody a person carries, from how the antibody-producing cells develop to how long the finished antibody survives once it is in the blood. One of them is a gene that rescues antibodies from being broken down and puts them back into circulation, which is the same mechanism drug companies exploit to make antibody-based medicines last longer in patients. Ordinary human variation and disease are the same biology at different strengths.

Declaration on the use of generative AI

The text of this thesis summary has been partly produced using the generative AI model Claude (Anthropic). I used it to produce draft text and to help organize arguments; every passage produced this way was checked against the primary sources and revised before inclusion. In the research itself I used the same model to help write Python code for plotting and data analysis, in the later part of my doctoral studies only, as such tools were not available for most of the period over which this work was carried out. Where code was drafted with assistance it was read, tested and checked against the underlying data in the same way as code written from scratch, and the analytical decisions it implements were mine. No result, figure or interpretation in this thesis was generated by a model. I have processed all generated text and take full responsibility for the content.

Chapter 1 – Conceptual background

The infinitesimal model and the foundations of modern genetics

> *"The law of combination of different characters which governs the development of the hybrids finds therefore its foundation and explanation in the principle enunciated, that the hybrids produce egg cells and pollen cells which in equal numbers represent all constant forms which result from the combinations of the characters brought together in fertilization."*
>
> - Gregor Mendel, "Experiments in Plant Hybridization" (1865)

Darwin failed to derive a working theory of inheritance. The common explanation for familial resemblance at the time was "blending", the idea that offspring traits would be an average of both parents – a tall father and a short mother would produce children of intermediate heights, a red flower crossed to a white flower would produce pink petals. However, such a theory is logically incompatible with the idea that environmental pressures would lead to the highly specialized forms that arise from natural selection, such as the exquisitely adapted beaks of the Galapagos finches. Blending inheritance should cause a "regression to the mean" in breeding populations, swamping out the inherited variation that selection needs to act on[1].

Mendel's breeding experiments with peas arrived at the truth of inheritance – gametes contained discrete particles that we now call alleles, and the combination of these discrete particles during fertilization gave rise to predictable ratios of traits in the offspring. Although a contemporary of Darwin, Mendel's work would be forgotten for decades and then rediscovered with much interest in the early 20th century.

The discovery of chromosomes and the paired way they separated during meiosis in the early 20th century provided a physical basis for Mendel's alleles – they resided on chromosomes[2,3]. Subsequently, Thomas Hunt Morgan's experiments with fruit flies proved that Mendel's heritable factors resided physically on specific locations (loci) on chromosomes and led to the development of linkage mapping, a method still widely used in modern genetics[4].

By the early 20th century Mendel's work had been fully rediscovered, and a community of committed Mendelians had emerged. However, it had still not been fully reconciled with Darwinian evolution. Darwinian evolution is a theory of gradual change over geological timescales, yet the traits being mapped by the early Mendelians were qualitative and binary – white flower petals in peas, red eyes in fruit flies.

Could evolution proceed by these phenotypic leaps and bounds instead of the gradual change over geological time proposed by Darwin's theory of evolution? Hugo de Vries, a botanist and pioneer of early genetics, rejected Darwinian gradualism, proposing an alternative theory called saltationism where large-effect mutations made new species and evolution proceeded by sudden leaps rather than graduallly[5].

The experimental evidence for discrete particles of inheritance located on chromosomes seemed overwhelming, but it could not be fully reconciled to the forms of heritable variation observed in natural populations. Most traits in natural populations are not binary or discrete but are continuous and normally distributed. How could Mendel's discrete particles give rise, for example, to the smooth Gaussian distribution of heights observed in human populations?

The key theoretical insight came from the work of Ronald Fisher, a foundational figure for both modern statistics and genetics. In his 1918 paper *"The Correlation between Relatives on the Supposition of Mendelian Inheritance"* he formulated a mathematical model where a large but finite number of Mendelian alleles contributed additively and independently to inherited traits, giving rise to the continuous, Gaussian distributions observed for quantitative traits in nature[6]. Darwinian evolution could then be reframed in terms of this theory: evolution arises due to changes in *population* allele frequencies due to natural selection. The gradualism observed in the paleontological record could then be reconciled to the empirical observations being made by cytologists and in the controlled breeding experiments of geneticists, and a complete theory of biology could be formulated many years before the discovery of the structure of DNA and the emergence of modern molecular biology.

In my doctoral work I conducted and interpreted genome-wide association studies of quantitative traits in human populations; work that was only possible due to the theoretical foundations laid by quantitative geneticists like Fisher.

The molecular basis of biological complexity

> "*Embryos derived from bicoid (bcd) females lack head and thorax. Cytoplasmic transplantation experiments reveal that bcd$^+$-dependent activity is localized at the anterior pole of wild-type embryos. The activity can induce anterior development in mutant embryos at any position along the antero-posterior axis and suppresses posterior development.*"

- Hans Georg Frohnhöfer & Christiane Nüsslein-Volhard, "Organization of anterior pattern in the Drosophila embryo by the maternal gene *bicoid*" (*Nature*, 1986)

Although classical genetics could locate Mendel's particles of inheritance on chromosomes, how these stretches of DNA give rise to complex biological systems remains an open question to this day. The discovery of the structure of DNA and its innate molecular capacity for self-replication ushered in the era of molecular biology, transforming genes from abstract units of inheritance into discrete molecular machines that seemed able to animate all the properties of a cell – the contractile motions of the cytoskeleton, the carefully regulated trafficking of material between cell compartments, the production of energy in chloroplasts and mitochondria, the replication of DNA, the translation of proteins by ribosomes[7].

Yet the gap between molecular understanding and biological complexity is vast and remains a largely open question to this day.

During animal development, a single fertilized egg cell gives rise to a fully patterned animal body plan with functioning organs. Using only our 20th century knowledge of genetics and molecular biology, we can conclude that the information needed to carry this out must be encoded in genes residing in chromosomes. But how does a protein encode the concept of an eye? If the scale of complexity in a single cell is hard enough to fathom, extending our molecular understanding into multicellular systems like embryos and brains seems insurmountable.

Experimental embryologists in the early 20th century began to elucidate some of the working principles of animal development – as early as 1905 the zoologist Edwin Conklin performed lineage tracing experiments in sea squirt embryos, observing a region of yellow pigmented cytoplasm that segregated predictably into tail muscle cells and acted as a natural lineage marker[8]. In contrast to the nuclear chromosomal factors of Mendelian inheritance, this early work pointed to the importance of cytoplasmic factors that determined the identity of cells and tissues.

Attempts to isolate these patterning factors using biochemical methods failed, but a breakthrough came in 1980 when Christiane Nüsslein-Volhard and Eric Wieschaus applied an alternative approach: forward genetic screens of random mutants in *Drosophila* larvae[9]. Forward genetic approaches start by measuring phenotypes – whether generated by mutagenesis or by natural variation – and work backwards to map and isolate the underlying genes through different forms of genetic analysis such as linkage mapping.

At the time the first screens for developmental mutants were carried out in *Drosophila* it was not clear that embryology would yield to such an approach – what if such processes were too complex, interlinked and buffered to yield interpretable single gene mutations? But the opposite was true – the initial screen identified 15 distinct gene mutations that altered the segmental pattern of the *Drosophila* larvae in a predictable, orderly fashion[10]. The patterning defects could also be placed into three

distinct categories that revealed different layers of organization – some affected contiguous blocks of segments while others affected alternating segments, for example. Thus, embryonic patterning is a *modular* system driven by a finite set of molecular machines encoded by genes[11]. More importantly, these results demonstrated that biological complexity could be genetically tractable – a phenotypic screen requires no prior knowledge of the underlying molecular biology, yet it can still enable the discovery of the molecular components that give rise to an observable trait.

Subsequent work helped to further define the molecular identities of these 15 genes – 10 of them were sequence-specific transcription factors, while the remainder were ligands and receptors for the Wnt and Hedgehog signaling pathways, which continue to be studied extensively in many areas of biology[12]. In our large-scale genetic analysis of human hematological and immunological traits, we observed a similar convergence towards transcription factors and signaling molecules.

In my doctoral work we probed the development of blood stem cells and the immune system through a forward genetics approach, but instead of using mutagenesis in model organisms we used natural variation in human populations. The conclusions we came to were much the same as those that emerged from the Heidelberg screens of *Drosophila* embryogenesis – biological complexity can be reduced to a finite set of molecular modules and is exquisitely tractable to forward genetic analysis[13].

The Human Genome Project and GWAS

> *"Five lines of evidence point to an increase in the complexity of the proteome from the single-celled yeast to the multicellular invertebrates and to vertebrates such as the human. Specifically, the human contains greater numbers of genes, domain and protein families, paralogues, multidomain proteins with multiple functions, and domain architectures. According to these measures, the relatively greater complexity of the human proteome is a consequence not simply of its larger size, but also of large-scale protein innovation."*
>
> - International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome" (Nature, 2001)

The forward genetic screens carried out in Heidelberg were performed on a model organism, allowing the experimenter to carry out controlled mutagenesis and breeding experiments. By scoring phenotypes in thousands of mutagenized embryos, they could saturate the entire genome with informative mutations. None of this is possible in humans, who cannot be mutagenized or crossed at will, and whose generation time is measured in decades rather than days.

For most of the 20th century human genetics was therefore confined to rare Mendelian disorders, traits caused by single genes of large effect that revealed themselves

through patterns of inheritance in family pedigrees according to the laws of Mendelian genetics. Linkage mapping, the same method that emerged from Thomas Hunt Morgan's fly experiments, could trace such genes to their location on a chromosome by following how a disease co-segregated with genetic markers through a family[14–17]. This worked well for conditions like cystic fibrosis and Huntington's disease, but it was powerless against the traits Fisher had explained. Common, continuously distributed phenotypes are not caused by single variants of large effect but by many variants of small effect, and their heritability is spread too thinly across the genome for any one family to reveal it.

Thus, to find genes of small effect one cannot follow a single locus through a single pedigree. One must compare allele frequencies across very large numbers of unrelated individuals and ask which positions in the genome differ systematically between people at the high and low ends of a trait. This is *association* rather than *linkage*, and it required two things that did not exist before the turn of the millennium: a complete reference sequence of the genome to serve as a coordinate system, and a catalog of the common variation that distinguishes one person from another. The Human Genome Project supplied the coordinate system and sequence map that make modern genetic analysis possible. Completed in 2003 after 13 years and roughly three billion dollars, it produced the reference against which every subsequent variant could be named and located[18–20].

The catalog of common variation came from the International HapMap Project, which set out to characterize the common single nucleotide polymorphisms (SNPs) in human populations and the way they are inherited together[21,22]. Because recombination shuffles the genome unevenly, nearby variants tend to be inherited as blocks – a correlation structure known as linkage disequilibrium (LD). Rather than sequencing every base in every participant, which remained prohibitively expensive, one could therefore genotype a few hundred thousand carefully chosen marker SNPs on a microarray and rely on LD so that each typed marker also reported on its untyped neighbors[23–25].

With these tools in place, the genome-wide association study (GWAS) became possible. The first convincing example arrived in 2005, when a common variant in the complement factor H gene was associated with age-related macular degeneration in a sample of fewer than 200 people[26]. The approach reached maturity in 2007 with the Wellcome Trust Case Control Consortium, which examined seven common diseases in a single shared design that set the basic template for the field[27]. Robust genetic associations with a plethora of traits and diseases accumulated quickly into thousands, and two features of them stood out. They were overwhelmingly polygenic, each variant explaining only a small fraction of the variance in a trait, and they fell predominantly outside protein-coding sequence, in the regulatory regions of the genome[28–30].

The work of this thesis applies this logic to hematopoiesis and the immune system. Immunoglobulin (Ig) levels and circulating stem-cell levels are quantitative traits in the Fisherian sense, varying continuously across the population and under substantial genetic control, which makes them well suited to a genome-wide association approach. Analyzing nine quantitative Ig traits in a large population sample, we fine-mapped 504 genetic associations. We recovered many known regulators of the B-cell lineage and implicated many that had not previously been reported. As in the Heidelberg screens, these signals converged on transcription factors and signaling molecules, in this case regulators of B-cell development and antibody production. Thus, a forward genetic dissection of human immunity through natural variation arrived at the same conclusion as the screens of the fly embryo: a complex biological system is built from a finite and discoverable set of molecular modules.

The future of human genetics

> *"There are deep challenges to fully understanding the impact of very small effects in organismal systems, so we believe there is great need to develop cell-based model systems that can recapitulate aspects of complex traits."*
>
> - Evan Boyle, Yang Li & Jonathan Pritchard, "An expanded view of complex traits: from polygenic to omnigenic" (Cell, 2017)

A genome-wide association study tells us which DNA sequence variants correlate with trait variation, but associations on their own cannot reveal molecular mechanism. The vast majority of variants identified in GWAS are non-coding – the signal can point to a stretch of regulatory DNA containing many correlated variants, any one of which might be the causal one, acting on a gene that is not necessarily the nearest, in a cell type the study cannot itself identify. How does a stretch of non-coding DNA reach forward through the molecular machinery of the cell to shape a phenotype? GWAS has been remarkably successful at telling us where to look. The harder problem of uncovering the underlying molecular mechanism is only beginning to be solved.

This is the central problem of post-GWAS genetics, and it is currently being approached from two directions.

Statistical fine-mapping uses the correlation structure of a locus to narrow a broad association down to a small credible set of variants that plausibly contain the true causal variant(s)[31]. Statistical fine-mapping relaxes the assumption that genetic variants are independent of each other and performs joint modeling, often using Bayesian approaches, of the likelihood of *combinations* of multiple alleles being causal for an observed shift in trait values.

Functional genomics then asks what those variants do on a molecular level[32]. Maps of open chromatin, transcription factor binding, and gene expression, increasingly resolved to single cells and to the particular states a cell passes through, allow a candidate variant to be placed in the regulatory context in which it acts, linked to the gene it most likely controls and the cell type in which it controls it[33,34]. Where these maps converge on the same variant, gene, and cell, a mechanistic hypothesis emerges from what began as nothing more than a statistical coordinate in the human reference genome.

The fruit of this approach is not only mechanistic understanding of the molecular basis of trait variation, but knowledge with genuine clinical impact. Drug targets supported by human genetic evidence succeed in clinical trials far more often than those without it[35–37]. Combining information from these same variants yields polygenic scores that estimate an individual's genetic liability to a disease, with a growing role in clinical risk prediction[38,39]. The catalog of GWAS associations is gradually becoming a catalog of therapeutic and diagnostic hypotheses and will become increasingly relevant for clinical practice as DNA sequencing becomes more accessible.

My doctoral research sits at this post-GWAS transition. Having mapped the genetic architecture of Ig and stem-cell levels, the natural next step is to ask what the strongest signals are doing on a molecular level. We have begun to resolve individual loci to their causal variants and to test them experimentally. I find in this a continuity with the questions that drew me to biology in the first place. The screens of the fly embryo showed that biological complexity, however daunting, can be reduced to a finite set of molecular modules. The genetics of the human hematopoietic system tell the same story.

The hematopoietic system

> *"The study of human hematopoiesis is often limited by the inability to manipulate this process in vivo and differences that exist between humans and commonly employed model organisms. However, human genetics provides a way to gain insight into natural variation in a variety of hematologic phenotypes."*
>
> - Vijay Sankaran & Stuart Orkin, "Genome-wide association studies of hematologic phenotypes: a window into human hematopoiesis" (*Current Opinion in Genetics & Development,* 2013)

Blood is the most prolific tissue in the human body. Hundreds of billions of blood cells are produced every day, and the entire complement of erythrocytes, platelets, and white blood cells is replaced continuously throughout life, all of them descended from a rare pool of stem cells. Of all the systems in which one might attempt the

forward genetics described in the previous sections, hematopoiesis is among the most tractable in humans. It is continuously accessible through a simple blood draw, its outputs are routinely quantified in clinical settings, and it is arranged in a hierarchy whose structure has been mapped in detail[40,41].

Till and McCulloch first demonstrated, by counting clonal colonies in the spleens of irradiated mice, that single cells could regenerate the blood, defining the hematopoietic stem cell operationally by its two essential properties of self-renewal and multipotency[42]. The functional definition of the human stem cell came from the work of John Dick and colleagues, who transplanted human cells into immunodeficient mice and scored their capacity to reconstitute hematopoiesis, resolving the hierarchy of lineage-restricted progenitors[43,44]. The same framework revealed that leukemias are themselves hierarchically organized, sustained by a leukemic stem cell, a finding that reshaped how blood cancers are understood[45]. The classical picture that emerged is a branching tree, in which the stem cell gives rise through a series of increasingly restricted progenitors to the committed myeloid and lymphoid lineages. Single-cell measurements have since revised this tree toward a more continuous landscape of priming and commitment, but the hierarchy remains the organizing concept of the field and provides the language in which its cell types are named[46].

These cells are identified by the surface glycoprotein CD34[47]. Hematopoietic stem and progenitor cells (HSPCs) make up less than 0.1% of the mononuclear cells in blood, and it is CD34 expression that allows them to be counted and sorted by flow cytometry. CD34 labels a mixed population of stem and progenitor cells and is a critical marker for clinical stem-cell transplantation: $CD34^+$ cells are the fundamental unit of stem-cell transplantation and the baseline blood $CD34^+$ level is one of the better predictors of successful transplantation[48,49].

Hematopoietic stem-cell transplantation is a cornerstone of treatment for blood malignancies, and modern grafts are collected not from the bone marrow but from peripheral blood by leukapheresis. HSPCs, however, normally reside in the marrow niche, retained there by the chemokine axis in which CXCL12 produced by stromal cells binds the receptor CXCR4 on the stem-cell surface[50]. To be harvested in the clinic, they must first be mobilized into the circulation, which is achieved with granulocyte colony-stimulating factor and, when that is insufficient, with the CXCR4 antagonist plerixafor[51,52]. Mobilization varies between individuals and a minority of patients mobilize so poorly that an adequate graft cannot be collected[53]. The genetic regulation of blood $CD34^+$ levels is therefore not only a question about stem-cell biology but a clinical one. A genetic variant that alters a person's baseline $CD34^+$ level is, in effect, running a natural experiment in mobilization, and a regulator discovered this way is a candidate drug target for improving it.

Human hematopoiesis is well-suited for forward genetic analysis – blood counts are routinely measured in the clinic, and because the lineage hierarchy is known, the cell type in which a variant acts can be inferred using cells sorted by lineage markers

or, more recently, using single-cell methods. Large studies such as that of Astle and colleagues mapped thousands of associations with mature blood-cell traits across the lineages[54–56]. The challenge in blood was the same one that defines the post-GWAS era everywhere: the move from a statistical association to the underlying cellular and molecular mechanism.

The tools to enable functional genetic interpretation in human hematopoiesis were assembled over the past decade. Corces, Buenrostro, and colleagues generated chromatin accessibility maps by ATAC-sequencing across sorted hematopoietic populations, building an atlas that ultimately spans 18 progenitor and mature cell types[57]. Ulirsch, Lareau, Bao, and Sankaran then applied these data to the interpretation of GWAS associations – fine-mapping the associations for 16 blood-cell traits in roughly 115,000 individuals from the UK Biobank to credible sets of plausible causal variants, overlapping those variants with the accessibility atlas to assign a cell type, identifying the disrupted transcription factor by the combination of motif prediction and measured occupancy, and nominating target genes through correlation between accessibility and expression[58]. They further introduced gchromVAR, a method that uses fine-mapped posterior probabilities together with quantitative accessibility to resolve which of several closely related cell states a variant acts in. This analysis converged on the binding sites of the master regulators of hematopoiesis – GATA1, RUNX1, PU.1 – factors that classical biology had already identified as major lineage determining factors in blood.

The efforts above mapped the mature blood-cell counts that are commonly assayed in the clinic – the erythrocytes, platelets, and leukocytes at the bottom of the hierarchy. The stem and progenitor compartment itself, the blood $CD34^+$ level, had never been subjected to a genome-wide association study, despite its central importance to transplantation. Our lab has built a platform for high-throughput, multicolor flow cytometry in human blood samples, and by coupling this with genetic association analysis we are able to assay highly resolved blood traits invisible to standard clinical tests – including circulating $CD34^+$ levels.

In a study of 13,167 individuals, we reported the first genetic associations with blood $CD34^+$ cell levels measured by flow cytometry of blood samples, identifying nine significant and two suggestive associations, accounted for by eight loci: *PPM1H*, *CXCR4*, *ENO1-RERE*, *ITGA9*, *ARHGAP45*, *CEBPA*, *TERT*, and *MYC*[59]. The recovery of *CXCR4*, with four independent associations at an approved drug target for stem-cell mobilization, was proof of principle that genuine regulators of the stem-cell compartment can be read out from natural variation in humans. Alongside *CXCR4* were genes never previously implicated in stem-cell biology, *PPM1H*, *ENO1*, *RERE*, and *ARHGAP45*, together with *ITGA9*, which had been only minimally studied in this context.

The most significant association mapped to *PPM1H*, a serine and threonine phosphatase that had never been implicated in hematopoiesis. My contribution to this

study was the functional dissection of this locus. The allele that confers higher blood $CD34^+$ levels was found to downregulate *PPM1H*, and the causal factor was localized to a single non-coding variant, rs772557-A, which abrogates a binding site for the master hematopoietic transcription factor MYB in the first intron of the gene. A companion study, which forms the second paper of this thesis, took one of the other novel loci further and found the same kind of mechanism acting through a different factor[60]. At the *ENO1-RERE* locus, the causal non-coding variant, rs10864368-T, abrogates a binding site for GATA2, the master regulator of the stem and progenitor state. Two loci from a single stem-cell GWAS thus resolve to the same kind of event, a common variant that creates or destroys a binding site for a master regulator and thereby tunes the output of the hierarchy.

The forward-genetic dissection of the stem-cell compartment can thus recover the master regulators of hematopoiesis from human variation, as well as revealing novel targets for clinical perturbation.

The B cell and the antibody response

The B-cell lineage produces Ig, a single family of secreted proteins whose diversity is large enough to bind essentially any molecular shape an organism might encounter[61]. The plasma cell that secretes antibody is one of the most specialized cells in the body and results from a complex and orderly developmental program[62]. The genetic analysis of circulating Ig levels – the final product of this developmental program – is the subject of the third and largest paper in this thesis[63].

After birth, B-cell development takes place in the bone marrow, along the lymphoid branch that descends from the hematopoietic stem cell through lymphoid-primed progenitors, including the common lymphoid progenitor[64]. The defining event of this development is the assembly of the B-cell receptor. Through V(D)J recombination, the RAG1/RAG2 recombinase introduces double-strand breaks at the recombination signal sequences that flank the Ig gene segments, and the broken ends are joined by the non-homologous end-joining pathway. Combinatorial assortment of a modest number of inherited segments, imprecise joining at the segment junctions, and the pairing of heavy and light chains together generate a receptor repertoire of enormous diversity, with a different specificity in essentially every clone[65]. A cell that assembles a functional receptor matures and leaves the marrow as a naïve B cell, displaying its receptor on the surface as membrane-bound IgM and circulating through secondary lymphoid organs until it encounters an antigen.

The encounter with antigen, in the presence of help from a $CD4^+$ T cell, triggers the reaction at the center of Ig biology: the germinal-center reaction[66]. Within the germinal center, antigen-activated B cells cycle between two compartments called the dark zone and the light zone. In the dark zone they proliferate and the enzyme

AICDA (AID) introduces point mutations into the rearranged Ig genes, a process of somatic hypermutation that further diversifies the receptor. In the light zone the mutated cells compete for antigen and for the limited help of follicular helper T cells, so that the clones whose mutations have improved their affinity are selected to survive and re-enter the cycle. This process is called affinity maturation, essentially a Darwinian process of variation and selection compressed into days and confined within a single tissue.

This reaction is governed by a small set of transcription factors, some of which are recovered by our genetic analysis[62]. The germinal-center program, maintained by *BCL6*, is opposed by the plasma-cell program driven by *IRF4* and *PRDM1* (BLIMP1), the master regulator of the plasma-cell state, while *PAX5* maintains B-cell identity across the lineage. *IRF4* acts in a concentration-dependent manner: low levels are required for the germinal-center response and for class switching, and high levels induce *PRDM1* and commit the cell to the plasma-cell fate. As a cell commits, *PRDM1* represses *BCL6* and *PAX5*, and the loss of PAX5 relieves its repression of *XBP1*, *JCHAIN* and the secreted form of the heavy chain. *XBP1*, induced downstream of *PRDM1*, then drives the expansion of the endoplasmic reticulum and secretory apparatus that allows the cell to export antibody at a rate of thousands of molecules per second [67]. A subset of these plasma cells home back to the bone marrow, where the niche sustains them for years, and it is this long-lived population that maintains the standing concentration of antibody in the serum[68,69]. The quantity that the third paper measures, a person's circulating Ig level, is therefore the integrated output of the entire program described above, from receptor assembly in early B cells through germinal-center selection to terminal differentiation and plasma-cell survival.

Ig itself is not a single molecule but a family of isotypes encoded by a group of constant region genes in the immunoglobulin heavy-chain (IGH) locus on chromosome 14[70]. The five classes – IgM, IgD, IgG, IgA, and IgE – share a common architecture but differ in their constant regions and therefore in their effector roles, and IgG and IgA are further divided into subclasses encoded by distinct genes[71,72]. IgM is the first antibody of a response and circulates as a pentamer joined by the J chain and in complex with CD5L[73,74]. IgG is the dominant antibody of the mature antibody response in blood. IgA has a specialized role in mucosal surfaces but is also found in circulation as serum IgA, which is what we measured in our study[75].

Which isotype a B cell produces is decided by class-switch recombination[76], and the survival of the plasma cells that secrete it depends heavily on a signaling axis built from the TNF-family ligands *TNFSF13B* (BAFF) and *TNFSF13* (APRIL) and their receptors, *TNFRSF13B* (TACI) and *TNFRSF17* (BCMA)[77]. This signaling axis regulates both class switching and plasma-cell survival, and our study of Ig levels recovers population variants mapping to all these genes, including two allelic series mapping to *TNFRSF13B* and *TNFSF13*. Although this aspect was not investigated

extensively in our study, class-switched Ig can also be generated through extrafollicular routes independently of germinal centers[78].

Once antibody is secreted, its fate in the circulation is set by a second set of molecules, the Fc receptors which bind the constant Fc regions of Ig[79]. *FCGR2B* (FcγRIIB) is the only inhibitory Fc receptor, carrying a signaling motif that raises the threshold for B-cell activation and acts as a negative feedback brake on the antibody response to IgG. *FCGRT* (FcRn), the neonatal Fc receptor, binds IgG in a pH-dependent manner within the endosome and recycles it back to the cell surface rather than allowing its degradation, and in doing so sets the long half-life of IgG in the blood[80]. Both molecules act directly on the levels of circulating antibody, one by tuning how much is made and the other by regulating how long it persists in circulation. In our study we find that population variants in the IgG-binding receptor genes *FCGR2B* and *FCGRT* are particularly important for heritable variations in IgG levels, consistent with their known biology.

Like the blood-cell counts of the previous section, Ig levels are quantitative traits in the Fisherian sense. They vary continuously across the population, they are measured routinely in the clinic, and they are substantially heritable, making them well suited to a genome-wide association approach. Their clinical extremes define disease at both ends – immunodeficiency where antibody is too low and a spectrum of autoimmune conditions and B-cell malignancies where the response is dysregulated, so that the genes and variants which set the normal range of these traits are also candidates for understanding their pathology. Analyzing nine Ig phenotypes in a population sample spanning Iceland, Sweden and the UK, we recovered known regulators of the B lineage at every stage from lineage commitment to antibody secretion, and implicated multiple novel genes. Additionally, we found that analyzing Ig levels at this sample size begins to reveal systems-level properties of the B cell system. These findings are discussed in more detail in the Results section.

The studies introduced in this chapter address two quantitative traits derived from human blood, $CD34^+$ stem-cell levels and Ig levels, and describe the same experiment performed at two ends of the hematopoietic hierarchy. The same forward-genetic method – association followed by fine-mapping followed by functional dissection – was applied to the top of the hematopoietic hierarchy, the stem and progenitor compartment, and to one of its terminal lymphoid outputs – the antibody-secreting plasma cell.

Chapter 2 – Aims

Genome-wide association studies have been remarkably successful at locating the regions of the genome in which variation tracks with a trait, but an association is not a mechanism, and the work of moving from a statistical coordinate to molecular mechanism remains a central challenge in the field. The overall aim of this thesis is to apply that full pipeline, from association through fine-mapping to functional dissection, across the hematopoietic hierarchy, taking quantitative blood traits as a forward-genetic discovery platform for the molecular machinery that builds blood and immunity. The traits chosen span the hierarchy from its origin to one of its terminal outputs: the blood $CD34^+$ stem and progenitor compartment at one end, and the circulating Ig produced by the plasma cell at the other.

The specific aims of the constituent papers were as follows.

Paper I. To define the genetic architecture of blood $CD34^+$ cell levels, a trait not previously subjected to genome-wide association studies despite its central role in transplantation, and to begin to resolve the resulting associations to their underlying genes and mechanisms. My focus in this study was the functional dissection of the *PPM1H* locus, the most significant and novel association.

Paper II. To carry one of these associations with blood $CD34^+$ cell levels from to a validated causal mechanism in its native chromosomal context, illustrating end-to-end functional interrogation of a non-coding association signal.

Paper III. To define the genetic architecture of the antibody compartment at scale, by fine-mapping the associations for nine Ig traits at population scale, and to resolve the strongest signals to the core genes and regulatory elements that govern B-cell development, germinal-center selection, and plasma-cell differentiation.

Chapter 3 – Methodological considerations

The analytical methods used throughout this work can be divided into three stages. The first is statistical genetics, the identification of association signals and their resolution to candidate variants and genes, performed largely through the collaboration with deCODE Genetics and through public molecular-genetic resources. The second is functional genomics, the use of chromatin and expression data – in some cases public and in some cases generated by our own lab – to assign each signal to a cell type, a regulator, and a target gene. The third is experimental – the direct testing of the resulting hypotheses in cells.

A recurring theme in post-GWAS analysis is that no single method is decisive on its own. Each step can introduce false positives or spurious associations. A causal variant or mechanism is nominated only where several independent lines of evidence – statistical, epigenomic, and experimental – converge on the same hypothesis.

3.1 Genome-wide association and statistical fine-mapping

The genome-wide association studies underlying Papers I and III were performed by the deCODE collaboration, and the methods are described in full in those papers. They are summarized here because a critical understanding of them is necessary to everything that follows.

In GWAS analysis a quantitative trait is regressed against allele dosage at millions of variants individually, and the variants whose association survives a stringent genome-wide significance threshold are taken forward[81,82]. Two distinct features of the deCODE approach are worth drawing out.

The first is the depth of genotype imputation. By taking advantage of known patterns of LD from high-quality reference genomes, imputation allows assignment of genotypes at variants that have not been assayed directly by microarray hybridization and thus greatly expands the number of sequence variants that can be analyzed

for association[83]. Association power and fine-mapping resolution both depend on the density and accuracy of the imputed variants, and the deCODE reference panel, built on whole-genome sequencing of a large fraction of the Icelandic population, imputes rare variation more accurately than the panels available to most studies[84]. In addition, deCODE has several high-quality Swedish reference genomes that enable rare variant imputation in the samples we provided. This allows our study to combine the cost advantages of microarray-based genotyping with imputation of ultra-rare SNPs and indels, whose effects can be meaningfully detected at the large sample sizes we achieved in our study. Such rare variants of strong effect can be particularly informative for understanding the molecular basis of trait heritability, as theory predicts that they will converge on core pathways regulating that trait[85].

In the context of my doctoral research, the most important step is not the association test but the resolution of each associated locus into independent signals and candidate variants. A genome-wide significant locus is not a resolved causal variant. It is a region of correlated variants, any of which might be causal or merely a passive hitchhiker. Although early GWAS work used simple heuristics based on LD thresholds to disentangle complex genetic signals, the field has converged on more sophisticated methods to solve these issues over time[31]. Two distinct problems must be separated. The first is that a single locus often contains more than one independent causal variant, a feature now known to be common. The second is that, within each independent signal, the true causal variant must be localized against its correlated neighbors.

The deCODE analysis addresses these concerns through a sequence of two analyses. Independent signals are first identified by stepwise conditional analysis, in which the lead variant of a locus is added as a covariate and the region re-tested, the next strongest residual signal is taken as a second independent association, and the procedure is repeated until no further signal survives[86]. Each conditionally independent signal is then fine-mapped by the Bayesian refinement approach of Maller and colleagues, which computes for each variant an approximate Bayes factor and, from it, a posterior probability of being the causal variant under the assumption of a single causal variant within that signal[87]. Summing these posteriors defines a credible set, the smallest group of variants whose cumulative posterior probability exceeds a chosen threshold, conventionally 95% or 99%.

The approach we adopted differs from the other major approach for fine-mapping used in the field – the sum of single effects regression implemented in the popular SuSiE package[88]. SuSiE decomposes a locus into a small number of single-effect components simultaneously, producing a credible set for each component without first conditioning on lead variants. The conditional-then-refine strategy used by deCODE anchors each signal on the empirical lead variant of a stepwise model; SuSiE infers the number and identity of independent effects jointly.

Neither approach is unambiguously better in principle, and both were considered for this work; the deCODE pipeline was used because the association analysis was carried out there. The conditional approach is transparent and robust when signals are well separated, but it can mislocalize the partition between two signals that are close in both position and strength, whereas SuSiE handles such cases more naturally at the cost of greater sensitivity to the reference LD panel. The distinction is not merely academic for this thesis, because the eQTL resources (discussed below) used for colocalization were fine-mapped with SuSiE, and bridging the two representations is a methodological problem addressed in Section 3.3.

Statistical fine-mapping continues to be an active area of research in the field. Although our analysis discarded non-European samples, trans-ancestry analysis has been shown to improve fine-mapping resolution by taking advantage of differences in LD patterns and allele diversity between populations[55,89]. Another recent development is genome-wide fine-mapping using Bayesian mixture models, which fit all variants jointly across the genome rather than treating each significant locus in isolation[90]. The combination of increasing sample sizes in national biobanks with better computational methods promises to substantially increase the yield of causal variants from population genetic studies in humans.

3.2 Expression quantitative trait analysis

A regulatory variant works by changing how much of a gene gets made. To find which gene, we look for expression quantitative trait loci, or eQTLs: a variant is an eQTL for a gene if individuals carrying different alleles have measurably different levels of that gene's transcript[91,92]. When the variant driving a trait association turns out to be the same variant driving an eQTL, that is strong evidence the trait acts by changing that gene's expression. The eQTL resources used here are therefore a foundation of the entire analysis in Paper III, and the choice among them was a substantive methodological decision that is discussed in more detail below.

The central issue for our work is the cell type in which expression is measured. The largest and most widely used eQTL resources measure expression in whole tissues or in whole blood. The GTEx project catalogs eQTLs across dozens of post-mortem bulk tissues, and eQTLGen aggregates whole-blood eQTLs across tens of thousands of individuals[93,94]. Their power is enormous, but for an immunological trait their resolution is limited in a specific way. Whole blood is a mixture of many cell types in varying proportions, and an eQTL detected in such a mixture may reflect a regulatory event in any of them, or an artifact of cell-type composition rather than of regulation within a cell[95,96]. For this reason, GTEx and eQTLGen were not used as eQTL resources in this work.

Our analysis instead rests on sorted-cell eQTL resources, in which expression is measured in purified, defined cell populations. The principal resource is ImmuNexUT, which provides eQTLs across 28 immune cell types sorted from a large cohort, including the B cell and plasmablast populations directly relevant to Ig production[97]. This is further complemented by the eQTL Catalogue, a uniformly reprocessed collection of eQTL studies that puts many sorted-cell and purified-population datasets through a unified analysis pipeline[98]. We selected a subset of the eQTL Catalogue that was relevant to the immune system, focused on sorted myeloid and lymphoid white blood-cell samples.

The advantage of sorted-cell data is that an eQTL is assigned to a known cell type at the point of measurement, so that a GWAS signal colocalizing in plasmablasts is a statement about plasmablasts and not about an unresolved mixture like whole blood. The cost is statistical power, since sorted-cell cohorts are smaller than whole-blood ones, and a real eQTL acting only in a rare sorted population may go undetected. We chose to prioritize specificity over sample size, as cell type specificity is so central to the immune system. To partially resolve this, we were able to use deCODE's internal eQTL resources which include magnetic bead enriched B, T and monocyte fractions from PBMCs at much higher sample sizes. This enabled the colocalization of rare non-coding variants to cell-type specific eQTLs, and provided evidence for a novel gene *ARHGAP15* in B-cell biology.

Single-cell eQTL resources became available during this work[99]. I chose not to use them, because sparse per-cell counts limit detection power for lowly expressed genes, including most transcription factors, and the sorted-cell resources gave better coverage of the genes of interest here[100]. A systematic comparison of single-cell and sorted-cell eQTL data for this application remains to be done.

3.3 Colocalization

The coincidence of an eQTL and a trait association in the same genomic region is not, on its own, evidence that the trait acts through the gene affected by the eQTL. The two signals may be driven by the same causal variant, in which case the regulatory hypothesis is supported, or by two distinct variants that happen to lie in LD, in which case the overlap is coincidental and the gene may be irrelevant[101]. Distinguishing these two situations is the problem of colocalization, and it is the methodological heart of the eQTL analysis in Paper III.

The colocalization problem arises because LD makes naive overlap uninformative. In a region of strong correlation, a trait variant and an unrelated eQTL variant will appear to coincide simply because everything in the region is correlated with everything else. The earliest attempts to address this, such as regulatory trait concordance and the conditional approaches that preceded formal colocalization, were

heuristic[102]. The field converged on the probabilistic framework of Giambartolomei and colleagues, implemented in the coloc package, which remains the standard[103]. Coloc enumerates the possible configurations of causal variants underlying two signals in a region and computes, under a Bayesian model, the posterior probability of each of five hypotheses: no association, association with one trait only, association with the other trait only, association with both but through distinct variants, and association with both through a single shared variant. It is the posterior probability of this last hypothesis that supports a shared mechanism. In our study we used a stringent posterior probability threshold (posterior probability of colocalization >0.9) to ensure that we had high-confidence target genes for each of our signals.

The original coloc method made the same simplifying assumption as the Maller fine-mapping above, i.e. there is at most a single causal variant per signal in the region. Where a locus contains multiple independent causal variants, which is common, this assumption is violated and the method can fail to detect a genuine colocalization or misattribute it[31]. The solution to the problem is the extension of coloc to operate on fine-mapped signals rather than on whole loci. In this approach, the two signals are each decomposed into independent components by SuSiE, and colocalization is tested pairwise between the resulting credible sets[104].

This requires, as input, not merely the marginal summary statistics but the per-component output of the fine-mapping, encoded in the log Bayes factor, or .lbf, files that SuSiE produces. These files carry the per-single-effect Bayes factors for every variant, and they are what allow colocalization to be evaluated for each independent genetic signal instead of an entire locus. The eQTL Catalogue distributes these files publicly, but ImmuNexUT does not; they were obtained for this work through a collaboration with the ImmuNexUT authors that I initiated. This made the colocalization in Paper III possible at a resolution that the public summary statistics alone would not support.

A methodological asymmetry is worth noting – the eQTL signals were fine-mapped with SuSiE, while the deCODE trait signals were resolved by the conditional-Bayesian approach of Section 3.1. The colocalization therefore bridges two different fine-mapping representations, and the deCODE side required construction of compatible per-signal Bayes-factor files, carried out within the collaboration, to interface with the coloc framework.

3.4 Trans associations and inverse-variance weighted regression

The *cis* methods above assign target genes to candidate variants within their own locus (usually defined as a 1 Mb window around the lead variant). A separate set of

trans analyses in Paper III asks how the Ig-associated variants relate to the rest of the proteome and to each other, and these *trans* associations reveal properties of the antibody system that would be invisible to single locus analyses.

Plasma proteins are measured at population scale by affinity proteomics, principally the Olink platform, which quantifies thousands of plasma proteins at once by converting each antibody-binding event into a DNA barcode that can be counted by sequencing[105]. Applied across human population samples, this yields a protein quantitative trait locus, or pQTL, map: the association between genetic variants and the circulating level of each protein[106]. The UK Biobank Pharma Proteomics Project, which measured roughly 3,000 plasma proteins in over 50,000 participants, is the primary resource used for this analysis[107].

A pQTL is termed *trans* when the variant lies far from the gene encoding the protein it affects, and it is the *trans*-pQTLs that carry information about how a trait connects to the wider proteome. Because the variant does not sit near the affected protein's gene, a *trans* association points to a systems-wide effect of a genetic perturbation[106]. The canonical illustration is the ABO locus, which associates in trans with the plasma levels of dozens of structurally unrelated proteins. The variant's position at ABO gives no hint of these targets; the connection runs through the ABO glycosyltransferase, whose activity alters the glycosylation and hence the circulating half-life of many proteins at once. A *trans*-pQTL therefore exposes regulatory links that the variant's target gene alone would never reveal, and recurring sets of *trans*-pQTLs for a given trait can trace the network of molecular relationships the trait participates in. It is this property that the Ig analysis exploits, using the *trans*-pQTLs of the Ig-associated variants to find the plasma proteins whose levels move in concert with antibody levels.

The *trans*-pQTL analysis in this thesis uses inverse-variance weighted (IVW) regression, the estimator underlying two-sample Mendelian randomization[108], in two separate applications. In IVW regression, the effects of a set of variants on one trait are regressed on their effects on another, through the origin, with each variant weighted by the inverse squared standard error of its effect on the outcome. Weighting by precision is necessary because the studies compared differ in sample size.

A causal interpretation of IVW regression (as in Mendelian randomization) requires the exclusion restriction, that the instruments affect the outcome only through the exposure and there is no horizontal pleiotropy[109]. However, as this assumption cannot be conclusively demonstrated to be true in the traits examined here, we dropped causal interpretations and do not refer to our analyses as Mendelian randomization. The regression is therefore used as a test of proportionality between two sets of genetic effects, not as a causal estimate. Observed proportionality is interpreted from prior knowledge of the proteins involved, or left open.

In the first application, Ig-associated variants were intersected with plasma pQTLs[107], and for each protein enriched among the variants for a given Ig trait, variant effects on the protein were regressed on their effects on the Ig trait. Where the protein is a known physical partner of the antibody, proportionality serves as a positive control on the association signals. Where no physical relationship is known, it identifies a candidate coupling of undetermined direction. Both kinds were found (Section 4.3).

The second application uses the same IVW regression on the IgG subclasses. IgG1 to IgG4 are encoded by separate constant-region genes in fixed tandem order within IGH, and a B cell selects among them by class-switch recombination, a deletional rearrangement that excises the intervening DNA[76]. Because the subclasses derive from one locus through a shared switching mechanism, a change favoring one may occur at the expense of another. A prior deCODE GWAS of IgG subclass levels reported one such case, a rare deletion spanning *IGHG1* that raises IgG3[110]. The larger variant yield at IGH here provided enough instruments to test whether such compensation is systematic.

Paper III measures total IgG, the sum of the four subclasses. Using the prior subclass GWAS[110], matched variants for each subclass were regressed on total IgG. The relationship is partly arithmetic, since a variant changing one subclass changes total IgG in proportion to that subclass's share, so the fitted line describes a variant that scales production without redistributing it. Departures were quantified as the difference between the observed subclass effect and the IVW prediction, divided by the standard error of the subclass effect, i.e., a standardized residual. Variants were flagged as outliers when the absolute value exceeded two standard deviations of the distribution across instruments in the panel. Flagged variants were retained if they also showed reciprocal effects across subclasses, raising one while lowering another. Noise produces deviations independently across subclasses, whereas redistribution produces opposite-sign departures in the same variant. The retained variants are described in Section 4.3.

3.5 Chromatin accessibility and ATAC-sequencing

The statistical methods above nominate a variant and, through eQTL colocalization, a gene. They do not show that the variant sits in a functional regulatory element, nor in which cell type that element is active. In post-GWAS analysis, this is routinely done using chromatin accessibility data.

Accessibility is central to eukaryotic gene regulation. Weintraub and Groudine showed in 1976 that transcriptionally active genes sit in chromatin preferentially sensitive to digestion by DNase I, establishing that the regulatory state of a locus is written in the physical accessibility of its DNA[111]. Discrete hypersensitive sites

within that accessible chromatin, marking the positions where regulatory proteins bind, were mapped at the end of the decade[112]. The underlying logic is structural – the genome is packaged into nucleosomes, which occlude the DNA they wrap, and the regions where regulatory proteins such as transcription factors bind are locally stripped of nucleosomes and laid open[113].

For some years accessibility was overshadowed by the concept of the histone code, the proposal that the many covalent modifications on histone tails form a combinatorial language that specifies the regulatory state of a region[114]. Henikoff and Shilatifard pushed back on this directly, arguing that the real question is causality: whether histone modifications drive differences in chromatin state, or merely register them[115]. Their interpretation was that the patterns of several key histone modifications are better explained as consequences of transcription and nucleosome disruption than as instructions causing it, and that the older accessibility paradigm, running back to Weintraub and Groudine, remains the mechanistically sound account of how nucleosomes regulate the genome. The genome-wide data supports this: ENCODE found that the large majority of transcription factor binding occurs within accessible chromatin[116].

The development of ATAC-sequencing broadened the use of genomic accessibility data. The earlier genome-scale method for assaying chromatin accessibility, DNase-sequencing, needs both large numbers of cells, which limits which cell types can be profiled, and a prohibitively expensive read depth[117]. ATAC-sequencing, developed by Buenrostro and colleagues, uses a transposase to insert sequencing adapters straight into accessible DNA[118]. By taking advantage of the enzyme's natural preference for accessible DNA, this cut both the input requirement and the cost by orders of magnitude. This made it possible to systematically profile sorted primary cell populations available only in small numbers, rather than relying on cell lines. Once accessibility is mapped across a panel of defined cell types, the open regions become lineage markers: an element open in erythroid progenitors but closed elsewhere marks the erythroid lineage, one open only in B cells marks the B lineage. Intersecting a credible set with such a panel therefore tests whether the candidate variants fall in accessible chromatin at all, and at the same time assigns the signal to the specific cell type(s) where that chromatin region is open. The two resources used throughout this thesis are the ATAC-sequencing atlases generated for hematopoietic progenitors by the Buenrostro and Sankaran labs[57], and a much more comprehensive atlas of differentiated immune cells generated by Calderon and colleagues[119].

ATAC-sequencing has been further extended into single-cell ATAC-sequencing (scATAC-sequencing), allowing chromatin accessibility profiling of complex solid tissues without sorting into purified cell populations[120]. In this work, public scATAC-sequencing data from tonsillar samples was particularly valuable[121], as the tonsil is a central site for B-cell development, and it allowed variant effects to be located in transitional developmental states such as light- and dark-zone germinal-center B cells.

3.6 Heritability partitioning and variant-set enrichment

Enrichment of GWAS signal in cell-specific open chromatin can be used to determine if heritability for that trait is enriched within a specific cell lineage. Two distinct methods were used to perform this analysis in our work.

Paper I used partitioned heritability, implemented using stratified LD score regression and carried out by one of my colleagues[122]. The method takes the genome-wide distribution of association statistics across all variants and asks what fraction of a trait's heritability falls within a given annotation, such as the accessible chromatin of a particular cell type. It works from summary statistics alone and needs no fine-mapping. For the $CD34^+$ trait it showed that heritability is concentrated in chromatin accessible in stem and progenitor populations, confirming that the trait acts intrinsically in HSPCs[59].

For Paper III I used gchromVAR instead[58]. Rather than the genome-wide signal, it takes the fine-mapped variants, weighted by their posterior probabilities, and measures their enrichment in the accessible chromatin of each cell type. For the Ig traits, gchromVAR showed that high probability variants are enriched in the accessible chromatin of B lineage cells.

The two methods use different inputs and estimate different quantities. Stratified LD score regression partitions genome-wide SNP heritability across binary annotations; gchromVAR tests whether fine-mapped posterior probability mass is concentrated in quantitatively weighted accessibility peaks, against a background matched for GC content and peak intensity. The choice was driven by the data available in each study: Paper I had genome-wide summary statistics but limited fine-mapping, while Paper III's dense credible sets suited a posterior-weighted method. Both analyses localized each trait's signal to the lineage its biology predicts: the stem and progenitor compartment for blood CD34+ levels, and the B-cell lineage for Ig levels.

3.7 Single-cell resolution of cell state

Bulk accessibility and bulk expression data can assign a signal to a sorted cell type, but a sorted population is itself heterogeneous, and a regulator that acts transiently at a particular developmental transition can be averaged out in bulk samples. Furthermore, many relevant developmental stages occur in inaccessible solid tissues such as the bone marrow or germinal centers[66]. Resolving such cases requires single-cell data, and single-cell expression and accessibility were used in Paper III to localize candidate regulators within the B-cell compartment at a resolution that bulk sorting cannot reach.

For the Ig GWAS the relevant tissue is the germinal center, where B cells pass through distinct states, the proliferative dark zone, the selective light zone, and the commitment to plasma-cell fate. These processes occur within solid secondary lymphoid tissues and cannot be directly assayed in circulating cells. To place the novel regulators emerging from the GWAS within this trajectory, single-cell datasets of the germinal-center reaction were used, principally the human tonsillar dataset of King and colleagues[121], which resolves the germinal center into its constituent states and provides readouts of both single-cell expression and accessibility.

This made it possible to show that the candidate genes *ZNF608* and *MCTP2* are specifically upregulated in germinal-center B cells relative to other B cell subsets[63]. Single cell ATAC-sequencing data further allowed the localization of credible set variants for these genes within accessible chromatin of these transitional developmental phases. Although further work will be needed to elucidate the precise roles of *ZNF608* and *MCTP2* in germinal-center development, it was only through use of single-cell data that we were able to localize these genetic effects within a transitional developmental state that would be invisible to measurements taken from circulating lymphocytes.

3.8 Prediction of transcription factor binding disruption

A causal regulatory variant in accessible chromatin often acts by altering transcription factor binding. Motif analysis predicts whether an allele strengthens or weakens the match to a factor's known binding preferences. Two tools were used, reflecting methods that became established over the course of this work. The $CD34^+$ work used PERFECTOS-APE, which scans databases of position weight matrices and estimates the change in binding probability between alleles[123]. The later papers used FABIAN, which aggregates multiple models per transcription factor rather than relying on a single matrix, including hidden Markov model representations that capture dependencies between adjacent positions[124], and which integrates curated ChIP-sequencing occupancy from ReMap[125,126].

Two limitations apply. First, a predicted disruption is a hypothesis about binding, not a measurement of it. A variant can disrupt a strong motif for a factor that is never expressed in the relevant cell, and a factor can bind through cofactors at a site where its own motif is weak. Motif prediction alone is therefore never sufficient to assign a causal transcription factor. Where a factor was nominated, occupancy was checked by retrieving published ChIP-sequencing for that specific factor in relevant cell lines and primary cells, rather than by mass intersection against all available data. This is how binding of EGR1 at *TNFSF13* and of MYB at *PPM1H* was established – in these cases I did not even use database resources like ReMap but performed manual

literature searches to identify relevant datasets[127,128]. Second, motif methods score a variant against a set of matrices without modeling the surrounding sequence. Sequence-based deep-learning models such as ChromBPNet[129], which predict accessibility directly from sequence and score a variant in its full context, are the natural successor to this approach; they were not used here.

All of these methods are biased by how unevenly transcription factors have been characterized. A few factors have extensive genome-wide binding data and many have none, so the available motifs and occupancy tracks are a skewed sample of the proteins that might bind a site. Protein-binding microarrays and related in vitro assays partly fill this gap, but in vitro preference is a poor surrogate for binding in situ, where chromatin context determines where a factor actually goes. Such data feed into JASPAR[130] and PERFECTOS-APE and therefore into these results unavoidably; what I avoided was treating an in vitro-only motif match as independent evidence for a causal factor.

3.9 Reporter assays and the dissection of regulatory dependence

The methods described so far are computational. They nominate a candidate variant, a regulator, a target gene, and a cell type, but none of them demonstrates that the variant alters regulatory activity. That requires an experiment, and the principal experimental method in this thesis is the luciferase reporter assay[131].

In this assay the candidate regulatory sequence, in each of its allelic forms, is cloned upstream of a luciferase gene and transfected into cells, and the luminescence produced reports the regulatory activity of that sequence. A difference in signal between the two alleles indicates that the variant alters enhancer or promoter activity in the cell type assayed. The assay has two limitations. The sequence is tested on a plasmid rather than at its native chromosomal position, so it reports the intrinsic activity of the element rather than its behavior *in situ*, and only one element is tested per construct. The unbiased alternative is the massively parallel reporter assay (MPRA), which tests thousands of sequences in a single pooled experiment and would have allowed many credible-set variants to be screened together[132,133]. An MPRA was not feasible within the cost and time constraints of this work, so prioritized variants were tested individually.

Combining the reporter assay with small interfering RNA (siRNA) knockdown[134] of the nominated transcription factor tests whether the allelic difference depends on that factor. Because the two constructs differ only at the variant, an indirect consequence of depleting the factor would be expected to affect both alleles; an allele-selective loss of signal therefore implicates the factor acting at that site. This was

the basis for showing that activity of the *PPM1H* element depends on MYB, where MYB knockdown reduced the signal from the allele predicted to bind it but not from the other allele. The same design was used to show that a rare non-coding variant reduces Ig levels by abolishing EGR1 binding at the *TNFSF13* promoter. These experiments test the factor dependence that motif analysis predicted.

3.10 Genome editing

Genome editing allows direct perturbation of non-coding variants in their chromosomal context: the element is removed or altered where it sits, and the consequence for the surrounding genes is quantified using methods like qPCR. Both Paper I and Paper II depend on this method to draw their conclusions.

The method used throughout was CRISPR/Cas9[135,136]. A short guide RNA directs the Cas9 nuclease to a chosen position, where it introduces a double-strand break; cellular repair then leaves a small insertion or deletion at that site, or, when two guides are used, excises the segment between them. The readout is the expression of neighboring genes. The perturbation occurs at the endogenous locus, with chromatin state and flanking sequence otherwise unchanged, so a change in expression can be attributed to the element in the position where it normally acts. A plasmid-based reporter cannot assay genetic perturbations in their native genomic context, which is why the two methods are used together rather than as alternatives.

Several limitations restrict what this evidence establishes. First, a deletion removes several base pairs and therefore more than the variant itself, so the experiment shows that the element is required, not that a particular variant within it is causal; the variant-level claim rests on the reporter assay (and on allelic imbalance, discussed below). Second, a deletion also changes the spacing between the sequences that flank it, which can perturb local looping independently of the deleted element's own activity, so a positive result can have an alternative explanation. Third, the editing was carried out in immortalized cell lines rather than primary cells. K562, used for these experiments, is an erythroleukemia line with an abnormal karyotype and its own regulatory state, so a target gene is assigned in a system that approximates rather than reproduces the cell type in which the variant acts.

Three alternatives address these limitations. CRISPR interference, in which a catalytically inactive Cas9 fused to a repressor domain is targeted to the element without cutting, silences it reversibly, leaves the sequence and its spacing intact, and thereby avoids the second limitation above[137]. Base editing and prime editing install a single substitution at the endogenous locus and would allow the variant-level question to be addressed directly[138,139]. Deleting the element in primary cells would place the perturbation in the correct cell type, but this can be very technically challenging and limits editing efficiency. None of these alternatives was used in this thesis.

3.11 Allelic imbalance

Allelic imbalance analysis provides complementary evidence to the reporter assays described above – it tests the variant directly in its native chromosomal position[140]. This form of analysis supplied some of the most direct causal evidence in this thesis, including the clearest pointer to a causal variant for *ZNF608*, a novel gene regulating B cell development.

The principle is that in a cell heterozygous at a candidate variant, both alleles sit in the same *trans* environment, exposed to the same factors at the same concentrations, and differ only in their own sequence. A systematic difference between the two alleles in a sequencing readout, whether in chromatin accessibility, transcription factor binding, or expression, is therefore attributable to sequence differences in *cis*. Reads carrying each allele are counted and tested against the ratio expected under no imbalance using a binomial test. A significant departure from balanced counts is evidence that the variant has a cis effect in its native chromatin context.

The method was used in two forms. For transcription factor binding, allelic imbalance in ChIP-sequencing reads was tested, as for MYB binding at the *PPM1H* variant in Jurkat cells[128] (Paper I) and GATA2 binding at the *ENO1* variant[141] (Paper II). For accessibility, allelic imbalance in ATAC-sequencing or DNase-sequencing reads was tested in primary cells, as in primary $CD34^+$ data[141] *at the ENO1 variant and in primary B-cell ATAC-sequencing data*[119], where the imbalance pointed to the causal variant for *ZNF608* (Paper III).

One statistical subtlety arises due to ploidy. The binomial test requires an expected allelic ratio, which is 0.5 at a diploid heterozygous site. Several of the cell lines used in functional genomics are not diploid, and where a locus is present in more than two copies the expected ratio depends on the copy number of each allele. K562 is not diploid at the ENO1 locus, so a test against a ratio of 0.5 would be incorrect. Imbalance was therefore evaluated against the measured copy state of the locus and confirmed to hold across the plausible copy number configurations, so that the inference does not rest on an assumed diploidy the cell does not have. Where the imbalance is large, as it was here, it remains significant across all plausible ploidies, which is the most defensible form of the claim. Copy number for K562 and Jurkat was taken from published whole-genome sequencing of these lines[142,143]. Primary human B-cell[119] and $CD34^+$ samples[141] were treated as diploid.

3.12 Ethical considerations

The work in this thesis used human blood samples and human genetic data, collected with ethical approval. The Swedish blood donors and primary care patients were

sampled under approvals from the Lund University Ethical Review Board (dnr 2018/2 for Paper I; dnr 2022-01414-02 for Paper III), and the umbilical cord blood used for the primary $CD34^+$ experiments in Paper I was collected at Skåne University Hospital and Helsingborg Hospital under regional ethical committee regulations with informed consent. The Icelandic data in Paper III were collected under approvals from the Icelandic National Bioethics Committee (no. 16-022) and the Icelandic Data Protection Authority, and the UK data were obtained from UK Biobank under application 56270, which operates under its own approval and participant consent. Genotype and phenotype data were handled in pseudonymized form throughout. Paper II used only publicly available data and established cell lines and required no separate approval.

Chapter 4 – Results

4.1 Paper I: The genetic architecture of the blood CD34+ cell compartment

My contribution to this study[59] was the functional characterization of the *PPM1H* locus, described below; the genome-wide association study and the other candidate-gene analyses were carried out by the wider collaboration.

Blood CD34+ cell levels had never been subjected to a genome-wide association study before this work, despite the centrality of the CD34+ count to stem-cell transplantation. Across 13,167 blood donors and primary care patients from southern Sweden, nine signals were significantly associated with circulating stem-cell levels as measured by flow cytometry. The central purpose was to test whether genuine regulators of the stem and progenitor compartment could be read out from natural genetic variation in humans.

As part of this study a colleague generated an eQTL dataset for circulating CD34+ cells through magnetic bead enrichment and RNA-sequencing, and associations between candidate variants from the GWAS and nearby genes were evaluated by multiple linear regression. Colocalization analysis, as performed in Paper III, was not fully developed at the time this paper was published and was not used here.

Our study design was validated by four of the nine signals mapping independently to CXCR4, a chemokine receptor that retains stem cells in the marrow niche and the target molecule for the approved mobilizing drug plerixafor[52]. Within each of the four credible sets a single regulatory variant was identified that either lies in the *CXCR4* promoter or loops to it[144], and the variants act on *CXCR4* expression in the expected direction, confirmed for the three common variants by *cis*-eQTL analysis in circulating $CD34^+$ cells and by CRISPR deletion in a cell line. The rare variant rs555647251 was not polymorphic in the expression cohort and could not be tested, highlighting the importance of larger sample size eQTL datasets for stem and progenitor cell populations. Partitioned heritability[122] showed that the heritability of the $CD34^+$ trait is concentrated in chromatin accessible in stem and progenitor populations, further evidence that the associations act intrinsically and cell-autonomously within HSPCs rather than through some downstream or systemic route.

Beyond *CXCR4* and the known myeloid regulator *CEBPA*, the study implicated four genes never previously implicated in stem-cell biology – including *PPM1H*, *ENO1*, *RERE* – together with *ITGA9*, which had been only minimally studied in this context. Two of the implicated genes underlie autosomal dominant disorders defined by aberrant regulation of the stem and progenitor compartment: *CXCR4* in WHIM syndrome, where gain-of-function mutations in the C-terminal region prevent receptor internalization and retain progenitors in the marrow, and *CEBPA* in familial acute myeloid leukemia[145,146]. Somatic mutations in both genes recur in blood malignancies[147,148]. Six of the nine significant variants also coincided with variants known to associate with mature blood-cell traits[54], placing the stem-cell signals on the same genetic axis as the differentiated compartments downstream of them.

The most significant association of the entire study mapped to PPM1H, a serine/threonine phosphatase that had never been implicated in hematopoiesis, and it became the focus of the functional work. Relative to other blood cell types, PPM1H is preferentially expressed in HSPCs, and the cis-eQTL in sorted CD34+ cells showed that the allele conferring higher blood CD34+ levels is associated with lower PPM1H expression. The credible set for the locus falls within the first intron of PPM1H, and integrating chromatin accessibility with expression across sorted HSPCs localized a roughly 500 bp segment within that intron whose accessibility correlates with PPM1H expression and which loops to the gene's promoter[144]. This segment is open specifically in the early populations where the trait acts – hematopoietic stem cells, multipotent progenitors, common myeloid progenitors, and megakaryocyte-erythroid progenitors.

Four credible-set variants lie in this HSPC-specific regulatory element. To find which carries the regulatory effect, I tested all four in luciferase reporters in K562 cells, and only rs772557 showed a significant allelic difference, with the G allele driving more reporter activity, in a direction consistent with the eQTL observed in primary CD34+ cells. Three independent lines of evidence identify MYB as the factor through which rs772557 acts. First, motif analysis with Perfectos-APE[123] predicted that the rs772557-G allele creates a binding site for the master hematopoietic transcription factor MYB[149] that the A allele abrogates. Second, ChIP-sequencing for MYB in Jurkat cells[128], which are heterozygous at rs772557 in whole-genome sequencing data, showed complete allelic imbalance towards the G allele, demonstrating *in situ* allele-specific binding. Third, knockdown of MYB by siRNA selectively inhibited the reporter activity of the G construct while leaving the A construct unchanged, establishing that the allelic difference in activity depends specifically on MYB. Separately, allele-specific CRISPR/Cas9 disruption of the region in K562 cells (these experiments were performed by a colleague), which are also heterozygous for rs772557, lowered *PPM1H* expression, confirming that the element itself is required for normal expression of the gene in a genotype-dependent manner. Together, reporter activity, motif prediction, native-context binding, factor

dependence, and perturbation of the element identify rs772557 as a causal variant that tunes *PPM1H* expression in $CD34^+$ by gaining or losing an MYB site.

The final step connected the gene back to the cellular phenotype in primary stem cells, conducted in collaboration with Jonas Larsson's group at Lund. Knockdown of *PPM1H* by shRNA in primary $CD34^+$ cord blood cells raised the proportion of $CD34^+$ and $CD34^+$ $CD90^+$ cells relative to control, the direction predicted by the eQTL, in which less *PPM1H* corresponds to higher $CD34^+$ levels.

Paper II takes a second of the novel $CD34^+$ loci, *ENO1-RERE*, to the same depth, and finds a mechanism of the same kind.

4.2 Paper II: Functional dissection of the 1p36.23/ENO1-RERE locus

This study was designed and carried out by me in its entirety, comprising both the computational and the experimental work[60]. Among the eight loci associated with blood CD34+ cell levels in Paper I was a strong association at 1p36.23, in the intergenic region between ENO1 and RERE. The lead variant, rs2047094, is common and ranked among the most significant signals in the study, and the allele associated with higher blood CD34+ levels was associated with higher expression of both ENO1 and RERE in sorted CD34+ cells in eQTL data generated by our lab[59]. These data established that the association acts through cell-intrinsic changes in expression at the locus but did not identify the causal variant, the mechanism, or the target gene, since the signal is composed of 17 variants in tight LD across approximately 14 kb.

Paper II resolves the signal to a single causal variant and a defined molecular mechanism. Assigning the target gene requires a note on the eQTL analysis, as both ENO1 and RERE are plausible candidates by proximity. In Paper I we used in-house eQTL data from magnetic-bead enriched CD34+ cells and multiple linear regression to test the association between rs2047094 genotype and expression of both genes, and technical choices about how the expression data are normalized or corrected, and which covariates are included, can significantly influence such tests[150,151]. Tested against uncorrected values, the variant shows a significant eQTL for ENO1 only, with RERE non-significant. Expression data of this kind contain substantial latent structure, arising from technical batch effects, differences in cell composition, and other unmeasured covariates, which introduces correlated noise across genes and can both generate spurious associations and obscure real ones[150]. The established correction is to estimate this structure from the expression matrix and adjust for it, for example by including principal components as covariates[152]. With the first 10 principal components included as covariates, the *RERE* eQTL reached significance while the *ENO1* eQTL was retained. The differential sensitivity reflects

expression level: *RERE*, a transcriptional cofactor for retinoic acid signaling[153], is expressed well below *ENO1*, a glycolytic enzyme. A low-expression transcript has a correspondingly lower signal-to-noise ratio, so its eQTL is more readily masked by unwanted variation and recovered only after adjustment. The corrected analysis supports both ENO1 and RERE as candidate targets of the locus.

To reduce the 17 candidate variants to a shorter list, I intersected them with ATAC-sequencing across sorted blood populations[57], on the principle that a regulatory variant should fall within accessible chromatin in the relevant cell type. Two variants, rs10864368 and rs2065603, both in near-complete LD with the lead variant, lie in accessible chromatin of primary $CD34^+$ populations. I tested both variants in luciferase reporters in K562 cells. The construct carrying rs10864368-C, the allele associated with higher *ENO1* expression and higher $CD34^+$ levels, produced significantly greater reporter activity than the rs10864368-T construct, whereas neither allele of rs2065603 gave signal above background. These results identify rs10864368 as the regulatory variant and establish the direction of its effect.

To determine which gene the element regulates, I deleted the rs10864368-harboring region in K562 cells using CRISPR/Cas9 with a guide RNA spanning the variant and measured the effect on both neighboring genes. Deletion reduced *ENO1* expression, consistent with the element acting as a transcriptional activator of *ENO1*, and produced no significant change in *RERE*. However, this does not rule out *RERE* as a target gene, as discussed further below.

To establish that rs10864368 exerts a regulatory effect in its native chromosomal context, I examined allelic imbalance in chromatin accessibility. In ENCODE DNase-sequencing data from primary $CD34^+$ cells of seven heterozygous donors[141], accessibility was significantly skewed toward the rs10864368-C allele, with a consistent bias across donors. Because both alleles occupy the same nuclei and are exposed to the same *trans* environment, this imbalance constitutes direct evidence that the C allele confers greater accessibility in primary human stem and progenitor cells.

Motif analysis identified the binding site most strongly disrupted by the variant as that of GATA2, with the rs10864368-T allele abolishing a canonical GATA motif. GATA2 is a central regulator of the stem and progenitor state: conditional deletion of *Gata2* in mice eliminates adult HSPCs and germline GATA2 mutations cause GATA2 deficiency syndrome in humans, a rare genetic disease where disrupted HSPC development causes immunodeficiency and blood malignancies[154–157].

Published ChIP-sequencing data[158] confirmed GATA2 occupancy directly over rs10864368 in primary $CD34^+$ HSPCs. To test whether binding is allele-specific in the native context, I examined allelic imbalance in GATA2 ChIP-sequencing reads from K562 cells, which are heterozygous at the variant in whole-genome sequencing[143]. Reads derived almost exclusively from the C allele, an imbalance that remains significant across all plausible copy-number states at the locus.

Multiple lines of evidence – reporter activity, accessibility imbalance in primary cells, motif disruption, direct occupancy, and binding imbalance – together converge on a single mechanism in which rs10864368-T abrogates a GATA2 binding site, reducing GATA2 binding and local chromatin accessibility and lowering target gene expression.

The behavior of the element through differentiation was examined using public time-course ChIP-sequencing data[159] from *ex vivo* primary HSPCs that were differentiated toward the erythroid lineage using defined culture conditions. GATA2 occupies the rs10864368-harboring element in the stem and progenitor state and is progressively lost as the cells commit to erythroid fate. Notably, it is not replaced by GATA1 and thus does not show the GATA switching behavior observed in erythroid differentiation[160]. In the canonical GATA switch of erythroid maturation, GATA1 succeeds GATA2 at shared sites during differentiation; at this element GATA2 departs without a GATA1 handover, identifying the element with the progenitor state rather than the erythroid program that follows.

The signaling transcription factor SMAD1 transiently occupies the same regulatory element containing rs10864368 during this differentiation time-course. SMAD1 is the intracellular effector of bone morphogenetic protein (BMP) signaling[161]. This indicates that the element integrates an extrinsic signal, BMP, with an intrinsic lineage-determining factor, GATA2. Regulatory elements that combine signaling and lineage-determining transcription factors in this manner have been reported to be enriched for GWAS variants[159], which provides a mechanistic rationale for the presence of a common variant of appreciable effect at this position: a variant situated at the convergence of a signaling input and a master regulator will have stronger effects on the output of the progenitor compartment and is therefore detectable by genome-wide association. The rs10864368-harboring element represents a specific instance of this class and connects the association to a defined node of hematopoietic signaling.

The phenotypic consequences of the variant extend beyond $CD34^+$ cell levels. Across other GWAS associations[162,163], the rs10864368-C allele, which increases $CD34^+$ cell levels, is associated with a higher frequency of myeloid-lineage cells and a lower lymphocyte percentage. This pattern is consistent with stronger GATA2 binding at the element expanding the myeloid-biased progenitor compartment, raising the circulating $CD34^+$ count while shifting hematopoietic output toward the myeloid lineage.

Deletion of the rs10864368 element in the K562 cancer cell line reduced *ENO1* but not *RERE*, and the paper reports *ENO1* as the primary candidate on that basis. However, the corrected eQTL derived from primary HSPCs supports both genes, and the biological case supports *RERE*. *ENO1* is a glycolytic enzyme, and its proposed capacity to influence stem-cell numbers depends on a moonlighting function of a shorter isoform for which published evidence is limited[164,165]. *RERE* is a

transcriptional coregulator of retinoic acid signaling[153], a pathway with an established role in hematopoietic stem cells, where it maintains dormancy and where reduced signaling lowers HSPC numbers[166]. *RERE* is therefore a strong candidate target alongside *ENO1*.

4.3 Paper III: The genetic architecture of the human B-cell system

This study[63] was carried out in collaboration with deCODE genetics. As first author I played a leading role in designing and choosing the downstream analyses. I performed all of the reporter assay experiments, and carried out the colocalization, chromatin-enrichment, *trans*-pQTL concordance, Ig subclass buffering, and transcription factor motif analyses. The genome-wide association testing, conditional resolution of signals, and fine-mapping were performed by our collaborators at deCODE Genetics, and the immune cell phenotypes were drawn from the BloodVariome compendium, a companion resource from our group[167].

The regulation of the human B-cell system *in vivo* is inaccessible to direct experimentation. Animal models – in particular mice – have supplied most of what is known about the transcription factors, cytokines, and receptors that govern antibody production[62]. Circulating Ig levels offer an indirect route to probe this process in humans. They have been measured in the clinic for decades as a coarse index of humoral function, and inter-individual differences in these levels are heritable[168], which raises the possibility that population variation in a simple blood measurement encodes information about the regulatory machinery that produces it. Whether that information is fine-grained enough to reconstruct the system, rather than merely to flag a handful of major loci, was unknown. Previous genetic studies of Ig traits were modest in scale – a previous study from our lab had 19,129 individuals[168] – and had resolved only a partial view of the underlying architecture.

In this study we analyzed nine Ig traits in 114,697 individuals from Iceland, Sweden, and the UK Biobank, and demonstrate that at this scale the coarse readout of Ig levels can resolve the system that generates it.

A central design decision was to analyze not only the three measured isotypes but six composite traits derived from them. This design was introduced in that earlier study, which analyzed the same nine traits and gave a first indication that the composites carried signal beyond the single isotypes. The composite traits are linear combinations of log-transformed primary isotype measurements (thus representing products and ratios of these measurements). Alongside IgA, IgG, and IgM, we constructed AGM (A × G × M, a product trait) as an index of total Ig, AG as an index of class-switched output, M/AG as a measure of the balance between pre-switched

and class-switched antibody, and A/M, G/M, and A/G as isotype-specific class-switching biases, each built from standardized, log-transformed measurements.

Combinations of noisy clinical measurements need not behave as coherent phenotypes, and there was no guarantee that a quantity such as A/G would map onto a biological process rather than dissolving into measurement error. In practice it did not: the composite traits increased the variant yield of the study, contributing 186 signals that were not captured by any single-isotype analysis.

They also appear to model the processes they were designed to proxy. A/G and G/M both uncover rare variants at the IGH locus that do not reach genome-wide significance for the primary Ig traits used to derive them. As these ratios were designed to model class-switch recombination and IGH is the physical site of class switching in the human genome[76], simple ratio traits appear to increase the power to detect functional variation related to switching. The product trait AGM recovered the largest number of rare variants connected to heritable immunodeficiency, indicating that it captures information about overall humoral immunity that single-isotype measurements do not. The composite traits therefore function as sensors of specific aspects of the antibody response.

We tested 62.2 million sequence variants against the nine traits. Stepwise conditional analysis identified 1,144 independent signals, which 95% credible sets and inter-trait colocalization collapsed to 504 unique variants, 472 of them previously unreported and 33 of them rare, with effect allele frequencies at or below 1%. Lead variants for each trait explained between 11 and 21% of trait variance, consistent with heritability estimates of 13 to 18% from LD score regression, and fine-mapping resolved 70 of the signals to a single putative causal variant. The remainder of the chapter describes what this variant set recovers about the B-cell system.

To move from variants to genes, we prioritized candidate effector genes at each locus using two conservative lines of evidence: nonsynonymous coding variants within the credible set with posterior inclusion probability above 10%, and *cis*-expression quantitative trait loci[97,98] in sorted immune cells and blood that colocalized with the association at posterior probability above 90%[104]. This implicated 397 genes across 290 signals, and for loci without molecular support we assigned the nearest protein-coding gene. The resulting gene set was enriched for Ig-related mouse phenotypes and for Reactome pathways, and across a reference panel of 81 human cell types it was most highly expressed in B cells and professional antigen-presenting cells[169–171].

The genes span the entire B-cell life cycle: transcription factors that regulate B-cell development and differentiation, among them *EBF1*, *NFKB1, IKZF1, IKZF3, IRF4, PAX5*, and *RUNX3*; cytokines and receptors for differentiation and survival, including *TNFSF13B*, *TNFSF13*, *TNFRSF13B*, and *TNFRSF17*; effectors of the terminal plasma cell, including *ATG5*, *ELL2*, and *SLAMF7*; and the machinery that clears

and recycles secreted antibody, including the Fcγ receptors and the neonatal Fc receptor *FCGRT*.

To ask whether these variants act cell-autonomously within B cells rather than through other lineages, I tested the enrichment of Ig-associated variants in accessible chromatin across sorted immune cell types using gchromVAR[58]. The genetic signal for all nine Ig traits was strongly concentrated in B-cell open chromatin, with a permutation P of 7.9×10^{-5}, indicating that Ig variation is mediated largely through B-cell-autonomous regulatory elements and that simple Ig traits can act as a discovery platform for B-cell physiology and regulation when analyzed at population scale.

Genetic correlations[172] between the Ig traits and found that the genetic architecture recapitulates the known differentiation hierarchy of isotype switching. IgM showed minimal genetic correlation with IgA and with IgG, at r_g of -0.012 and 0.13, consistent with largely separate networks governing pre-switched and class-switched antibody. IgA and IgG were more strongly correlated, at r_g of 0.36, reflecting the pathways they share downstream of class switching. Twenty-three candidate genes, 5.8% of the total, were shared across all three isotypes and converged on core B-cell regulators, including *IKZF1, IKZF3, IRF4, TNFRSF13B, TNFSF13, and BCL2*. The isotype-specific loci, by contrast, tracked specialized function: loci exclusive to IgA supported mucosal immunity, including *LTBR* and *INAVA*[173,174]; loci for IgG mirrored effector and clearance function, including the Fcγ receptors and the glycosylating enzyme *ST6GAL1*[175,176]; and loci exclusive to IgM included the early B-cell development transcription factor *PAX5*[177].

Because a GWAS surveys the system without prior selection of loci, it can identify novel effector genes that had not been previously implicated in humoral immunity, and two of these emerge as candidate regulators of the germinal-center reaction with a strong weight of converging evidence.

The strongest evidence for a novel regulator is found for *ZNF608*. Two independent associations upstream of the gene reduce IgM while showing no significant effect on IgA or IgG. One of these associations (rs12517864-A) colocalizes with increased *ZNF608* expression in circulating memory B-cell subsets in ImmuNexUT[97], i.e. higher *ZNF608* is associated with lower IgM levels. Using public ATAC-sequencing data, I examined allelic imbalance in chromatin accessibility in heterozygous primary B cells[119] and found strong skew toward the A allele of rs12517864, at an allele fraction of 0.92 and a binomial P of 2.14×10^{-14}, i.e. the high-expressing allele is strongly associated with a localized increase in chromatin accessibility in heterozygous individuals.

The variant overlaps a binding site for the transcription factor SPI1 in primary B cells[178], and motif analysis predicts that the associated allele increases SPI1 affinity, in the direction concordant with the observed upregulation, supplying a plausible mechanism for the expression effect. SPI1 has known roles in the development of

germinal center B cells[179,180]. Single-cell RNA-sequencing data[121] from tonsillar samples show that *ZNF608* is preferentially expressed in germinal-center B cells, mirroring the expression of the germinal-center master regulator *BCL6*[181]. *ZNF608* is also known to be recurrently mutated in follicular lymphomas and acts as a tumor suppressor in CRISPR screens of diffuse large B-cell lymphoma (DLBCL) cell lines[182,183]. Colocalization, allele-dependent accessibility, a concordant motif prediction, a germinal-center expression pattern, and recurrent somatic mutation in the corresponding malignancy thus converge on *ZNF608* as a regulator of the germinal-center response, a gene with no prior published role in antibody production.

MCTP2 – a calcium binding protein with no published role in the immune system[184] – is supported by a similar body of evidence. A variant at the locus (rs6497185-C) reduces IgM, in the same direction as the *ZNF608* signals. The C allele upregulates *MCTP2* in unswitched memory B cells[97], i.e., increased expression of the target gene in B cell subsets leads to lower IgM levels analogously to *ZNF608*. Like *ZNF608*, *MCTP2* is upregulated in germinal-center B cells, and the lead variant overlaps an accessible binding site for BCL6. The shared phenotype, the shared germinal-center expression, and the location of both variants within the accessible chromatin bound by germinal-center transcription factors place *ZNF608* and *MCTP2* together as candidate components of the network governed by BCL6 and SPI1 that controls the germinal-center reaction.

Another novel gene was implicated by a signal that reduces M/AG and downregulates the vitamin D receptor gene *VDR* in plasmablasts. The lead variant for this signal (rs886441) lies within a plasmablast-accessible regulatory element[119]. Notably this regulatory element is co-bound by IRF4 and PRDM1 in public ChIP-sequencing data from *ex vivo* cultured plasma cells[185]. Motif analysis predicts that this lead variant reduces binding of IKZF1 – a key transcription factor for B-cell development[186] – in the direction of lower *VDR* expression, and rs886441 directly overlaps an IKZF1 binding site in a B cell line[187]. *VDR* is upregulated in tonsillar plasmablasts in single-cell RNA-sequencing data[121]. Vitamin D signaling has been implicated in autoimmunity, but there is limited and conflicting evidence for its role in B cells[188–190]. Our results provide human genetic evidence that vitamin D signaling regulates antibody production through an unknown mechanism in terminally differentiated plasmablasts. A further novel gene, *ARHGAP15*, is supported by a rare noncoding variant that reduces IgA and colocalizes with reduced *ARHGAP15* expression in B cells, implicating this Rac GTPase-activating protein in the regulation of B-cell signaling.

The association signals identified in our study overlap extensively with the genetics of immune-mediated disease, but the value of that overlap differs sharply between common and rare variants. Only for a small number of rare alleles does the overlap support the idea that quantitative Ig variation lies on the same continuum as immune pathology.

Cross-referencing with the GWAS Catalog and ClinVar[162,191] showed that 148 Ig-associated variants were in high LD (R2 > 0.8) with risk variants for 141 disease outcomes, most heavily for autoimmune and allergic conditions (99 variants), for B-cell malignancies (34 variants), and for immunodeficiencies or infection susceptibility (17 variants), with some variants falling into more than one category. However, it should be noted that for the majority of these overlapping disease associations we do not find clear evidence for shared mechanisms between the humoral immune system and heritable disease risk. It is not entirely surprising that variants that affect risk for B-cell-related disease will also detectably perturb circulating antibody levels via multiple mechanisms. Ig variation is highly pleiotropic – perturbations to multiple systems and stages of B-cell development affect Ig levels – and it is difficult to establish the direction of causality for overlapping disease risk variants without additional evidence. Thus, at the level of common, low-risk disease variants, the direct clinical value of Ig GWAS analysis is somewhat limited – overlap of common lupus disease risk variants with Ig-altering variants does not directly inform the diagnosis or treatment of lupus, for example.

However, a different story emerges for rare variants, which are able to identify known risk factors for uncommon diseases. For example, two alleles at *IRF4-EXOC2* and one at *GLRX5* that markedly raise IgM in our study have been previously linked to Waldenström's macroglobulinemia, a disorder defined by monoclonal IgM overproduction[192–195]. Furthermore, we recover two rare *loss-of-function variants in TNFRSF13B that reduce* total Ig (AGM) and are established risk factors for common variable immunodeficiency (CVID), a disease characterized by low antibody levels[196,197]. For both these diseases, a clear causal link can be seen between the underlying Ig trait and disease etiology. These results suggest that novel rare variants identified in our study are likely to be enriched for causal factors mediating heritable diseases related to B cells. For example, we identify multiple rare hypomorphic alleles in *FCGR2B* associated with increased IgG levels that could increase the risk of autoimmune diseases such as lupus[198].

Rare variants require high sample sizes to detect significant effects in GWAS analysis, but for rare diseases it remains logistically challenging to recruit sufficiently large cohorts. The use of population-scale quantitative trait measurements to identify clinically impactful rare variants is thus a promising future direction for the field, particularly for the identification of rare non-coding, intergenic variants that are invisible to the exome-sequencing and coding-variant focused approaches that currently dominate the field[199,200]. As biobank cohorts and whole-genome sequence databases continue to expand in size, simple quantitative immune traits like Ig levels could help uncover ultra-rare and private mutations that mediate human disease and uncover novel targets for clinical treatment and diagnosis.

Serum *trans*-pQTL analysis looks at the effect of a genetic variant across the entire serum proteome, capturing variant effects on molecules such as circulating cytokines or cell markers correlated with specific cell type frequencies. To trace these,

I integrated the Ig association signals with plasma proteomics from 54,265 UK Biobank participants[107]. A total of 237 Ig-associated variants correlated with 1,354 *trans*- and 59 *cis*-protein quantitative trait loci, influencing 730 distinct proteins, of which 12 showed significant enrichment for one or more Ig traits. To detect possible causal or mechanistic relationships between specific serum proteins and Ig isotype levels, I performed IVW regression between the effect sizes of matched variants taken from the two studies.

Thus, IVW regression can recover known stoichiometric relationships between pentameric serum IgM and specific serum proteins. IgM-associated loci overlapped strongly with the protein J-chain, across 30 variants, and with CD5L, across 23 variants, with highly significant concordance between Ig and protein effects. J-chain polymerizes IgM monomers into pentamers, and CD5L associates with IgM through J-chain, so these are the known stoichiometric subunits of circulating pentameric IgM[74]. Recovering these two strict stoichiometric relationships from genetics alone validates the inverse-variance weighted regression approach as a means of recovering meaningful biological associations between Ig isotypes and specific serum proteins.

IgA-associated loci were enriched for *trans*-pQTL effects on LY96, across 20 variants, with a concordance comparable to that of the stoichiometric IgM subunits. LY96 is a secreted co-receptor for Toll-like receptor 4 that is essential for the recognition of bacterial lipopolysaccharide[201], and although canonically myeloid it is also expressed in plasma cells and is upregulated in *ex vivo* stimulated B cells[119]. It has no known physical association with IgA, yet the strength of the coupling, on the scale of a stoichiometric subunit, points to a previously unrecognized molecular relationship between IgA and LY96. Separately, loci associated with class-switched Ig traits were enriched for *trans* effects on marker proteins of mature B and plasma cells, including FCRL5, IL5RA, MZB1, and TNFRSF17, consistent with variation in the size and activity of the mature B- and plasma-cell pool rather than with direct molecular coupling. FCRL5 and TNFRSF17 are targets of multiple myeloma immunotherapy, and plasma IL5RA and TNFRSF17 are elevated in carriers of myeloma risk alleles[202], linking these proteomic signatures to the pathology of the plasma cell.

To identify specific cell types affected by Ig association signals, I integrated our results with the BloodVariome compendium, a companion resource cataloging inherited effects on 1,005 flow cytometry traits across 127 immune cell subsets in 11,983 individuals[167]. 22 Ig-associated variants colocalized with at least one immune cell trait. Twelve of these affected B-cell traits, indicating B-cell-intrinsic effects, while the remaining 10 affected traits of other lineages, consistent with indirect regulation, primarily through helper T cells, regulatory T cells, and dendritic cells. The intrinsic effects resolved discrete control points along the differentiation axis: a variant in the *untranslated region of TNFSF13B raised* total Ig, B-cell counts, and plasma TNFSF13B together, and increased the risk of tonsillectomy, consistent

with TNFSF13B-driven B-cell expansion; a missense variant in *NIBAN3* reduced IgM and lowered naive and transitional B-cell frequencies, implicating the poorly characterized gene in early human B-cell development. Among the indirect effects, a missense variant in *INAVA* reduced the A/G ratio while decreasing CCR6 surface expression on follicular helper T cells.

For multiple traits, the rare variants of large effect accumulate not diffusely but on the core genes or signaling axes that directly regulate the process the trait models. The class-switching proxy A/G converges on the Ig heavy-chain locus, where class switching physically occurs. Total IgG converges on the Fcγ receptors and on *FCGRT*, the machinery that clears and recycles IgG. The product trait AGM, a proxy for total Ig production, converges on the TNFRSF13B–TNFSF13 axis, the ligand-receptor pair that controls plasma-cell survival and differentiation. This pattern is particularly striking for the composite traits, as it clearly demonstrates that they are effectively capturing the biological processes they were designed to model.

The heavy-chain locus on chromosome 14q32.33 encodes the constant-region genes that undergo class-switch recombination, and its complexity has long frustrated human genetic study[203,204]. We identified 17 independent signals at the locus, 12 of them novel. These signals were enriched 2.5-fold in the constant region, with five of the 17 lead variants mapping to the 3'RR1 super-enhancer that governs isotype selection[205,206]; 15 of 17 showed their strongest association with the composite ratio traits that capture class-switching dynamics. Four rare variants associated with A/G had particularly large effects, including one that tags a 28-kb deletion spanning *IGHG1* that is causal for familial IgG1 deficiency[110]. Only two of the 17 signals contained coding variants in their credible sets, indicating that heritable Ig variation at the heavy-chain locus acts overwhelmingly through non-coding or structural variation.

The heavy-chain locus also displays an emergent property that is one of the principal findings of the study. The four IgG subclasses are encoded by separate constant-region genes within the locus, and in earlier work[110] we had observed that a deletion tagged by one variant reduced IgG1 while raising IgG3, leaving total IgG unchanged. To ask whether such compensation is general, I evaluated 89 IgG-associated variants against this prior IgG1 to IgG4 subclass data using an inverse-variance weighted outlier framework, reasoning that a variant acting on total IgG should shift all subclasses proportionally, whereas a variant with compensatory effects should produce reciprocal deviations from that expectation. The analysis returned 10 outliers, of which five showed the reciprocal pattern of buffering, and all five mapped to the IGH locus. One, a splice variant in *IGHG2* that causes familial IgG2 deficiency[207], reduced IgG2 sharply while raising IgG1. No buffering signal was detected at any locus outside the heavy chain. Subclass compensation is therefore not a general feature of IgG regulation but an intrinsic systems property of the IGH locus itself, one that stabilizes total IgG output against subclass-specific

perturbation. The precise molecular mechanism by which this buffering is achieved remains to be resolved.

Rare variants influencing IgG converge on the Fcγ receptors that bind IgG[176]. At the Fcγ receptor locus we identified 24 independent signals, 20 of them novel, of which 23 influence IgG or IgG-derived composite traits. Notably, the Fcγ locus carries the four rare variants with the largest effects on IgG in the study. These include a novel rare coding variant in *FCGR2B*, p.Pro47Ser, that markedly raises IgG but is classified as neutral by computational variant effect predictors such as AlphaMissense[208]. The substituted proline lies in a disordered interdomain linker, where variant classifiers based on evolutionary conservation are known to lose sensitivity[209]. Proline is known to be enriched in such disordered regions and thought to confer structural rigidity in these domains[210]. As *FCGR2B* is an inhibitory receptor, the strong increase in IgG levels suggests that the variant acts as a hypomorph.

Outside the canonical locus, the neonatal Fc receptor FCGRT protects IgG from lysosomal degradation inside cells and recycles it into circulation, giving IgG its long serum half-life[80]. We identified a rare coding variant in *FCGRT*, p.Lys132Asn, that markedly reduces circulating IgG and is predicted to impair protein function by AlphaMissense. Analysis of a published crystal structure[211] places the affected lysine at the interface with β2-microglobulin, an interaction that is known to be essential for FCGRT function. To our knowledge, this is the first loss-of-function human population variant described in the neonatal Fc receptor. As this variant would plausibly reduce the serum half-life of clinically administered IgG, it is likely to have implications for personalized medicine[212].

The product trait AGM converged on the TNFRSF13B–TNFSF13 axis, and here the associations form allelic series at both the receptor and the ligand, with four independent alleles at each. TNFSF13, produced by myeloid cells, binds TNFRSF13B on mature B and plasma cells to promote class switching and antibody production. At *TNFRSF13B,* two rare missense variants that abrogate signaling, p.Cys104Arg in the extracellular domain and p.Ala181Glu in the transmembrane domain, sharply reduce AGM, expand pre-switched memory B cells, raise the pre-switch marker FCRL2, and lower the plasma-cell markers TNFRSF17 and IL5RA. These variants are known risk factors for CVID[196,197], and the phenotypic pattern observed here is suggestive of a developmental block at the pre-switched memory B-cell stage. A common intracellular missense variant in *TNFRSF13B* (p.Pro251Leu) produces the mirror image phenotype – raising AGM, lowering plasma FCRL2, and raising TNFRSF17 and IL5RA – consistent with enhanced TNFRSF13B signaling. Notably, the p.Pro251Leu variant is a risk factor for multiple myeloma[213], a cancer characterized by the overproliferation of plasma cells. The *final member of the TNFRSF13B allelic series* is a novel rare non-coding variant that has not been previously described – unlike the other alleles it only significantly associates with IgM without affecting IgA or IgG. The lead variant intersects an accessible element upstream of *TNFRSF13B,* and the distinct phenotypic effect

from the coding variants suggests that this regulatory element modulates a restricted stage of B-cell development.

At the TNFSF13 locus, two common non-coding variants act through separate myeloid compartments: one raises total Ig by upregulating TNFSF13 in monocytes, and the other lowers it by downregulating TNFSF13 in neutrophils. Only the monocyte variant raised plasma TNFSF13 protein levels, as I confirmed by ELISA in 793 blood donors, genetically identifying monocytes rather than neutrophils as the principal source of circulating TNFSF13 in humans[214]. Two rare non-coding *TNFSF13* variants reduce total Ig and circulating TNFSF13 levels, and thus act as hypomorphs. For one of these, which was predicted to disrupt EGR1 binding[125,127], I confirmed disruption of transcriptional activity by luciferase reporter assay and by EGR1 knockdown using siRNA. Thus, loss of EGR1 binding at the *TNFSF13* promoter leads to reduced TNFSF13 and total Ig levels. The concordant effects of all four *TNFSF13* alleles – with all four affecting the product traits and all three primary isotypes – highlights the central role of this cytokine in controlling overall Ig levels.

Across these two allelic series we find a clear logical concordance between the variant mechanisms and their downstream phenotypic effects on the B-cell system. The variants that reduce TNFRSF13B signaling and lower total Ig are the variants that raise the risk of CVID, while the variant that enhances signaling and raises Ig raises the risk of multiple myeloma. A single receptor therefore behaves as a graded rheostat whose setting maps onto opposite poles of B-cell pathology, with immunodeficiency at one end and plasma-cell malignancy at the other. Similarly, the four TNFSF13 varian*ts* show strikingly similar phenotypic signatures, with all of them influencing all three primary isotype traits and the product traits derived from them in a direction consistent with their effect on TNFSF13 expression.

Taken together, these results establish that circulating Ig levels, analyzed at population scale across nine traits, resolve the regulatory architecture of the human B-cell system. The 504 variants recover known regulators at every stage from B-cell commitment to antibody secretion. Natural variation, read out through a coarse and routine clinical measurement, thus functions as a perturbation screen of the human immune system. The logic is the same one applied in Papers I and II, in which association was carried through fine-mapping to mechanism at a single stem-cell locus; here that pipeline is applied not to one locus but to an entire system, and through an expanded sample size it is able to resolve that system genetically.

Chapter 5 – Discussion

Each of the three studies treats a quantitative trait measured in a large population as a record of perturbations that natural variation has already performed, and asks what can be inferred from those perturbations about the underlying biology. What differs between them is the scale at which that question is asked.

In the discussion that follows, I lay out some larger themes that emerge from these studies and their relation to modern human genetics as a field.

5.1 Quantitative variation as a discovery platform for clinical genetics

Ig concentrations are among the cheapest phenotypes in medicine. They are measured by routine clinical chemistry, require no specialized assay or cell isolation, and can be obtained on every participant in a cohort rather than on a selected subset. Applied at population scale across nine traits, this yielded 504 independent associations, 472 of which had not been reported before, and 397 candidate genes assigned across 290 signals.

One consequence of working at this scale is the recovery of significant numbers of rare alleles of strong effect. The rarity of such variants is a consequence of natural selection – most strong-effect variants are likely to be deleterious for fitness, so over time they are weeded out by selection and are only maintained at low frequencies[215]. This gives the characteristic relationship between allele frequency and effect size consistently observed across GWAS for multiple traits.

Such rare variants can play critical roles in heritable disease, and rare and private variants are one of the explanations for the missing heritability problem in human genetic studies[216,217]. Unfortunately, the ability to measure the effects of rare variation is limited by sample size. Although rare disease genetics can partially access this variant space through whole-genome or exome sequencing, such studies are often limited to clearly interpretable missense or structural variation, rendering the vast space of non-coding variation invisible to these study designs[199].

The main obstacle to variant discovery in rare disease genetics is recruitment rather than sequencing costs. Heritable immunodeficiencies like CVID, severe combined

immunodeficiency and the agammaglobulinemias are individually uncommon, and assembling a cohort large enough to test any particular allele requires international consortia working over years. Even the largest such collections contain at most a few thousand patients, so a variant carried by one patient in a hundred appears in a few dozen people, limiting statistical power[218]. Ascertainment is by clinical diagnosis, which is variable in its criteria and often delayed, and tends to select the severe end of the distribution. Heterozygous carriers, individuals with subclinical reductions in antibody, and people who never came to medical attention are invisible to this design.

Measuring a quantitative trait in an unselected population inverts the problem. Everyone has an Ig level, so carriers of a variant are captured independently of any disease diagnosis, and the number of carriers scales with cohort size rather than with the prevalence of a diagnosis. In a cohort of 114,697, an allele at a frequency of 0.07% is carried by roughly 160 people, an allele at 0.34% by around 780, and an allele at 0.7% by about 1,600. These are sample sizes sufficient to estimate an effect precisely, and because the phenotype is continuous, the estimate carries magnitude and direction.

In Paper III, 33 of the 504 associations involve variants at an effect allele frequency of 1% or below. Notably, these strong-effect associations tend to cluster on genes and loci known to play a central role in B-cell biology and uncover several disease-associated rare variants. At the Ig heavy-chain locus, they include a 28-kb deletion spanning *IGHG1*, which causes IgG1 deficiency, and a splice variant in *IGHG2* responsible for familial IgG2 deficiency. At *TNFRSF13B*, the coding variants p.Cys104Arg and p.Ala181Glu are established risk factors for CVID, and both reduce total Ig in our results. At *DOCK2*, a rare missense variant appears among the alleles with the largest positive effects on total Ig levels, in a gene where biallelic loss of function causes a severe combined immunodeficiency[219] – the positive effect size suggests that this could be a rare gain-of-function variant in *DOCK2*, although more work is needed to determine the effect of this coding mutation. At *FCGRT*, the rare coding variant p.Lys132Asn lies at the interface with β2-microglobulin and lowers IgG, and is the first identified population hypomorph in this gene. At *FCGR2B*, the rare p.Pro47Ser variant displays one of the strongest positive effects on IgG observed in the study, and thus likely acts as a loss-of-function variant in this inhibitory receptor. Uncommon non-coding alleles mapping to *IRF4* raise IgM and are known risk variants for B-cell malignancy, including Waldenström's macroglobulinemia, whose defining feature is overproduction of monoclonal IgM.

Our results demonstrate that subclinical shifts in Ig levels can recover known disease variants without measuring disease status directly. This suggests that novel rare variants identified in our study can be re-evaluated with a lowered multiple testing burden in clinically defined cohorts. For example, rare hypomorphic alleles in *FCGR2B* could be evaluated for their effects on the pathophysiology of autoimmune

disease, and the rare variant identified in *FCGRT* could be assessed in patients receiving IgG-based immunotherapies.

Paper I shows the same logic operating on a different trait and at a fraction of the sample size. Blood $CD34^+$ cell level is measurable by flow cytometry in any donor, and 13,167 individuals were enough to recover *CXCR4*, the gene mutated in WHIM syndrome and the drug target for plerixafor, through four independent signals including a rare variant. Additionally, a common variant affecting *PPM1H,* a phosphatase with no published role in stem-cell biology, was the most significant association identified in the study. Pilot work in our lab has begun to assess if allele carriers for these variants show altered response to stem-cell-mobilizing drugs in the clinic[220]. Although these stem-cell-influencing variants are not rare, the same issue with cohort recruitment for rare immunodeficiencies arises – recruiting 10,000 patients receiving stem-cell transplantation for a full GWAS analysis would be prohibitively time-consuming and costly, but GWAS results from a quantitative trait correlated to the clinical phenotype can enable hypothesis-driven evaluation of predefined variants at a lower multiple testing burden.

5.2 Discovery of novel B-cell genes and pathways

Of the 504 associations identified in Paper III, 472 had not been reported previously, and several of the candidate genes assigned to them have no described role in antibody production. Three are worth setting out in detail, because in each case the genetic signal is accompanied by regulatory and expression evidence that places the gene in a specific stage of B-cell differentiation.

ZNF608 and *MCTP2* are uncovered through variants influencing IgM levels, and both genes are upregulated in the germinal center. At *ZNF608*, the IgM-decreasing allele of rs12517864 raises *ZNF608* expression in memory B-cell subsets. The variant falls in a region accessible in light-zone germinal-center B cells and occupied by SPI1 in B cells, and the substitution alters a base within the SPI1 recognition motif. Allelic imbalance in ATAC-sequencing reads from heterozygous primary B cells is strong, at an allele fraction of 0.92, so the two alleles differ substantially in chromatin accessibility in the cells where the gene acts. At *MCTP2*, the IgM-decreasing allele of rs6497185 raises *MCTP2* expression, most clearly in unswitched memory B cells, and the variant lies in a region accessible in light-zone germinal-center B cells and occupied by BCL6 in germinal-center B cells. Single-cell expression data from tonsil place both genes at their highest levels in light-zone germinal-center B cells, alongside BCL6 itself. *ZNF608* is additionally a recurrent somatic mutation target in DLBCL and follicular lymphoma[182,183], the malignancies of the germinal-center reaction, so a gene identified from a blood test in healthy people coincides with one identified from tumor sequencing.

VDR is the clearest instance of a gene relocated to a new compartment by these data. The vitamin D receptor has been studied in immunology almost entirely through T cells, where it mediates the effects of vitamin D on helper T-cell differentiation, suppresses Th1 and Th17 responses, and promotes regulatory T-cell phenotypes[221]. Its function in B cells has attracted far less attention[222], and *in vivo* evidence for a role in antibody production has been lacking[190]. Here, the causal variant rs886441-G reduces M/AG and *VDR* expression specifically in plasmablasts[97]. The variant lies within a binding motif for IKZF1, which the A allele is predicted to disrupt, in a region occupied by IKZF1 in a B cell line and accessible by single-cell ATAC-sequencing in plasmablasts[121]. Chromatin immunoprecipitation in plasmablasts differentiated *ex vivo* from memory B cells shows that the same region is occupied by IRF4 and PRDM1 and carries H3K27 acetylation[185], so the variant sits in an active regulatory element bound by the two transcription factors that drive terminal plasma-cell differentiation. Single-cell expression data from tonsil[121] are consistent, with *VDR* expression concentrated in the plasmablast compartment rather than in germinal center or naive B cells.

The *VDR* result therefore rests on two independent kinds of evidence pointing to the same cell type. The expression quantitative trait locus establishes that the variant changes *VDR* transcript levels in plasmablasts and not in other sorted immune populations. Together with the occupancy of the element by IRF4 and PRDM1, this constitutes human *in vivo* genetic evidence that *VDR* acts in the antibody-secreting compartment, at the plasmablast stage, and that natural variation in its regulation shifts the balance between unswitched and class-switched Ig. Given how widely vitamin D status varies across humans and its known role in autoimmunity, these results could turn out to have clinical implications for the treatment of autoimmune disease, as most of the research on vitamin D in the immune system has focused on its role in the T-cell lineage.

5.3 Trait engineering as a general principle

The composite traits employed in Paper III are simply linear combinations of standardized, log-transformed Ig measurements, originally introduced in our previous GWAS of Ig levels[168]. The ratios – A/M, G/M, and A/G – are meant to capture isotype-specific class-switching biases, by separating them from genetic effects that act on total Ig levels. The products, AG and AGM, do the opposite and reinforce what the components share. The ninth trait, M/AG, sets the unswitched output against the switched output and therefore senses pre-switch versus post-switch bias with statistical support drawn from all three isotypes. While this logic maps intuitively to our understanding of B-cell development, it is also plausible that the composite traits do not carry any more information than the primary traits used to derive

them. This criticism was even raised by our colleagues at deCODE in initial discussions about the results.

Our results show conclusively that these simple linear combinations are able to increase genetic signal in a biologically interpretable and coherent manner, a finding that has implications for the study of quantitative traits more broadly.

The composite traits increase signal yield – association testing across the nine traits identified 1,144 independent signals, which inter-trait colocalization consolidated to 504 unique variants, and the ratio traits contribute a substantial share of these; A/G alone carries 51 associations found for no other trait, more than any primary isotype contributes exclusively. The ratio traits also show an enrichment of signal at the IGH locus, where 15 of the 17 independent signals showed their strongest association with a composite ratio trait rather than with any individual isotype. As the IGH locus is the physical location where class switching occurs, this provides further evidence that the ratio traits are more capable of detecting functional variation related to class switching.

The principle generalizes beyond Ig. Wherever a panel of correlated quantitative measurements is available, derived traits can be constructed to isolate the processes that convert one measurement into another, or to aggregate what several measurements share: ratios among blood-cell types that read out lineage commitment, metabolite ratios that read out enzymatic steps, lipid subfraction ratios that read out particle remodeling, and frequencies of cell states derived from flow-cytometric panels of the kind assembled in BloodVariome[167], which read out transitions between states. In each case the derived trait can, in principle, bring a specific process into genetic focus that the raw measurements leave blurred. The limit of the principle is that composites can be constructed without end, and most possible combinations will correspond to no biological process at all. The discipline that makes trait engineering more than a fishing expedition is that each composite must be motivated by a known process before it is tested, so that a discovered signal has an interpretation waiting for it rather than an interpretation invented after the fact. Applied without that discipline, the method would inflate the multiple-testing burden and return combinations whose associations meant nothing.

5.4 Systems-level properties: buffering and compensation

The clearest systems-level finding in the thesis is the subclass buffering observed at IGH. A prior study of IgG1 to IgG4 subclass levels identified a deletion spanning *IGHG1* that reduced IgG1 while raising IgG3, which suggested compensation of one subclass for the other[110]. The present study is able to test this isolated

observation systematically. Across 89 IgG-associated variants that could be mapped onto the earlier subclass analysis, the effect of each variant on an individual subclass was regressed against its effect on total IgG, and outliers from that relationship were classified according to whether their subclass effects were reciprocal, as compensation requires, or merely disproportionate.

Five out of ten outlier variants in this analysis showed reciprocal subclass effects, and all five mapped to IGH. The clearest is rs191766497-T, a splice variant in *IGHG2* that causes familial IgG2 deficiency: it lowers IgG2 by a large margin and is accompanied by compensatory increases in the other three subclasses. The remaining IGH outliers behave in the same way. Outliers were also identified outside IGH – at the Fcγ receptor locus, at HLA, and at KPNB1 – but none of them showed reciprocal behavior; they depart from the regression without redistributing among subclasses. Subclass compensation is therefore not a general property of IgG regulation but an intrinsic property of the Ig heavy-chain locus.

The underlying mechanism for Ig subclass buffering is not resolved by the study; the paper itself notes that the molecular basis warrants future mechanistic work. The general question the finding raises is larger than the locus. If a locus can redistribute its effects among components while holding an aggregate constant, then a study that measures only the aggregate will fail to detect this, and the genetic architecture of the components will be invisible at the level ordinarily assayed. This is a specific instance of a broad problem. Much of quantitative genetics measures aggregate outputs, and compensation of the kind seen at IGH would cancel real effects before they reach the measured phenotype, so that the apparent absence of a genetic contribution reflects the choice of what to measure rather than the absence of underlying variation.

5.5 Forward genetics with natural variation as the mutagen

A mutagenesis screen of the kind that Nüsslein-Volhard and Wieschaus carried out in *Drosophila* saturates a genome with induced alleles, scores each line for a phenotype of interest, and thereby recovers the genes required to build the phenotype without any prior hypothesis about which genes those are[10]. Population-scale human genetics can be read as the same experiment with two substitutions. Standing variation replaces induced mutation as the source of alleles, and a biobank replaces controlled crosses as the population in which phenotypes are scored.

The transition to natural variation from controlled breeding and induced mutagenesis costs a great deal scientifically. There is no control over the allelic spectrum, since one takes the variation the population happens to carry. Natural selection has

already removed the most damaging alleles, so the strongest perturbations available to a mutagenesis screen are systematically depleted in a human population, and effect sizes are correspondingly small for the majority of alleles identified[215]. There is no saturation in the strict sense, because one cannot guarantee that every gene has been hit by an informative allele; absence of a signal is not evidence that a gene is uninvolved. The rarest and most informative alleles are in addition the hardest to follow up – in Paper I the rare *CXCR4* variant rs555647251 was not polymorphic at all in the $CD34^+$ cell expression cohort, so its effect on *CXCR4* expression could not even be tested.

However, natural variation at large population scales also adds considerable scientific value that model-organism studies cannot provide. The perturbations are in the right species, so no cross-species inference is required to reach human biology. They act in the right cell types, *in vivo,* over a human lifetime, so the phenotype integrates development and homeostasis as they occur rather than as they are reconstructed in a model raised in a controlled environment. And the phenotype is measured in the physiological range rather than at the knockout extreme, so the genetics recovered is the genetics of tuning a working system rather than of breaking it, which is the regime relevant to common human variation and to most of medicine.

Paper II illustrates the difference between model organism findings and the more subtle effects that can be discovered using human population variation. Conditional deletion of *Gata2* in mouse hematopoietic cells causes complete loss of adult stem and progenitor cells[157], which establishes that the factor is necessary and says nothing about how its activity is graded in a functioning system. The human common variant rs10864368 affects stem and progenitor levels but does not directly touch GATA2 itself; rather it abrogates one GATA2 binding site among the thousands the factor occupies, reducing local chromatin accessibility at a single regulatory element and lowering expression of the target gene, and the phenotypic consequence is a shift in circulating $CD34^+$ cell levels within the normal range. These weak, common non-coding effects can thus help elucidate physiological gene regulation at the level of individual transcription factor binding sites, in a manner that would be invisible to traditional knockout studies or Mendelian disease genetics.

In Paper III, the increased sample size enables detection of allelic series and dosage effects in core Ig-regulating genes. The variants *comprising the TNFRSF13B–TNFSF13 allelic series* range from loss-of-function variants that abrogate signaling and cause immunodeficiency, to a common intracellular variant that raises antibody output and increases risk for multiple myeloma, tracing a graded continuum of pathway activity that no single engineered knockout would produce.

Natural variation also supplies allele classes that could be missed by computational predictions of variant effect that are now widely used to assess variant pathogenicity[208]. The *FCGR2B* missense variant p.Pro47Ser has an effect allele frequency of 0.07% and raises IgG by 0.77 standard deviations, one of the largest effects in the

study, in the direction expected for loss of the receptor's inhibitory function. Computational predictors classify it as benign, with an AlphaMissense score of 0.085, because the residue lies in a disordered interdomain linker adjacent to the first C2-type domain of the ectodomain, and variant effect predictors are known to have reduced sensitivity in disordered regions where no structural template constrains the inference.

5.6 Quantitative traits and disease gene discovery

A thread running through all three papers is that studying variation in the healthy range recovers genes that matter in disease. Variation within the normal clinical range and pathological disease are not separate phenomena but the same axis of biology sampled at different points, so a gene whose common variation nudges a trait within the normal range is a candidate for the gene whose rarer or more severe perturbation carries the trait out of that range into pathology. A study of healthy variation is therefore a study of disease genes conducted at the mild end of the same distribution.

Paper I demonstrates the point at small sample size. Its Mendelian disease genes, its somatic cancer drivers, and its overlap with mature blood-cell traits are set out in Section 4.1. All of these connections to disease are derived from nine association signals from a flow cytometry trait measured in blood donors and primary care patients, with no clinical phenotype examined.

Paper II reaches a disease gene by a different route. The causal variant does not lie in a known disease gene, it disrupts a single binding site for GATA2, a factor whose germline mutations cause GATA2 deficiency, a heritable syndrome combining immunodeficiency with myelodysplasia. A common non-coding variant carried by roughly half the population therefore recovers a Mendelian disease gene not by perturbing it but by perturbing one of the sites at which it acts. Where a coding variant implicates a gene, a regulatory variant of this kind implicates the *trans*-acting factor that binds it. The downstream consequences of a single altered GATA2 binding site could also be elucidated – the same variant associates with higher myeloid cell frequencies and lower lymphocyte percentage in large blood-cell studies, so a regulatory change acting in the progenitor compartment propagates to the composition of the mature blood.

The disease associations for Paper III have already been addressed in Section 5.1 in terms of rare variant discovery, but these results can also be seen through a more gene-centric perspective. Just as in Paper I, we consistently find a biologically coherent connection between the direction of effect on the quantitative trait and the disease phenotype. Three uncommon alleles that raise IgM, two at *IRF4* and one at *GLRX5*, associate with Waldenström's macroglobulinemia, a disorder whose

hallmark is monoclonal IgM production, so the direction of the healthy-range effect and the direction of the disease correspond. Two rare loss-of-function variants in *TNFRSF13B* that lower total Ig are established risk factors for CVID, and a novel rare missense variant in *DOCK2*, an established severe immunodeficiency gene, increases total Ig suggesting that it may act as a gain-of-function mutation.

The TNFRSF13B–TNFSF13 series elegantly demonstrates how gene dosage and direction of effect can create logically concordant effects on heritable disease risk. *The TNFRSF13B receptor variants do not act* in a single direction. The two rare missense alleles that abrogate signaling reduce total Ig, expand pre-switched memory B cells, raise plasma FCRL2, and lower the mature B cell and plasma-cell markers TNFRSF17 and IL5RA, a pattern consistent with a developmental bottleneck at the class-switch transition, and they raise the risk of CVID. The common intracellular variant p.Pro251Leu shows the reciprocal pattern, raising AGM, lowering FCRL2, raising TNFRSF17 and IL5RA, and raising the risk of multiple myeloma. In this case the mirrored effects of these two alleles are consistent with the mirrored nature of the two diseases themselves – CVID is characterized by very low Ig levels while multiple myeloma is an overproliferation of Ig-producing plasma cells.

The quantitative trait is thus not a weaker substitute for a disease study but a different instrument that returns signed, graded information that connects specific physiological endophenotypes to clinical disease risk, in a manner inaccessible to case-control study designs. As biobanks increase in sample size and diversity, it is likely that these aspects of quantitative trait analysis will increasingly impact clinical discovery.

5.7 Therapeutic leverage

Human genetic support for a target raises the probability that a drug against it will succeed in clinical development, a relationship documented across large numbers of programs and noted in Chapter 1[35]. Several of the genes implicated by this work could in principle be pursued as drug targets. Similarly, several variants could be assessed for clinical impact to facilitate personalized medicine. Two are worth setting out in detail, because in each case the direction of effect is logically concordant with the underlying clinical application, and the supporting evidence goes beyond the association itself.

PPM1H emerges from Paper I as a candidate inhibition target for mobilizing HSPCs, and the case for it has three parts. The direction of effect suggests that inhibition or degradation could mobilize stem cells, since the allele associated with higher blood $CD34^+$ cell levels lowers *PPM1H* expression in $CD34^+$ cells. The direction of effect is supported by a *cis*-expression quantitative trait locus in sorted

cells from 155 blood donors and by short hairpin RNA knockdown of *PPM1H* in primary $CD34^+$ umbilical cord blood cells, which raised the proportions of $CD34^+$ and $CD34^+$ $CD90^+$ cells relative to control. Furthermore, *PPM1H* is preferentially expressed in hematopoietic stem and progenitor populations relative to other blood-cell types. Its credibility is further supported since the same study recovers *CXCR4*, whose antagonist plerixafor is an approved mobilizing agent.

FCGRT is an established therapeutic axis, already the target of agents that modulate IgG recycling[212]. The rare *FCGRT* variant p.Lys132Asn lowers circulating IgG significantly, and a co-crystal structure of the receptor in complex with β2-microglobulin places Lys132 at the interface between them. As this interaction is known to be essential for protein function, it is likely that the reduced IgG levels observed in allele carriers result from a loss of FCGRT function. Given its known role in stabilizing circulating IgG, it is plausible that a partial loss of receptor function in *FCGRT* could impact individual response to IgG-based immunotherapies and autoimmune diseases mediated by pathogenic IgG. Further work is needed to determine the impact of this variant in these clinical contexts.

5.8 Limitations

The most consequential limitation of these studies is that they are restricted to individuals of European ancestry. Paper I was conducted in southern Sweden, with a discovery set of 10,949 individuals of Swedish ancestry and a follow-up set of 2,218 individuals of other European ancestry. Paper III draws on Iceland, on Swedish cohorts, and on the British-Irish subset of UK Biobank. The cohorts are therefore predominantly of Northern European ancestry throughout. Effect sizes, allele frequencies, and patterns of LD will not transfer uniformly to other populations, and the rare variants are, by their nature, population-specific. Paper I provides the concrete instance: all of its lead variants were polymorphic across the geographic super-populations of the 1000 Genomes compendium except for the rare *CXCR4* variant rs555647251, which was polymorphic only in the European and American super-populations[223]. This limits the generality of the architectural claim, because the demonstration that large-effect variation clusters at process-defining loci has been made in one ancestral background, and it limits the equity of any clinical application derived from the work[224].

Most of the mechanistic inferences in Paper III rest on genetic and molecular association rather than on experimental perturbation, and the strength of the evidence is uneven across the loci discussed in this chapter. The strongest evidence chains in the thesis are those at *PPM1H* and at *ENO1*, where a causal variant was supported by reporter assays, by transcription factor occupancy, by knockdown or CRISPR disruption, and by allelic imbalance in chromatin accessibility in primary cells. In

Paper III only the *TNFSF13* result reaches a comparable standard, combining a reporter assay with siRNA knockdown of EGR1, and supported by transcription factor binding data in primary macrophages derived from monocytes. At *ZNF608*, *MCTP2*, and *VDR* the evidence is a concordant assembly of eQTL direction, chromatin accessibility, and transcription factor occupancy with a matching motif prediction, supported at *ZNF608* by allelic imbalance in accessibility in primary B cells. These are predictions consistent with the data rather than demonstrations, and they should not be weighed equally with the MYB and GATA2 dissections or with the EGR1 result, which were ultimately validated through wet lab experiments. The systems-level claims rest on the aggregate of many such assignments and inherit the uncertainty of the weaker ones.

Several causal assignments remain unresolved, and two kinds of ambiguity should be distinguished. The first concerns which variant acts. Statistical fine-mapping alone rarely reaches a single variant: fine-mapping resolved 70 of the 504 Ig signals to a single putative causal variant, and at the 1p36.23 locus of Paper II the association comprised 17 variants in high LD across approximately 14 kb, where it was the chromatin accessibility filter, not the association statistics, that narrowed the field to two candidate variants before a reporter assay could distinguish them. Even where a variant has been functionally distinguished, the possibility that a linked variant contributes independently is not excluded, since the elimination is by assay rather than by exhaustive editing of all linked variants.

The second concerns which gene is affected. In Paper III, candidate genes were assigned to 290 of the 504 signals on the basis of coding variation or eQTL colocalization, which leaves 214 signals for which the nearest gene was used in the absence of molecular support. Proximity is an unreliable guide, since a regulatory element frequently acts on a gene other than its nearest neighbor. Where two genes share a regulatory element, the ambiguity may be irreducible from association data alone, as at 1p36.23, where the element contacts both the *ENO1* and the *RERE* promoter and the lead variant is a *cis*-expression quantitative trait locus for both.

The *trans*-pQTL analyses use inverse-variance weighted regression, a method borrowed from Mendelian randomization, but no causal claim is made from it, because the exclusion restriction that would justify a causal interpretation is unfalsifiable in this setting. The internal calibration described in Section 4.3 is what makes this analysis usable at all, and it also marks the boundary of what can be concluded. The IgM relationships are interpretable because the physical basis is known; the IgA and LY96 relationship is of comparable statistical strength but has no known physical basis, since LY96 is a secreted co-receptor for Toll-like receptor 4 with no described interaction with IgA. Whether the relationship reflects a genuine coupling between IgA regulation and innate immune signaling is not determined by these data and can only be addressed through functional work.

For Paper I our lab generated its own eQTL data in circulating $CD34^+$ cells, so the expression measurement was made in the cell type the trait is defined by. Paper III had no equivalent, and its colocalization rests on public sorted-cell resources. Those resources cover a range of immune cell types but not every state relevant to Ig biology, and in particular not the germinal center, which is where the candidate genes at several of the most fully worked loci are most highly expressed. Absence of colocalization is therefore weak evidence of absence, and this limitation can only be resolved through more comprehensive single-cell eQTL resources spanning diverse tissue types at large sample sizes.

Finally, both trait classes are aggregate outputs that do not identify which underlying variable a variant has moved. Circulating Ig reports the steady-state level of antibody in the blood and captures neither antigen specificity nor context-dependent responses, a limitation the study states itself. Blood $CD34^+$ cell level is subject to the same ambiguity, since it can reflect the rate at which progenitors leave the marrow, the time they spend in the circulation, or the size of the stem-cell pool, and an association identifies the gene without identifying which of these it acts on. In both trait classes, the genetics recovered is only the genetics of a system's resting physiological output at the time of measurement. Further highlighting this point, recent genetic analysis of conventional blood traits has shown that individuals display stable setpoints for hematological traits over their lifetime, but these can only be ascertained if they are measured longitudinally over time[225]. Our study design would fail to capture this component of variation.

5.9 Contribution

As with any genetic study, the primary contribution of these studies is the identification of novel loci, variants and genes that influence defined biological processes. Paper I is the first genome-wide association study of its trait, Paper III resolves the architecture of the Ig traits at a scale earlier studies could not reach, and integration with epigenomic data refines these findings into specific gene regulatory networks of transcription factors and regulatory elements central to the analyzed trait. These discoveries are now raw material for future functional work in hematopoietic stem cell and B-cell biology – we have only begun to scratch the surface of the novel findings made in these papers.

The second contribution is a resource rather than a result. The fine-mapped signals, their credible sets, the candidate genes and cell types assigned to them, and the summary statistics deposited in the GWAS Catalog are all available for reuse. For the Ig traits in particular this constitutes a reference against which variants found in patients with humoral immune disorders can be interpreted, since a variant of uncertain significance in a gene such as *TNFRSF13B* or *FCGRT* can be checked

against a measured, signed effect on antibody levels in an unselected population. The value of that reference is bounded by the ancestry limitation set out in Section 5.8, since it will interpret some patients better than others.

Chapter 6 – Conclusions and future perspectives

6.1 Conclusions

The work in this thesis supports four conclusions. First, a quantitative trait measured in a large population can be carried from a genome-wide association through fine-mapping to a defined molecular mechanism, as demonstrated at *PPM1H* through MYB, at 1p36.23 through GATA2, and at *TNFSF13* through EGR1. Second, at sufficient sample size the same approach resolves not merely a list of loci but the regulatory organization of a system, including properties such as the buffering of IgG subclasses that belong to a locus rather than to any individual variant. Third, composite traits constructed in advance from a known biological process recover that process genetically, so that the design of the phenotype is as consequential for what a genome-wide association study returns as the size of the cohort. Fourth, variation within the normal clinical range and disease-causing variation lie on the same axis, and the direction of effect on the quantitative trait carries information about the direction of clinical risk, as seen most clearly for *TNFRSF13B*.

6.2 Completing the mechanistic chain for novel genes

Paper III assigns a candidate gene to 290 of the 504 signals, and for most of them the evidence is statistical and correlative rather than experimental. Converting any one of these candidates into an established mechanism requires the same three steps: editing the variant at its endogenous position, connecting the element to the gene independently of the variant, and determining what the gene does in the cell type where the trait is set. *ZNF608* is the candidate for which this chain is furthest advanced, and it serves here as the example.

The evidence assigning *ZNF608* to the germinal-center reaction combines an expression quantitative trait locus, a concordant motif prediction, and near-complete allelic imbalance in chromatin accessibility in primary B cells. The last of these is a controlled observation rather than a prediction, but the causal role of the variant has still not been shown directly, and the immediate work is to close that gap. Editing

rs12517864 at its endogenous position in a germinal-center-like B-cell line, with *ZNF608* expression as the readout, would establish the variant as causal in its native chromosomal context, and silencing the element itself would confirm the link between element and gene independently of the variant. Beyond the variant, the function of *ZNF608* in B cells is unknown. CUT&RUN[226] for *ZNF608* in primary germinal-center B cells sorted from tonsil would establish whether the protein binds chromatin directly and, if so, where, allowing it to be placed within the transcriptional network operating in the germinal center alongside BCL6 and SPI1. Given the recurrent somatic mutation of *ZNF608* in germinal-center-derived lymphomas[183], this would also bear on its role in malignancy.

6.3 Testing variants and elements in parallel

Locus-by-locus dissection of the kind described above proceeds at a rate of roughly one locus per project, and the 504 signals identified here will not be resolved that way. Two approaches offer throughput. An MPRA tests thousands of credible-set variants for allele-dependent regulatory activity in a single experiment, allowing direct experimental evaluation of causal variants in loci where supporting evidence is currently lacking[133]. A base editor screen installs the coding alleles themselves rather than disrupting the gene entirely, allowing systematic evaluation of a broad spectrum of coding alleles across the entire protein[227]. The rare *FCGRT* variant p.Lys132Asn is a natural test case: it is predicted to sit at the interface with β2-microglobulin and to lower IgG, and installing it alongside related alleles across the receptor would test that prediction against a measured phenotype in a single experiment.

6.4 Further dissection of regulatory organization

Paper III describes properties that belong to loci and to the system rather than to individual variants: the buffering of IgG subclasses at IGH, the allelic series at TNFRSF13B and TNFSF13, and the convergence of process-defining genes on the trait that matches their process. These are descriptions of regulatory organization, and none of them yet has a mechanism. Establishing what such organization consists of molecularly is a distinct line of work from resolving individual loci, and the compensation observed at IGH is the clearest instance available here.

The molecular basis of the compensation observed at IGH remains open, and it is a considerably harder experimental problem than the loci above. Testing it requires a system in which class-switch recombination can be manipulated and the resulting subclass distribution measured, and no such system exists for human B cells.

Primary human B cells are difficult to maintain and to switch efficiently in culture, and the CRISPR screens that have dissected class switching to date were carried out in the mouse CH12 line[228–230], which switches to a single isotype and cannot report redistribution among the four human IgG subclasses. Resolving the mechanism is therefore better suited to laboratories with established expertise in primary human B-cell culture and in the biology of the heavy-chain locus than to the genetic approach taken here. The general requirement for such functional work is a human B-cell system in which the process that generates a trait can be perturbed, and the trait itself measured as the readout. No such system exists for human class switching, and the same gap applies to many of the other processes described here. Building them is the step that would convert the systems-level observations of this thesis from descriptions into testable mechanisms.

6.5 Clinical follow-up

A quantitative trait association supplies something that a target identified in a model system does not: the direction of effect of altering the gene in humans, measured in the tissue where the gene acts. Developing the identified genes and variants into potential targets for therapy or diagnosis is the natural translational follow-up from all three papers.

Three lines of clinical work follow from the targets set out in the previous sections. A program to identify inhibitors of *PPM1H* is underway in collaboration with the Chemical Biology Consortium Sweden. The variants influencing blood $CD34^+$ cell levels are being tested by colleagues in our lab against response to mobilizing agents in patients undergoing stem-cell collection, as a small pre-specified set rather than a new genome-wide study[220]. The rare Fc receptor variants are suited to the same kind of directed follow-up in modestly sized cohorts, *FCGR2B* p.Pro47Ser in autoimmune disease and *FCGRT* p.Lys132Asn in patients receiving IgG-based therapies.

6.6 Closing remarks

The work presented here began with a simple premise, that measurements already made in large numbers of people contain information about how the underlying biology is built, and that this information can be extracted if the associations are followed carefully enough. The three papers show what that yields at two different scales, from a single regulatory element in the stem-cell compartment to the genetic architecture of the antibody response. Much of what has been identified remains to

be characterized, and the loci and genes described here are best regarded as a starting point for the functional work that follows.

References

1. Fisher, R. A. *The Genetical Theory of Natural Selection*. (Clarendon Press, Oxford, 1930).
2. McIntosh, J. R. & Hays, T. A Brief History of Research on Mitotic Mechanisms. *Biology* **5**, 55 (2016).
3. Jones, G., Kleckner, N. & Zickler, D. Meiosis through three centuries. *Chromosoma* **133**, 93–115 (2024).
4. Moore, J. A. Thomas Hunt Morgan—The Geneticist. *Integr. Comp. Biol.* **23**, 855–865 (1983).
5. Allen, G. E. Hugo de Vries and the reception of the 'mutation theory'. *J. Hist. Biol.* **2**, 55–87 (1969).
6. Fisher, R. A. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* **52**, 399–433 (1918).
7. Alberts, B. *et al. Molecular Biology of the Cell*. 6th edn. (Garland Science, New York, 2014).
8. Knoblich, J. A. Asymmetric cell division: recent developments and their implications for tumour biology. *Nat. Rev. Mol. Cell Biol.* **11**, 849 (2010).
9. Peifer, M. 40 years since the Heidelberg genetic screen that revolutionized developmental and cell biology. *Development* **151**, (2024).
10. Wood, J. A. *et al.* Mutations affecting segment number and polarity in Drosophila. *Nature* **287**, 795–801 (1980).
11. Wagner, G. P., Pavlicev, M. & Cheverud, J. M. The road to modularity. *Nature Reviews Genetics* **8**, 921–931 (2007).
12. Taipale, J. & Beachy, P. A. The Hedgehog and Wnt signaling pathways in cancer. *Nature* **411**, 349–354 (2001).
13. Appleby, M. W. & Ramsdell, F. A forward-genetic approach for analysis of the immune system. *Nature Reviews Immunology* **3**, 463–471 (2003).
14. Marx, J. L. The Cystic Fibrosis Gene Is Found. *Science* **245**, 923–925 (1989).

15. Riordan, J. R. *et al.* Identification of the cystic fibrosis gene: Cloning and characterization of complementary DNA. *Science* **245**, 1066–1073 (1989).

16. MacDonald, M. E. *et al.* A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971–983 (1993).

17. Miki, Y. *et al.* A Strong Candidate for the Breast and Ovarian Cancer Susceptibility Gene BRCA1. *Science* **266**, 66–71 (1994).

18. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

19. Craig Venter, J. *et al.* The Sequence of the Human Genome. *Science* **291**, 1304–1351 (2001).

20. Abdellah, Z. *et al.* Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).

21. Belmont, J. W. *et al.* A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).

22. Belmont, J. W. *et al.* The International HapMap Project. *Nature* **426**, 789–796 (2003).

23. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* **22**, 139–144 (1999).

24. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95–108 (2005).

25. Johnson, G. C. L. *et al.* Haplotype tagging for the identification of common disease genes. *Nat. Genet.* **29**, 233–237 (2001).

26. Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).

27. Burton, P. R. *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

28. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).

29. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* **9**, 356–369 (2008).

30. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565–569 (2010).

31. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics* **19**, 491–504 (2018).

32. Lappalainen, T. & MacArthur, D. G. From variant to function in human disease genetics. *Science* **373**, 1464–1468 (2021).

33. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nature Reviews Genetics* **21**, 292–310 (2020).

34. Cuomo, A. S. E., Nathan, A., Raychaudhuri, S., MacArthur, D. G. & Powell, J. E. Single-cell genomics meets human genetics. *Nature Reviews Genetics* **24**, 535–549 (2023).

35. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nature Genetics* **47**, 856–860 (2015).

36. King, E. A., Wade Davis, J. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet.* **15**, e1008489 (2019).

37. Minikel, E. V., Painter, J. L., Dong, C. C. & Nelson, M. R. Refining the impact of genetic evidence on clinical success. *Nature* **629**, 624–629 (2024).

38. Kullo, I. J. Clinical use of polygenic risk scores: current status, barriers and future directions. *Nature Reviews Genetics* **27**, 246–263 (2025).

39. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* **19**, 581–590 (2018).

40. Sankaran, V. G. & Orkin, S. H. Genome-wide association studies of hematologic phenotypes: A window into human hematopoiesis. *Curr. Opin. Genet. Dev.* **23**, 339–344 (2013).

41. Doulatov, S., Notta, F., Laurenti, E. & Dick, J. E. Hematopoiesis: A human perspective. *Cell Stem Cell* **10**, 120–136 (2012).

42. Becker, A. J., McCulloch, E. A. & Till, J. E. Cytological Demonstration of the Clonal Nature of Spleen Colonies Derived from Transplanted Mouse Marrow Cells. *Nature* **197**, 452–454 (1963).

43. Notta, F. *et al.* Isolation of single human hematopoietic stem cells capable of long-term multilineage engraftment. *Science* **333**, 218–221 (2011).

44. Notta, F. *et al.* Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* **351**, (2016).

45. Kreso, A. & Dick, J. E. Evolution of the cancer stem cell model. *Cell Stem Cell* **14**, 275–291 (2014).

46. Laurenti, E. & Göttgens, B. From haematopoietic stem cells to complex differentiation landscapes. *Nature* **553**, 418–426 (2018).

47. Krause, D. S., Fackler, M. J., Civin, C. I. & May, W. S. CD34: Structure, Biology, and Clinical Utility. *Blood* **87**, 1–13 (1996).

48. Berenson, R. J. *et al.* Engraftment After Infusion of CD34+ Marrow Cells in Patients With Breast Cancer or Neuroblastoma. *Blood* **77**, 1717–1722 (1991).

49. Gambell, P. *et al.* Peripheral Blood CD34+ Cell Enumeration as a Predictor of Apheresis Yield: An Analysis of More Than 1,000 Collections. *Biology of Blood and Marrow Transplantation* **18**, 763–772 (2012).

50. Morrison, S. J. & Scadden, D. T. The bone marrow niche for haematopoietic stem cells. *Nature* **505**, 327–334 (2014).

51. Petit, I. *et al.* G-CSF induces stem cell mobilization by decreasing bone marrow SDF-1 and up-regulating CXCR4. *Nature Immunology* **3**, 687–694 (2002).

52. DiPersio, J. F. *et al.* Phase III prospective randomized double-blind placebo-controlled trial of plerixafor plus granulocyte colony-stimulating factor compared with placebo plus granulocyte colony-stimulating factor for autologous stem-cell mobilization and transplantation for patients with non-Hodgkin's lymphoma *Journal of Clinical Oncology* **27**, 4767–4773 (2009).

53. Olivieri, J. *et al.* Predicting failure of hematopoietic stem cell mobilization before it starts: the predicted poor mobilizer (pPM) score. *Bone Marrow Transplantation* **53**, 461–473 (2018).

54. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415-1429.e19 (2016).

55. Chen, M. H. *et al.* Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell* **182**, 1198-1213.e14 (2020).

56. Vuckovic, D. *et al.* The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell* **182**, 1214-1231.e11 (2020).

57. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics* **48**, 1193–1203 (2016).

58. Ulirsch, J. C. *et al.* Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nature Genetics* **51**, 683–693 (2019).

59. Lopez de Lapuente Portilla, A. *et al.* Genome-wide association study on 13 167 individuals identifies regulators of blood CD34+ cell levels. *Blood* **139**, 1659–1669 (2022).

60. Ali, Z. *et al.* Functional dissection of the gene regulatory mechanism underlying variation in blood CD34+ cell levels at 1p36.23/ENO1. *Scientific Reports* **16**, 558 (2026).

61. Schroeder, H. W. & Cavacini, L. Structure and function of immunoglobulins. *Journal of Allergy and Clinical Immunology* **125**, S41–S52 (2010).

62. 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 (2015).

63. Ali, Z. *et al.* Population-scale immunoglobulin genetics resolves the human B-cell system. *bioRxiv* 2026.03.30.715216 (2026) doi:10.64898/2026.03.30.715216.

64. Pieper, K., Grimbacher, B. & Eibel, H. B-cell biology and development. *Journal of Allergy and Clinical Immunology* **131**, 959–971 (2013).

65. Jung, D., Giallourakis, C., Mostoslavsky, R. & Alt, F. W. Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. *Annu. Rev. Immunol.* **24**, 541–570 (2006).

66. Victora, G. D. & Nussenzweig, M. C. Germinal Centers. *Annu. Rev. Immunol.* **40**, 413–442 (2022).

67. Shaffer, A. L. *et al.* XBP1, downstream of Blimp-1, expands the secretory apparatus and other organelles, and increases protein synthesis in plasma cell differentiation. *Immunity* **21**, 81–93 (2004).

68. Slifka, M. K., Antia, R., Whitmire, J. K. & Ahmed, R. Humoral immunity due to long-lived plasma cells. *Immunity* **8**, 363–372 (1998).

69. Halliley, J. L. *et al.* Long-Lived Plasma Cells Are Contained within the CD19−CD38hiCD138+ Subset in Human Bone Marrow. *Immunity* **43**, 132–145 (2015).

70. Schroeder, H. W. & Cavacini, L. Structure and function of immunoglobulins. *Journal of Allergy and Clinical Immunology* **125**, S41–S52 (2010).

71. Vidarsson, G., Dekkers, G. & Rispens, T. IgG subclasses and allotypes: from structure to effector functions. *Front. Immunol.* **5**, (2014).

72. Steffen, U. *et al.* IgA subclasses have different effector functions associated with distinct glycosylation profiles. *Nature Communications* **11**, 120 (2020).

73. Keyt, B. A., Baliga, R., Sinclair, A. M., Carroll, S. F. & Peterson, M. S. Structure, Function, and Therapeutic Use of IgM Antibodies. *Antibodies* **9**, 53 (2020).

74. Wang, Y., Su, C., Ji, C. & Xiao, J. CD5L associates with IgM via the J chain. *Nature Communications* **15**, 8397 (2024).

75. de Sousa-Pereira, P. & Woof, J. M. IgA: Structure, Function, and Developability. *Antibodies* **8**, 57 (2019).

76. Stavnezer, J., Guikema, J. E. J. & Schrader, C. E. Mechanism and regulation of class switch recombination. *Annu. Rev. Immunol.* **26**, 261–292 (2008).

77. MacKay, F. & Schneider, P. Cracking the BAFF code. *Nat. Rev. Immunol.* **9**, 491–502 (2009).

78. Elsner, R. A. & Shlomchik, M. J. Germinal Center and Extrafollicular B Cell Responses in Vaccination, Immunity, and Autoimmunity. *Immunity* **53**, 1136–1150 (2020).

79. Nimmerjahn, F. & Ravetch, J. V. Fcγ receptors as regulators of immune responses. *Nature Reviews Immunology* **8**, 34–47 (2008).

80. Roopenian, D. C. & Akilesh, S. FcRn: the neonatal Fc receptor comes of age. *Nature Reviews Immunology* **7**, 715–725 (2007).

81. Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* **1**, 59 (2021).

82. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* **32**, 381–385 (2008).

83. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics* **11**, 499–511 (2010).

84. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics* **47**, 435–444 (2015).

85. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).

86. Cordell, H. J. & Clayton, D. G. A Unified Stepwise Regression Procedure for Evaluating the Relative Effects of Polymorphisms within a Gene Using Case/Control or Family Data: Application to HLA in Type 1 Diabetes. *Am. J. Hum. Genet.* **70**, 124 (2001).

87. Maller, J. B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genetics* **44**, 1294–1301 (2012).

88. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the "Sum of Single Effects" model. *PLoS Genet.* **18**, e1010299 (2022).

89. Kuchenbaecker, K. & Navoly, G. Ancestral diversity in complex disease genetics: from discovery to translation. *Nature Reviews Genetics* **27**, 405–419 (2026).

90. Wu, Y. *et al.* Genome-wide fine-mapping improves identification of causal variants. *Nature Genetics* **58**, 940–951 (2026).

91. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755 (2002).

92. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16**, 197–212 (2015).

93. Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

94. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics* **53**, 1300–1310 (2021).

95. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414.e24 (2016).

96. Kim-Hellmuth, S. *et al.* Cell type–specific genetic regulation of gene expression across human tissues. *Science* **369**, (2020).

97. Ota, M. *et al.* Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. *Cell* **184**, 3006-3021.e17 (2021).

98. Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nature Genetics* **53**, 1290–1299 (2021).

99. Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**, (2022).

100. Sarkar, A. & Stephens, M. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nature Genetics* **53**, 770–777 (2021).

101. Plagnol, V., Smyth, D. J., Todd, J. A. & Clayton, D. G. Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. *Biostatistics* **10**, 327–334 (2009).

102. Nica, A. C. *et al.* Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations. *PLoS Genet.* **6**, e1000895 (2010).

103. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, e1004383 (2014).

104. Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genet.* **17**, e1009440 (2021).

105. Assarsson, E. *et al.* Homogenous 96-Plex PEA Immunoassay Exhibiting High Sensitivity, Specificity, and Excellent Scalability. *PLoS One* **9**, e95192 (2014).

106. Suhre, K., McCarthy, M. I. & Schwenk, J. M. Genetics meets proteomics: perspectives for large population-based studies. *Nature Reviews Genetics* **22**, 19–37 (2020).

107. Sun, B. B. *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).

108. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* **37**, 658–665 (2013).

109. Sanderson, E. *et al.* Mendelian randomization. *Nature Reviews Methods Primers* **2**, 6 (2022).

110. Olafsdottir, T. A. *et al.* Sequence variants influencing the regulation of serum IgG subclass levels. *Nature Communications* **15**, 1–13 (2024).

111. Weintraub, H. & Groudine, M. Chromosomal subunits in active genes have an altered conformation. *Science* **193**, 848–856 (1976).

112. Wu, C. The 5′ ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. *Nature* **286**, 854–860 (1980).

113. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics* **20**, 207–220 (2019).

114. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41–45 (2000).

115. Henikoff, S. & Shilatifard, A. Histone modification: Cause or cog? *Trends in Genetics* **27**, 389–396 (2011).

116. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).

117. Boyle, A. P. *et al.* High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* **132**, 311–322 (2008).

118. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10**, 1213–1218 (2013).

119. Calderon, D. *et al.* Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nature Genetics* **51**, 1494–1505 (2019).

120. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

121. King, H. W. *et al.* Integrated single-cell transcriptomics and epigenomics reveals strong germinal center-associated etiology of autoimmune risk loci. *Sci. Immunol.* **6**, (2021).

122. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47**, 1228–1235 (2015).

123. Vorontsov, I. E., Kulakovskiy, I. V., Khimulya, G., Nikolaeva, D. D. & Makeev, V. J. PERFECTOS-APE: predicting regulatory functional effect of SNPs by approximate P-value estimation. in *Proceedings of the 6th International Conference on Bioinformatics Models, Methods and Algorithms* 102–108 (2015) doi:10.5220/0005189301020108.

124. Mathelier, A. & Wasserman, W. W. The Next Generation of Transcription Factor Binding Site Prediction. *PLoS Comput. Biol.* **9**, e1003214 (2013).

125. Steinhaus, R., Robinson, P. N. & Seelow, D. FABIAN-variant: predicting the effects of DNA variants on transcription factor binding. *Nucleic Acids Res.* **50**, W322–W329 (2022).

126. Hammal, F., De Langen, P., Bergon, A., Lopez, F. & Ballester, B. ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res.* **50**, D316–D325 (2022).

127. Trizzino, M. *et al.* EGR1 is a gatekeeper of inflammatory enhancers in human macrophages. *Sci. Adv.* **7**, (2021).

128. Mansour, M. R. *et al.* An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373–1377 (2014).

129. Pampari, A. *et al.* ChromBPNet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants. *bioRxiv* 2024.12.25.630221 (2024) doi:10.1101/2024.12.25.630221.

130. Castro-Mondragon, J. A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **50**, D165–D173 (2022).

131. Smale, S. T. Luciferase assay. *Cold Spring Harb. Protoc.* **5**, (2010).

132. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology* **30**, 271–277 (2012).

133. Tewhey, R. *et al.* Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519–1529 (2016).

134. Elbashir, S. M. *et al.* Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411**, 494–498 (2001).

135. Cong, L. *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* **339**, 819–823 (2013).

136. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).

137. Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442 (2013).

138. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).

139. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).

140. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* **16**, (2015).

141. Dunham, I. *et al.* An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* **489**, 57 (2012).

142. Gioia, L., Siddique, A., Head, S. R., Salomon, D. R. & Su, A. I. A genome-wide survey of mutations in the Jurkat cell line. *BMC Genomics* **19**, (2018).

143. Zhou, B. *et al.* Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. *Genome Res.* **29**, 472–484 (2019).

144. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics* **47**, 598–606 (2015).

145. Smith, M. L., Cavenagh, J. D., Lister, T. A. & Fitzgibbon, J. Mutation of CEBPA in familial acute myeloid leukemia. *N. Engl. J. Med.* **351**, 2403–2407 (2004).

146. Hernandez, P. A. *et al.* Mutations in the chemokine receptor gene CXCR4 are associated with WHIM syndrome, a combined immunodeficiency disease. *Nature Genetics* **34**, 70–74 (2003).

147. Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N. Engl. J. Med.* **374**, 2209–2221 (2016).

148. Hunter, Z. R. *et al.* The genomic landscape of Waldenstrom macroglobulinemia is characterized by highly recurring MYD88 and WHIM-like CXCR4 mutations, and small somatic deletions associated with B-cell lymphomagenesis. *Blood* **123**, 1637–1646 (2014).

149. Greig, K. T., Carotta, S. & Nutt, S. L. Critical roles for c-Myb in hematopoietic progenitor cells. *Semin. Immunol.* **20**, 247–256 (2008).

150. Leek, J. T. & Storey, J. D. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genet.* **3**, e161 (2007).

151. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies. *PLoS Comput. Biol.* **6**, e1000770 (2010).

152. Zhou, H. J., Li, L., Li, Y., Li, W. & Li, J. J. PCA outperforms popular hidden variable inference methods for molecular QTL mapping. *Genome Biology* **23**, 210 (2022).

153. Vilhais-Neto, G. C. *et al.* Rere controls retinoic acid signaling and somite bilateral symmetry. *Nature* **463**, 953–957 (2010).

154. Tsai, F. Y. *et al.* An early haematopoietic defect in mice lacking the transcription factor GATA-2. *Nature* **371**, 221–226 (1994).

155. Hsu, A. P. *et al.* Mutations in GATA2 are associated with the autosomal dominant and sporadic monocytopenia and mycobacterial infection (MonoMAC) syndrome. *Blood* **118**, 2653–2655 (2011).

156. Spinner, M. A. *et al.* GATA2 deficiency: a protean disorder of hematopoiesis, lymphatics, and immunity. *Blood* **123**, 809–821 (2014).

157. Lim, K. C. *et al.* Conditional Gata2 inactivation results in HSC loss and lymphatic mispatterning. *J. Clin. Invest.* **122**, 3705–3717 (2012).

158. Shearstone, J. R. *et al.* Chemical Inhibition of Histone Deacetylases 1 and 2 Induces Fetal Hemoglobin through Activation of GATA2. *PLoS One* **11**, e0153767 (2016).

159. Choudhuri, A. *et al.* Common variants in signaling transcription-factor-binding sites drive phenotypic variability in red blood cell traits. *Nature Genetics* **52**, 1333–1345 (2020).

160. Bresnick, E. H., Lee, H. Y., Fujiwara, T., Johnson, K. D. & Keles, S. GATA Switches as Developmental Drivers. *J. Biol. Chem.* **285**, 31087 (2010).

161. Miyazono, K., Kamiya, Y. & Morikawa, M. Bone morphogenetic protein receptors and signal transduction. *The Journal of Biochemistry* **147**, 35–51 (2010).

162. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).

163. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology* **31**, 1102–1111 (2013).

164. Subramanian, A. & Miller, D. M. Structural Analysis of α-Enolase. *Journal of Biological Chemistry* **275**, 5958–5965 (2000).

165. Feo, S., Arcuri, D., Piddini, E., Passantino, R. & Giallongo, A. ENO1 gene product binds to the c-myc promoter and acts as a transcriptional repressor: relationship with Myc promoter-binding protein 1 (MBP-1). *FEBS Lett.* **473**, 47–52 (2000).

166. Cabezas-Wallscheid, N. *et al.* Vitamin A-Retinoic Acid Signaling Regulates Hematopoietic Stem Cell Dormancy. *Cell* **169**, 807-823.e19 (2017).

167. Lopez de Lapuente Portilla, A. *et al.* BloodVariome: a high-resolution atlas of inherited genetic effects in human immune cells. *bioRxiv* 2026.03.30.715213 (2026) doi:10.64898/2026.03.30.715213.

168. Jonsson, S. *et al.* Identification of sequence variants influencing immunoglobulin levels. *Nature Genetics* **49**, 1182–1191 (2017).

169. Blake, J. A. *et al.* Mouse Genome Database (MGD): Knowledgebase for mouse-human comparative biology. *Nucleic Acids Res.* **49**, D981–D987 (2021).

170. Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* **50**, D687–D692 (2022).

171. Karlsson, M. *et al.* A single–cell type transcriptomics map of human tissues. *Sci. Adv.* **7**, (2021).

172. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nature Genetics* **47**, 1236–1241 (2015).

173. Mohanan, V. *et al.* C1orf106 is a colitis risk gene that regulates stability of epithelial adherens junctions. *Science* **359**, 1161–1166 (2018).

174. Fütterer, A., Mink, K., Luz, A., Kosco-Vilbois, M. H. & Pfeffer, K. The lymphotoxin β receptor controls organogenesis and affinity maturation in peripheral lymphoid tissues. *Immunity* **9**, 59–70 (1998).

175. Lauc, G. *et al.* Loci associated with N-glycosylation of human immunoglobulin G show pleiotropy with autoimmune diseases and haematological cancers. *PLoS Genet.* **9**, (2013).

176. Takai, T. Roles of Fc receptors in autoimmunity. *Nature Reviews Immunology* **2**, 580–592 (2002).

177. Cobaleda, C., Schebesta, A., Delogu, A. & Busslinger, M. Pax5: the guardian of B cell identity and function. *Nature Immunology* **8**, 463–470 (2007).

178. Minderjahn, J. *et al.* Mechanisms governing the pioneering and redistribution capabilities of the non-classical pioneer PU.1. *Nat. Commun.* **11**, (2020).

179. Wei, F., Zaprazna, K., Wang, J. & Atchison, M. L. PU.1 Can Recruit BCL6 to DNA To Repress Gene Expression in Germinal Center B Cells. *Mol. Cell. Biol.* **29**, 4612 (2009).

180. Wang, H. *et al.* Transcription factors IRF8 and PU.1 are required for follicular B cell development and BCL6-driven germinal center responses. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 9511–9520 (2019).

181. Basso, K. & Dalla-Favera, R. BCL6. master regulator of the germinal center reaction and key oncogene in B Cell lymphomagenesis. *Adv. Immunol.* **105**, 193–210 (2010).

182. Reddy, A. *et al.* Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. *Cell* **171**, 481-494.e15 (2017).

183. Krysiak, K. *et al.* Recurrent somatic mutations affecting B-cell receptor signaling pathway genes in follicular lymphoma. *Blood* **129**, 473–483 (2017).

184. Shin, O. H., Hau, W., Wang, Y. & Südhof, T. C. Evolutionarily conserved multiple C2 domain proteins with two transmembrane regions (MCTPs) and

unusual Ca2+ binding properties. *Journal of Biological Chemistry* **280**, 1641–1651 (2005).

185. Cocco, M. *et al.* A dichotomy of gene regulatory associations during the activated B-cell to plasmablast transition. *Life Sci. Alliance* **3**, (2020).

186. Sellars, M., Kastner, P. & Chan, S. Ikaros in B cell development and function. *World J. Biol. Chem.* **2**, 132–9 (2011).

187. Lu, X. *et al.* MTA2/NuRD Regulates B Cell Development and Cooperates with OCA-B in Controlling the Pre-B to Immature B Cell Transition. *Cell Rep.* **28**, 472-485.e5 (2019).

188. Dankers, W., Colin, E. M., van Hamburg, J. P. & Lubberts, E. Vitamin D in autoimmunity: Molecular mechanisms and therapeutic potential. *Front. Immunol.* **7**, 236441 (2017).

189. Chen, S. *et al.* Modulatory effects of 1,25-dihydroxyvitamin D3 on human B cell differentiation. *J. Immunol.* **179**, 1634–1647 (2007).

190. Plum, L. A. *et al.* Antibody production in mice requires neither vitamin D, nor the vitamin D receptor. *Front. Immunol.* **13**, 960405 (2022).

191. Landrum, M. J. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Res.* **48**, D835–D844 (2020).

192. McMaster, M. L. *et al.* Two high-risk susceptibility loci at 6p25.3 and 14q32.13 for Waldenström macroglobulinemia. *Nature Communications* **9**, 4182 (2018).

193. Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nature Genetics* **53**, 1415–1424 (2021).

194. Güler, M. & Canzian, F. Clustering of lymphoid neoplasms by cell of origin, somatic mutation and drug usage profiles: a multi-trait genome-wide association study. *Blood Cancer Journal* **15**, 147 (2025).

195. Owen, R. G. *et al.* Clinicopathological definition of Waldenstrom's macroglobulinemia: Consensus Panel Recommendations from the Second International Workshop on Waldenstrom's Macroglobulinemia. *Semin. Oncol.* **30**, 110–115 (2003).

196. Salzer, U. *et al.* Mutations in TNFRSF13B encoding TACI are associated with common variable immunodeficiency in humans. *Nature Genetics* **37**, 820–828 (2005).

197. Castigli, E. *et al.* TACI is mutant in common variable immunodeficiency and IgA deficiency. *Nature Genetics* **37**, 829–834 (2005).

198. Floto, R. A. *et al.* Loss of function of a lupus-associated FcγRIIb polymorphism through exclusion from lipid rafts. *Nature Medicine* **11**, 1056–1058 (2005).

199. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).

200. Do, R., Kathiresan, S. & Abecasis, G. R. Exome sequencing and complex disease: practical aspects of rare variant association studies. *Hum. Mol. Genet.* **21**, R1–R9 (2012).

201. Shimazu, R. *et al.* MD-2, a Molecule that Confers Lipopolysaccharide Responsiveness on Toll-like Receptor 4. *J. Exp. Med.* **189**, 1777 (1999).

202. Went, M. *et al.* Deciphering the genetics and mechanisms of predisposition to multiple myeloma. *Nature Communications* **15**, 6644 (2024).

203. Watson, C. T. & Breden, F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. *Genes & Immunity* **13**, 363–373 (2012).

204. Lefranc, M. P., Lefranc, G. & Rabbitts, T. H. Inherited deletion of immunoglobulin heavy chain constant region genes in normal human individuals. *Nature* **300**, 760–762 (1982).

205. Cogné, M. *et al.* A class switch control region at the 3′ end of the immunoglobulin heavy chain locus. *Cell* **77**, 737–747 (1994).

206. Pinaud, E. *et al.* The IgH Locus 3′ Regulatory Region: Pulling the Strings from Behind. *Adv. Immunol.* **110**, 27–70 (2011).

207. Zhao, Y., Pan-Hammarström, Q., Zhao, Z., Wen, S. & Hammarström, L. Selective IgG2 deficiency due to a point mutation causing abnormal splicing of the Cγ2 gene. *Int. Immunol.* **17**, 95–101 (2005).

208. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, (2023).

209. Luppino, F., Lenz, S., Chow, C. F. W. & Toth-Petroczy, A. Deep learning tools predict variants in disordered regions with lower sensitivity. *BMC Genomics* **26**, 367 (2025).

210. George, R. A. & Heringa, J. An analysis of protein domain linkers: their classification and role in protein folding. *Protein Eng.* **15**, 871–879 (2002).

211. Oganesyan, V. *et al.* Structural insights into neonatal Fc receptor-based recycling mechanisms. *Journal of Biological Chemistry* **289**, 7812–7824 (2014).

212. Pyzik, M., Kozicky, L. K., Gandhi, A. K. & Blumberg, R. S. The therapeutic age of the neonatal Fc receptor. *Nat. Rev. Immunol.* **23**, 415–432 (2023).

213. Swaminathan, B. *et al.* Variants in ELL2 influencing immunoglobulin levels associate with multiple myeloma. *Nature Communications* **6**, 7213 (2015).

214. Weldon, A. J. *et al.* Surface APRIL Is Elevated on Myeloid Cells and Is Associated with Disease Activity in Patients with Rheumatoid Arthritis. *J. Rheumatol.* **42**, 749–759 (2015).

215. Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nature Genetics* **50**, 746–753 (2018).

216. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

217. Gibson, G. Rare and common variants: twenty arguments. *Nature Reviews Genetics* **13**, 135–145 (2012).

218. Gathmann, B. *et al.* Clinical picture and treatment of 2212 patients with common variable immunodeficiency. *Journal of Allergy and Clinical Immunology* **134**, 116-126.e11 (2014).

219. Dobbs, K. *et al.* Inherited DOCK2 Deficiency in Patients with Early-Onset Invasive Infections. *N. Engl. J. Med.* **372**, 2409–2422 (2015).

220. La, Y. P. *et al.* Genetic variation at PPM1H predicts mobilization efficiency in stem cell donors. *Blood Adv.* **9**, 4683–4686 (2025).

221. Prietl, B., Treiber, G., Pieber, T. R. & Amrein, K. Vitamin D and Immune Function. *Nutrients* **5**, 2502–2521 (2013).

222. Rolf, L., Muris, A. H., Hupperts, R. & Damoiseaux, J. Illuminating vitamin D effects on B cells – the multiple sclerosis perspective. *Immunology* **147**, 275 (2016).

223. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

224. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).

225. Foy, B. H. *et al.* Haematological setpoints are a stable and patient-specific deep phenotype. *Nature* **637**, 430–438 (2024).

226. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* **6**, (2017).

227. Schmidt, R. *et al.* Base-editing mutagenesis maps alleles to tune human T cell functions. *Nature* **625**, 805–812 (2023).

228. Dai, P. *et al.* Protocol for genetic dissection of class switch recombination using genome-editing tools. *STAR Protoc.* **6**, 103882 (2025).

229. Feng, Y. *et al.* FAM72A antagonizes UNG2 to promote mutagenic repair during antibody maturation. *Nature* **600**, 324–328 (2021).

230. Rogier, M. *et al.* Fam72a enforces error-prone DNA repair during antibody diversification. *Nature* **600**, 329–333 (2021).

Acknowledgements

To Björn Nilsson, my supervisor, for taking me into the group and giving me problems worth several years of work. The freedom to follow a result where it led, and the expectation that I would justify where it led me, thank you for helping me rediscover my passion and self-confidence as a scientist.

To Aitzkoa Lopez de Lapuente Portilla, my co-supervisor, who designed BloodVariome from scratch. Everything in this thesis rests on that framework. Thank you for all the patience, support and kindness you gave me these years.

To Caterina Cafaro, for long evenings troubleshooting wet lab experiments, for good advice, for your humour and kindness. To Ludvig Ekdahl and Antton Lamarca, for the countless times I brought a bioinformatics problem to your desk and left with it solved, or at least correctly diagnosed. Much of what I know about doing this work carefully I learned from watching how the two of you do it. To Ram Ajore, who trained me to use the cell lab and whose mastery of molecular biology saved me on multiple occasions. To Maroulitsa, to Laura, to Yuwen, to Yen, to Tereza, to everyone in the Nilsson lab – thank you for creating a wonderful environment where I could rediscover my passion and creativity, and for so, so many wonderful lunches.

To everyone at deCODE, for hosting me in Iceland and for making it possible for us to pursue projects of this scale. To Gudmar Thorleifsson, who has my greatest thanks. The subclass analysis was your idea, and it became the result I am most pleased with in the immunoglobulin paper. To Gísli H. Halldórsson, for helping with the colocalization procedure and for the eQTL analysis of deCODE's internal data. To Þórunn Ólafsdóttir and Ingileif Jónsdóttir, whose immunological expertise repeatedly kept our interpretation of the genetics anchored to how the immune system actually behaves.

To Kristijonas Žemaitis, who performed the stem cell experiments in the CD34+ study, and to Jonas Larsson for hosting that work. To Parashar Dhapola, for the single-cell data and help with analysis.

To Mineto Ota, for the generosity with which he set up our collaboration, and to Hamish King, who helped by email with the processing of the single-cell tonsil data and asked nothing in return.

This work was funded by the CanFaster programme at Lund University, which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 754299.

Thank you to my family for supporting me through this process, and to my perfect daughters, Amani and Arani, everything I do is for you.

Paper I

An update to this article is included at the end

blood® Regular Article

HEMATOPOIESIS AND STEM CELLS

Genome-wide association study on 13 167 individuals identifies regulators of blood CD34$^+$ cell levels

Aitzkoa Lopez de Lapuente Portilla,[1,2] Ludvig Ekdahl,[1,2,*] Caterina Cafaro,[1,2,*] Zain Ali,[1,2,*] Natsumi Miharada,[1,2] Gudmar Thorleifsson,[3] Kristijonas Žemaitis,[1,2] Antton Lamarca Arrizabalaga,[1,2] Malte Thodberg,[1,2] Maroulio Pertesi,[1,2] Parashar Dhapola,[1,2] Erik Bao,[4-6] Abhishek Niroula,[1,2,6] Divya Bali,[1,2] Gudmundur Norddahl,[3] Nerea Ugidos Damboriena,[1,2] Vijay G. Sankaran,[4-6] Göran Karlsson,[1,2] Unnur Thorsteinsdottir,[3] Jonas Larsson,[1,2] Kari Stefansson,[3] and Björn Nilsson[1,2,6]

[1]Lund Stem Cell Center; [2]Department of Laboratory Medicine, Lund University, Lund, Sweden; [3]deCODE genetics/Amgen Inc., Reykjavik, Iceland; [4]Division of Hematology and Oncology, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA; [5]Harvard Stem Cell Institute, Cambridge, MA; and [6]Broad Institute, Cambridge, MA

KEY POINTS

- **We report a large-scale genome-wide association study on blood CD34$^+$ cell levels, identifying 9 significant and 2 suggestive variants.**
- **The most significant associations map to *CXCR4* and *PPM1H*, the latter a potential inhibition target for stem cell mobilization.**

Stem cell transplantation is a cornerstone in the treatment of blood malignancies. The most common method to harvest stem cells for transplantation is by leukapheresis, requiring mobilization of CD34$^+$ hematopoietic stem and progenitor cells (HSPCs) from the bone marrow into the blood. Identifying the genetic factors that control blood CD34$^+$ cell levels could reveal new drug targets for HSPC mobilization. Here we report the first large-scale, genome-wide association study on blood CD34$^+$ cell levels. Across 13 167 individuals, we identify 9 significant and 2 suggestive associations, accounted for by 8 loci (*PPM1H, CXCR4, ENO1-RERE, ITGA9, ARHGAP45, CEBPA, TERT*, and *MYC*). Notably, 4 of the identified associations map to *CXCR4*, showing that bona fide regulators of blood CD34$^+$ cell levels can be identified through genetic variation. Further, the most significant association maps to *PPM1H*, encoding a serine/threonine phosphatase never previously implicated in HSPC biology. *PPM1H* is expressed in HSPCs, and the allele that confers higher blood CD34$^+$ cell levels downregulates *PPM1H*. Through functional fine-mapping, we find that this downregulation is caused by the variant rs772557-A, which abrogates an MYB transcription factor–binding site in *PPM1H* intron 1 that is active in specific HSPC subpopulations, including hematopoietic stem cells, and interacts with the promoter by chromatin looping. Furthermore, *PPM1H* knockdown increases the proportion of CD34$^+$ and CD34$^+$90$^+$ cells in cord blood assays. Our results provide the first large-scale analysis of the genetic architecture of blood CD34$^+$ cell levels and warrant further investigation of PPM1H as a potential inhibition target for stem cell mobilization.

Introduction

Human hematopoiesis is defined by the outgrowth of mature blood cells through successive proliferation and differentiation of hematopoietic stem and progenitor cells (HSPCs).[1] In blood, HSPCs constitute a tiny (~0.1%) subset of mononuclear white blood cells. Circulating HSPCs express the surface protein CD34, allowing quantification by flow cytometry.[2] Epidemiologic studies[3,4] indicate that blood CD34$^+$ cell levels are partly genetically determined. However, the underlying DNA sequence variants and genes remain unknown, and no genome-wide significant associations have been reported.[3]

To identify DNA sequence variants that influence blood CD34$^+$ cell levels, we conducted a genome-wide association study (GWAS). We identified 9 significant and 2 suggestive association signals, including 4 independent associations with *CXCR4* and an association with *PPM1H*. Notably, *PPM1H* encodes a serine/threonine phosphatase previously not implicated in HSPC biology. Relative to other blood cell types, *PPM1H* is preferentially expressed in HSPCs. The allele that confers higher blood CD34$^+$ cell levels downregulates *PPM1H* in CD34$^+$ cells due to loss of an MYB transcription factor–binding site. Our results provide new insight into the regulation of blood CD34$^+$ cell levels and warrant further studies of PPM1H as a potential inhibition target to facilitate stem cell mobilization.

Methods

The methods and materials are described in detail in the supplemental Methods (available on the *Blood* Web site).

Sample collection

We collected 16 931 peripheral blood samples from blood donors (n = 7773) and primary care patients (n = 9158) in 3 phases:

November 2015 to April 2016 (Phase I); January 2017 to November 2017 (Phase II); and August 2018 to April 2019 (Phase III) (Lund University Ethical Review Board; dnr 2018/2) (supplemental Table 1; supplemental Figure 1). Information about sex was available for all participants and age for 8251 participants. After quality control, 13 167 unique individuals remained.

Flow cytometry phenotyping

After red blood cell lysis, white blood cells were washed twice, stained with antibody cocktail containing anti–CD45-APC-H7 and either anti–CD34-PerCP-Cy5.5 or anti–CD34-PE-CF594 (supplemental Table 2), and analyzed by using a BD FACS Canto II (Phase I), BD LSR Fortessa (Phase II), or Bio-Rad ZE5 (Phase III) (supplemental Table 3). Samples were processed within 36 hours of blood draw.

Gating of flow cytometry data for association study

To quantify blood $CD34^+$ cell levels, we gated singlet peripheral blood mononuclear cells (PBMCs) based on forward scatter area, forward scatter height, and side scatter area, and then $CD34^+45^{low}$ cells and $CD45^+$ cells within PBMCs. The $CD34^+$ level was defined as the number of $CD34^+45^{low}$ cells divided by the number of $CD45^+$ PBMCs (supplemental Figure 2).

To facilitate analysis of the large volume of flow cytometry data, we developed pattern recognition software (AliGater; https://github.com/LudvigEk/AliGater; gating strategies at https://github.com/LudvigEk/HSPC-regulators-in-human-blood) (supplemental Methods; supplemental Figure 3). Phase I and II samples were gated by using AliGater; Phase III samples were gated manually by using FlowJo (FlowJo LLC). Minor differences in $CD34^+$ levels were observed between phases (median, 9.11×10^{-4} $CD34^+$ per $CD45^+$PBMCs in Phase I; 9.66×10^{-4} in Phase II and 1.0×10^{-3} in Phase III; Kruskal-Wallis test, $P = 4.0 \times 10^{-88}$) (supplemental Table 3), likely because of the use of different flow cytometers and antibody panels.

Genotyping and association analysis

Samples were genotyped by using Illumina OmniExpress-24 and Global Screening single-nucleotide polymorphism (SNP) microarrays. For imputation, we used pre-existing whole-genome sequence data for 17 408 individuals of European origin, including 3704 Swedish individuals. Using ancestry analysis, the association data were partitioned into a discovery set of 10 949 individuals of Swedish ancestry and a follow-up set of 2218 individuals of non-Swedish European ancestry. Association testing was performed by using linear regression in the discovery and follow-up sets separately; the results were combined through meta-analysis using a fixed effects inverse variance method.[5] To adjust for multiple testing, class-based Bonferroni significance thresholds were used.[6] To identify underlying independent signals, we used step-wise conditional analysis. To define the 99% credible set of plausible causal variants for each association signal, we used a Bayesian refinement approach.[7]

Chromatin accessibility data

To identify candidate causal regulatory variants, we used Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) data for sorted blood cell types.[8,9] To screen for genomic regions that drive candidate gene expression, these data were integrated with messenger RNA–sequencing (RNA-seq) data for the same cell types, by testing for correlation across blood cell types between gene expression and ATAC-seq signal intensity (averaged across a 100-bp sliding window positioned at every 50 bp). To map chromatin accessibility in MOLM13 cells, ATAC-seq was performed on 50 000 cells[10] (raw sequencing data in Sequence Read Archive, accession number PRJNA761267). To test for allele-dependent accessibility at rs772557, we used publicly available ENCODE DNase-sequencing data for primary adult $CD34^+$ cells from a single heterozygous individual.[11]

Heritability estimation

Total narrow-sense heritability was estimated with the linkage disequilibrium score regression (LDSC) software.[12,13] To estimate cell type–specific partitioned heritability based on chromatin accessibility, we used LD scores based on ATAC-seq data for sorted blood cell types available for LDSC.[12] This was extended with LD scores for myeloid and plasmacytoid dendritic cells computed by using LDSC from published ATAC-seq data[8] (National Center for Biotechnology Information Gene Expression Omnibus accession number GSE119453).

Promoter capture Hi-C data analysis

To identify variants that interact with gene promoters, we used published promoter capture Hi-C (PCHi-C) data for $CD34^+$ cells (ArrayExpress accession number E-MTAB-2323).[14]

Chromatin immunoprecipitation–sequencing data

To test allele-specific binding of MYB to rs772557, we used MYB chromatin immunoprecipitation with massively parallel DNA sequencing data[15] and whole-genome sequencing data[16] for Jurkat cells (Sequence Read Archive accession numbers SRR1603653 and SRR5349449).

Expression quantitative trait locus analysis in blood $CD34^+$ cells

We purified $CD34^+$ cells from 155 blood donors using immunomagnetic bead enrichment followed by fluorescence-activated cell sorting within 8 hours of blood draw (14 000 to 358 000 cells per sample; average, 122 000). After RNA extraction and complementary DNA synthesis, samples were sequenced on an Illumina NovaSeq 6000 sequencer. To test for associations between variants and gene expression, we used linear modeling with the variant genotype as independent variable and 10 expression principal component covariates. For *CXCR4*, the 3 common variants were included as independent variables in the same model. *P* values were calculated by using Student *t* test for the independent variables. The sequencing data and counts of Fragments Per Kilobase of transcript per Million mapped reads are available from the European Genome Archive repository (accession numbers EGAS00001005655 and EGAD00001008194).

Gene expression data and analysis

To map candidate gene expression, we used single-cell RNA-seq data for 35 582 mononuclear cells from blood and bone marrow,[17] and bulk RNA-seq data for sorted blood cell types.[8] In addition, we used single-cell Cellular Indexing of Transcriptomes and Epitopes by Sequencing data on 4905 $CD34^+$ cells from adult bone marrow.[18] For dimension reduction, we used Uniform Manifold Approximation and Projection[19] and imputed gene expression using MAGIC were used.[20]

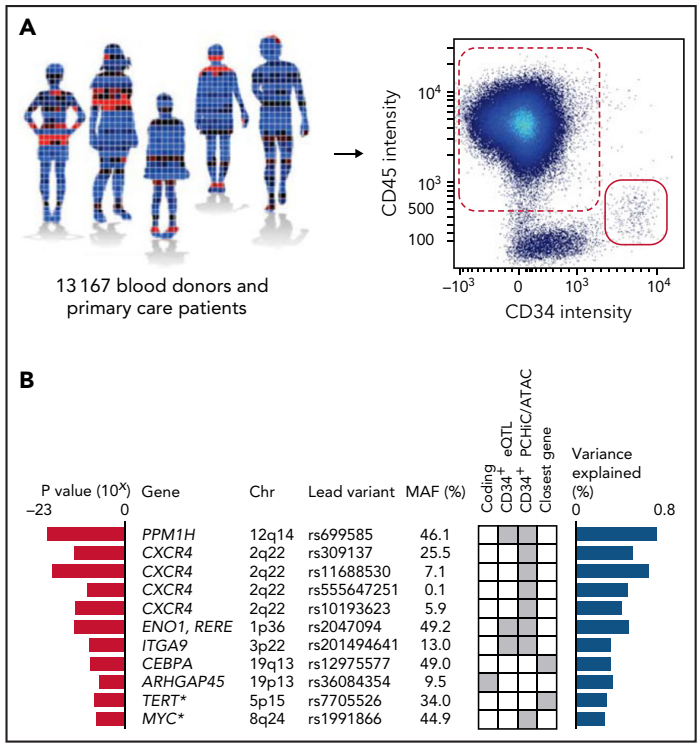


Figure 1. GWAS. (A) A GWAS was performed on blood CD34$^+$ cell levels in 13 167 individuals, including 10 949 of Swedish ancestry and 2218 of non-Swedish European ancestry. The CD34$^+$ level was defined as the number of CD34$^+$ cells (red solid) divided by the number of CD45$^+$ mononuclear cells (red dashed) (supplemental Figure 2). (B) In a combined analysis of the individuals of Swedish ancestry and the individuals of non-Swedish European ancestry, 9 significant and 2 suggestive (asterisks) associations (having *P* values within one order of magnitude from Bonferroni thresholds) were identified (supplemental Tables 5-8). The listed variants are the most significant (lead) variants for each association. Red bars indicate *P* values for association with blood CD34$^+$ cell levels for the lead variants in the combined analysis. Blue bars indicate proportion of variance explained. Gene names indicate the likely candidate genes of each association. Genes were prioritized as candidate genes if they: (1) had a coding variant within the 99% credible set of probable causal variants (supplemental Table 9); (2) had a *cis*-eQTL in CD34$^+$ cells (supplemental Table 10); or (3) the credible set contained a regulatory variant that maps either to the promoter or to a region with a chromatin looping interaction with the promoter in CD34$^+$ cells. If none of these criteria were fulfilled, the closest gene was prioritized. The criterion used to call each gene a candidate gene is indicated by gray squares in the matrix. MAF, minor allele frequency.

CRISPR/Cas9 perturbation of variant regions

We used dual-single guide RNA (sgRNA) CRISPR/Cas9 to delete the regions harboring the 4 candidate causal variants at *CXCR4*, and an allele-specific single-sgRNA CRISPR/Cas9 approach to perturb rs772557 at *PPM1H* (supplemental Table 4). sgRNAs were cloned into the pSpCas9(BB)-2A-GFP PX458 vector and transfected into MOLM13 cells (*CXCR4* variants) or K562 cells (*PPM1H* rs772557 variant). At 24 hours posttransfection, green fluorescent protein–positive cells were isolated by fluorescence-activated cell sorting. Deletion in MOLM13 cells was verified by polymerase chain reaction and electrophoresis in a 1% agarose gel. Deletion efficiency was estimated as the intensity of the deletion band divided by the sum of the intensity of the deletion band and the intensity of the wild-type band. For allele-specific CRISPR/Cas9 toward rs772557 in K562 cells, perturbation was verified by using Sanger sequencing and the Tracking of Indels by Decomposition (TIDE, https://tide.nki.nl/). As control, we used sgRNAs targeting a random noncoding region on a different chromosome. *CXCR4* and *PPM1H* expression after CRISPR/Cas9 perturbation was quantified by using quantitative polymerase chain reaction.

Luciferase experiments

For the *PPM1H* variants, 200-bp sequences representing the reference and alternative allele of rs772555, rs772556, rs772557, and rs772559 in their genomic contexts were synthesized as gBlocks and cloned into the pGL3-Basic plasmid using *Kpn*I and *Bgl*II restriction sites.

Sequences were centered on the variant, and the 2 constructs differed only for the variant. Renilla luciferase constructs were cotransfected with firefly construct into K562 cells. At 24 hours postelectroporation, luciferase and Renilla activities were measured. Based on the readings, $\log_2$ scores were calculated for each variant reflecting the luciferase activity of the alternative allele relative to the reference allele. For coelectroporation experiments with *MYB* small interfering RNA (siRNA), luciferase plasmids were cotransfected with Qiagen FlexiTube siRNA solution

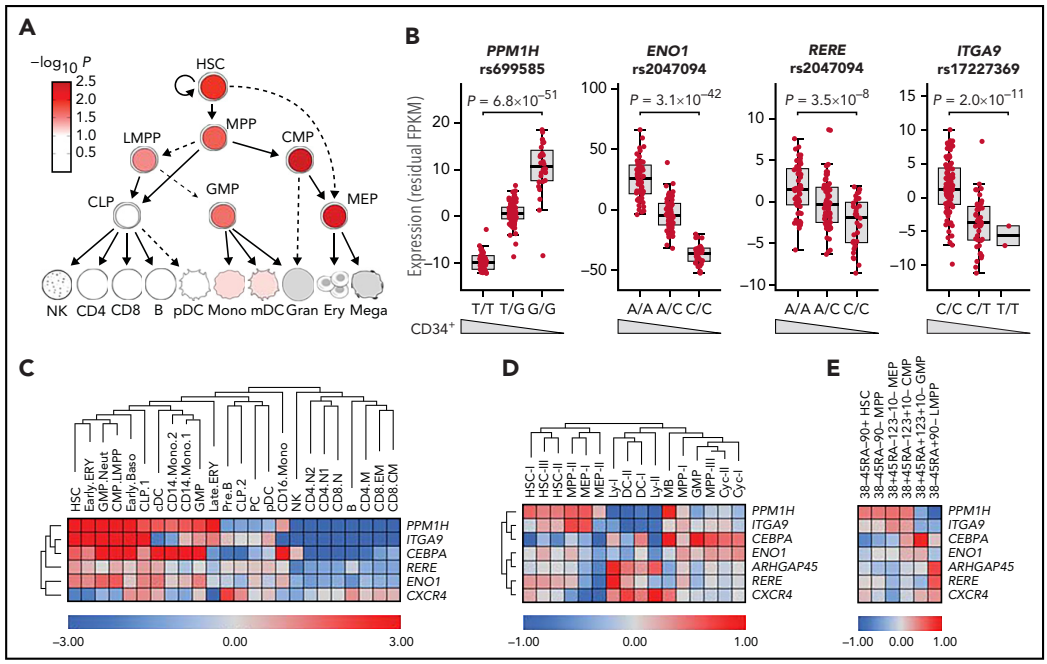


Figure 2. Effects on gene expression in HSPCs. (A) LD score regression shows enrichments of heritability in regions with accessible chromatin in HSPC subpopulations (red nuance indicates $-\log_{10} P$ value). (B) *cis*-eQTLs at *PPM1H*, *ENO1*, *RERE*, and *ITGA9* identified RNA-seq data for CD34$^+$ cells from 155 blood donors (supplemental Table 10). Data are residual fragments per kilobase of transcript per million mapped reads (FPKM) values after correction for 10 expression principal components. Wedges indicate directions of effects on blood CD34$^+$ cell levels for the same variant. (C) Candidate gene expression in single-cell messenger RNA-seq data from 35 582 blood and bone marrow mononuclear cells[17] (supplemental Figures 9 and 10). *ARHGAP45* was not represented in this data set. (D-E) Average expression of candidate genes across different cell clusters in single-cell Cellular Indexing of Transcriptomes and Epitopes by Sequencing data for 4905 lineage-negative CD34$^+$ cells from adult bone marrow.[18] In panel D, the 4905 cells have been clustered according to RNA-seq pattern (supplemental Figure 11). In panel E, the cells were instead clustered by gating using the sequence counts for the tags derived from antibodies to the CD38, CD45RA, CD90, CD123, and CD10 cell surface markers, as indicated at the upper edge of the heatmap (supplemental Figure 12). B, B cells; Baso, basophil; CD4, CD4$^+$ T cells; CD8, CD8$^+$ T cells; CLP, common lymphoid progenitors; Cyc, cycling cells; DC, dendritic cells; ERY, erythroid progenitors; GMP, granulocyte-monocyte progenitors; LMPP, lymphoid-primed multipotent progenitors; Mono, monocyte; PC, plasma cells; PreB, pre–B cells; Neut, neutrophil; NK, natural killer cells.

targeting *MYB* or Qiagen Negative Control siRNA. Lysates for luciferase activity measurement and Western blot were collected simultaneously, 24 hours after electroporation. Knockdown was confirmed by western blot using a recombinant c-Myb antibody.

Motif analysis

For motif analysis, we used PERFECTOS-APE (http://opera.autosome.ru/perfectosape) with the HOCOMOCO-10, JASPAR, HT-SELEX, SwissRegulon, and HOMER motif databases.

shRNA-knockdown in umbilical cord blood cells

Umbilical cord blood samples were obtained from newborns at Skåne University Hospital (Lund and Malmö, Sweden) and Helsingborg Hospital (Helsingborg, Sweden), in compliance with regulations set by the regional ethical committee and informed consent. Mononuclear cells were isolated by density gradient centrifugation within 48 hours of sample collection and kept at −80°C. Thawed cord blood–derived CD34$^+$ cells were sorted, lentivirally transduced with *PPM1H*-targeting or nontargeting (control) shRNA, and maintained in serum-free expansion medium supplemented with stem cell factor, thrombopoietin, and FMS-like tyrosine kinase 3 ligand. Using flow cytometry, the percentages and absolute counts (using CountBright beads; Thermo Fisher Scientific) of CD34$^+$ and CD34$^+$90$^+$ cells at days 7, 14, and 21 were measured. Cell enrichment was calculated by dividing CD34$^+$ and CD34$^+$90$^+$ counts at each time point by initial cell counts, and then normalizing the enrichment values to the control shRNA-transduced cells at the same time point. At each time point, 3 transduction replicates were recorded. The experiment was repeated 3 times.

For statistical analysis, we compared normalized enrichment values at days 7, 14, and 21 for shRNA-transduced cells vs nontargeting shRNA control. To ensure a conservative analysis, we first averaged the transduction replicates for each experiment, then calculated *P* values using permutation testing with day 7, 14, and 21 data for the same experiment being permuted within each day stratum (100 000 permutations).

Results

Genome-wide association study

Blood CD34$^+$ cell levels were analyzed in 13 167 blood donors and primary care patients from southern Sweden

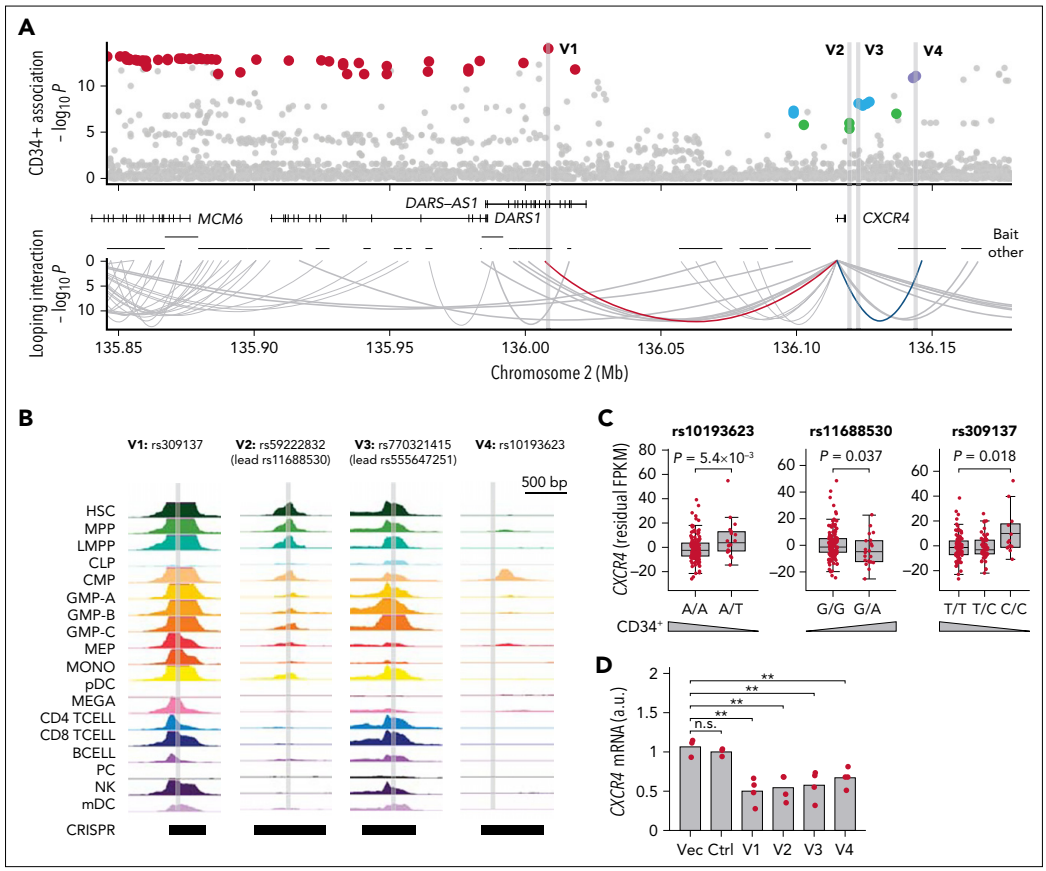


Figure 3. Associations with *CXCR4*. (A) Four independent associations were detected at 2q22. Credible sets (supplemental Table 9) indicated in red (lead variant rs309137), green (rs11688530), cyan (rs555647251), and blue (rs10193623). Using ATAC-seq and PCHi-C data, a single plausible causal variant was identified within each set. rs59222832 ("V2"; in credible set of rs11688530) and rs770321415 ("V3"; in credible set of rs555647251) map to the *CXCR4* promoter region, 4.7 and 1.3 kb upstream of the transcription start site, respectively. In the other 2 credible sets, the lead variants rs309137 ("V1") and rs10193623 ("V4") were identified as likely causal based on looping interactions with the promoter (red and blue arches; y-axis indicates PCHi-C *P*-score). (B) Chromatin accessibility at putative causal variants in 18 blood cell types (y-axis indicates ATAC-seq signal): HSCs, MPPs, lymphoid-primed multipotent progenitors (LMPP), common lymphoid progenitors (CLP), common myeloid progenitors (CMP), granulocyte-monocyte progenitors (GMPs), MEPs, monocyte (MONO), megakaryocyte (MEGA), CD4$^+$ T cells (CD4 TCELL), CD8$^+$ T-cells (CD8 TCELL), B cells (BCELL), natural killer cells (NK), myeloid and plasmacytoid dendritic cells (mDC and pDC, respectively), and plasma cells (PC). (C) Conditional *CXCR4 cis*-eQTLs for the 3 common variants in our CD34$^+$ cell RNA-seq data (n = 155). Wedges indicate directions of effects on blood CD34$^+$ cell levels. Data are residual Fragments Per Kilobase of transcript per Million mapped reads (FPKM) values after correction for the 2 covariate lead variants and 10 principal components. *P* values for multivariable regression, one-sided in the direction opposite the effect on blood CD34$^+$ cell levels of the illustrated variant. (D) *CXCR4* expression in MOLM13 cells subjected to CRISPR/Cas9 editing with empty vector (Vec), a nontargeting sgRNA pair (Ctrl), or sgRNA pairs designed to delete 587 to 1421 bp regions harboring the 4 putative causal variants (3-4 biological replicates per condition) (supplemental Table 4; supplemental Figure 14). a.u., arbitrary units; mRNA, messenger RNA; n.s., not significant; $^{**}P < .01$.

(aged 18-71 years) (Figure 1A; supplemental Table 1; supplemental Figure 1). In each sample, we analyzed up to 1 million white blood cells and defined the CD34$^+$ cell level as the number of CD34$^+$ cells divided by the number of CD45$^+$ mononuclear cells (supplemental Tables 2 and 3; supplemental Figures 2 and 3). To assess reproducibility, 660 individuals were sampled twice, with 3 to 36 months between samplings; significant correlation was found between replicates (Spearman $P = 1.3 \times 10^{-84}$; $r^2 = 0.44$) (supplemental Figure 4). Higher blood CD34$^+$ cell levels were observed in male subjects than in female subjects (supplemental Figure 5), but no association with age was noted.

The participants were genotyped for 18 million variants by using SNP microarrays and imputation. For association analysis, the data were partitioned into a discovery set of 10 949 individuals of Swedish ancestry and a follow-up set of 2218 individuals of other European ancestry (supplemental Table 1; supplemental

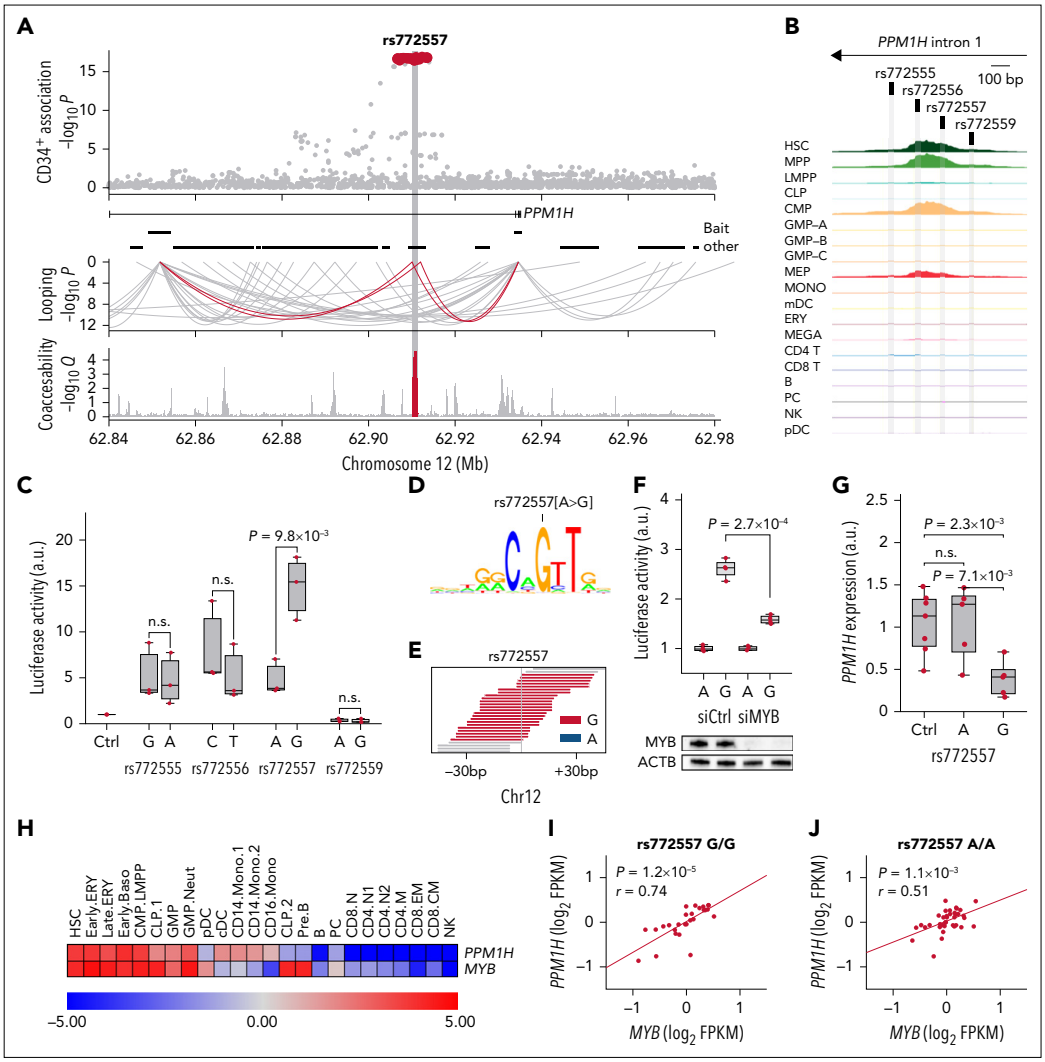


Figure 4. Association with *PPM1H*. (A) Top: close-up of the 12q14 signal. Middle: chromatin looping interactions in CD34$^+$ cells with standard and internal promoter (red arches; y-axis indicates PCHi-C *P*-score). Bottom: positive correlation between *PPM1H* expression across sorted blood cell populations and ATAC-seq signal[45] (100 bp sliding window) in an ~500-bp-long chromosomal segment in intron 1 (red peak)[45] (y-axis indicates false discovery rate for Pearson correlation). (B) Four credible set variants (supplemental Table 9) map to the identified segment, which is accessible in HSCs, MPPs, CMPs, and MEPs (y-axis indicates ATAC-seq signal). (C) Luciferase activities of the 4 candidate causal variants in K562 cells (3 biological replicates). *P* values for one-sided Student *t* test in the *cis*-eQTL direction. Data normalized to empty vector control (Ctrl). (D) MYB motif logo; rs772557[A>G] alters a critical recognition base. (E) Chromatin immunoprecipitation with massively parallel DNA sequencing for MYB in Jurkat cells (rs772557-heterozygous) showing exclusive pull-down of rs772557-G reads. (F) siRNA knockdown of MYB in K562 cells attenuates rs772557-G luciferase activity (3-4 biological replicates). *P* value is for one-sided Student *t* test. (G) CRISPR/Cas9 disruption of rs772557 (supplemental Figure 16) in K562 cells (5 biological targeting replicates; 7 biological control replicates). *P* values are for one-sided Student *t* test. (H) Coexpression of *PPM1H* and *MYB* across blood cell types. Data are bulked single-cell messenger RNA-seq data from 35 582 blood and bone marrow mononuclear cells.[17] (I-J) CD34$^+$ cell RNA-seq data revealed stronger correlation between *MYB* and *PPM1H* expression in rs772557-G homozygotes (n = 27) than in rs772557-A homozygotes (n = 37). Data are $\log_2$-transformed fragments per kilobase of transcript per million mapped reads (FPKM) values, median-centered per genotype group. *P* and *r* values are for Pearson correlation. a.u., arbitrary units; B, B cells; Baso, basophil; CD4, CD4$^+$ T cells; CD8, CD8$^+$ T cells; CLP, common lymphoid progenitors; Cyc, cycling cells; DC, dendritic cells; ERY, erythroid progenitors; GMP, granulocyte-monocyte progenitors; LMPP, lymphoid-primed multipotent progenitors; mDC, myeloid dendritic cells; Mono, monocyte; n.s., not significant; PC, plasma cells; pDC, plasmacytoid dendritic cells; PreB, pre–B cells; Neut, neutrophil; NK, natural killer cells.

Figures 6 and 7). To correct for multiple testing, variants were partitioned into 5 classes based on genomic annotations, and weighted Bonferroni adjustment was applied, taking into account the predicted functional impact of variants within each class (supplemental Methods).[6]

In a combined analysis of the 2 data sets, 6 association signals reached significance, and 2 were suggestive (within one order of magnitude from Bonferroni thresholds). Conditional analysis uncovered an additional 3 independent signals at the 2q22 locus, increasing the number of significant signals to 9 (Figure 1B), whereas no additional signals were detected at the other loci after accounting for their respective lead variants. No heterogeneity was detected in effect estimates between the discovery and follow-up data sets (Bonferroni-adjusted Cochran's Q P_{het} = not significant for all lead variants) (supplemental Table 5), nor between the different sample collection phases (Bonferroni-adjusted P_{het} = not significant for all lead variants) (supplemental Table 6); only the rare variant rs555647251 displayed significant heterogeneity between blood donors and primary care patients (Bonferroni-adjusted P_{het} = 0.046) (supplemental Table 7). The identified lead variants were polymorphic in all geographic ancestry super-populations in the 1000 Genomes Phase III compendium.[21] The exception was rs555647251, which was only polymorphic in the European (EUR) and American (AMR) super-populations (supplemental Table 8). The proportion of variance explained by the lead variants of the 9 significant signals was 4.6% in the combined data set. Using LD score regression,[12] the total SNP heritability was estimated at 12.7%. We detected enrichments of heritability in chromatin regions accessible in HSPCs[8,9] (Figure 2A), indicating that we preferentially identify variants that act by altering gene regulation intrinsically in these cell types.

Identification of candidate genes

To identify candidate genes based on HSPC-intrinsic effects, we generated expression quantitative trait locus (eQTL) data for sorted $CD34^+$ cells from 155 blood donors using RNA-seq. In addition, we retrieved promoter capture Hi-C (PCHi-C) data[14] for $CD34^+$ cells, and ATAC-seq data for sorted blood cells, including 7 HSPC populations.[8]

We defined the 99% credible sets of probable causal variants (supplemental Table 9) and prioritized genes as candidate genes if they: (1) had a non-synonymous coding variant within the credible set; (2) had a *cis*-eQTL in $CD34^+$ cells (Figure 2B; supplemental Table 10); or (3) the credible set contained a regulatory variant in the promoter region, or in a region with a chromatin looping interaction with the promoter in $CD34^+$ cells[14] (supplemental Figure 8). As regulatory variants, variants in regions with accessible chromatin in HSPCs were considered.[8] Strikingly, the identified candidate genes include both known HSPC-relevant genes (*CXCR4* and *CEBPA*) and genes never previously implicated in HSPC biology (*PPM1H, ENO1*, *RERE*, and *ARHGAP45*) or only minimally studied in this area (*ITGA9*)[22] (Figure 1B).

Genetic overlap with human diseases and traits

To investigate further the impact of the identified variants, we questioned if these variants associate with other traits and diseases. Hence, we searched for coincident associations among variants in LD ($r^2 > 0.8$) with the lead variants in the UK Biobank and the GWAS Catalog.[23-27] We detected coincident associations with variants known to associate with mature blood cell traits (6 of the significant variants, at 2q22/*CXCR4*, 12q14/*PPM1H*, 1p36/*ENO1*-*RERE*, 19p13/*ARHGAP45*, 19q13/*CEBPA*), and hematologic malignancies and autoimmune disorders (variants at 2p22/*CXCR4* and 19p13/*CEBPA*) (supplemental Tables 11 and 12). Of note, 2 of the significant candidate genes underlie autosomal-dominant disorders defined by aberrant HSPC regulation: WHIM (warts, hypogammaglobulinemia, immunodeficiency and myelokathexis) syndrome (*CXCR4*)[28,29] and familial acute myeloid leukemia (*CEBPA*).[30] Furthermore, somatic mutations in *CXCR4* and *CEBPA* have been reported in blood malignancies.[31-33]

Gene expression in human hematopoiesis

We next explored the expression of the candidate genes across hematopoietic cell types. We observed enrichments of expression in HSPC vs non-HSPC populations, both in single-cell RNA-seq data for 35 582 mononuclear blood and bone marrow cells[17] (one-sided Wilcoxon rank sum test, $P = 2.1 \times 10^{-10}$) (Figure 2C; supplemental Figure 9) and in bulk RNA-seq data[8] (one-sided Wilcoxon rank-sum test, $P = 7.1 \times 10^{-5}$) (supplemental Figure 10).

To map expression within the $CD34^+$ compartment, we used single-cell Cellular Indexing of Transcriptomes and Epitopes by Sequencing data for 4905 lineage-negative $CD34^+$ bone marrow cells.[18] This revealed distinct expression biases for the different candidate genes (Figure 2D-E; supplemental Figures 11 and 12; supplemental Table 13), including greater than fourfold enrichments in hematopoietic stem cells (HSCs), multi-potential progenitors (MPPs), megakaryocyte-erythrocyte progenitors (MEPs), and mast cell/basophil progenitors (MBs) for *PPM1H* (5.86- to 14.1-fold enrichment; one-sided Wilcoxon rank sum test, $P = 2.01 \times 10^{-4}$ to 1.35×10^{-11}); in HSC, MPP, and MEP populations for *ITGA9* (4.74- to 11.1-fold; $P = 1.98 \times 10^{-3}$ to 2.38×10^{-10}); in MB, dendritic cell, and lymphoid (Ly) precursors for *CXCR4* (4.00- to 9.79-fold; $P = 1.56 \times 10^{-18}$ to 6.17×10^{-67}); in Ly populations for *RERE* (4.28-fold; $P = 6.40 \times 10^{-7}$); and in MPP, MB, dendritic cell, Ly, and granulocyte-monocyte progenitor populations for *CEBPA* (4.38- to 24.3-fold; $P = 2.55 \times 10^{-3}$ to 2.32×10^{-80}). In addition, we observed greater than twofold enrichment of expression in MPPs for *ENO1* (2.18- to 2.23-fold; $P = 2.98 \times 10^{-95}$ to 1.25×10^{-101}) and in Ly for *ARHGAP45* (2.28- to 2.99-fold; $P = 4.96 \times 10^{-5}$ to 1.16×10^{-5}) (supplemental Table 13). These data further support that the identified candidate genes are relevant to HSPCs.

Associations with *CXCR4*

Four signals map to *CXCR4* (C-X-C chemokine receptor type 4) at 2q22. This receptor binds stromal-derived factor-1 (SDF-1; also known as CXCL12). Internalization of CXCR4 is required for HSPC egression from the bone marrow.[34] In the WHIM syndrome, gain-of-function mutations in the C-terminal region prevent internalization after stimulation, leading to retention of HSPCs and other white blood cells in the bone marrow.[35] CXCR4 inhibitors are one of the current methods for mobilizing $CD34^+$ cells for leukapheresis.[36] The fact that we identify association with *CXCR4* provides proof-of-principle for our idea that regulators of blood $CD34^+$ cell levels can be exposed in vivo in humans through genetic variation.

The *CXCR4* signals represent 3 common and 1 rare variant (Figure 1B). Within each of their credible sets (supplemental Table 9), we identified a single regulatory variant that either maps to, or has a looping interaction with, the *CXCR4* promoter, making them plausible causal variants (Figure 3A; supplemental Figure 8A). In the sets of rs11688530 and rs555647251, rs59222832 and rs770321415 were identified as regulatory promoter variants. In the other 2 sets, the lead variants rs309137 and rs10193623 were identified as likely causal based on looping interactions.

Because CXCR4 is a known negative regulator of HPSC regression, we looked for effects of the identified variants on *CXCR4* expression in $CD34^+$ cells in the direction opposite their effects on blood $CD34^+$ cell levels. Indeed, multivariable analysis of our $CD34^+$ cell RNA-seq data for blood donors unveiled conditional *CXCR4 cis*-eQTLs for the 3 common variants (one-sided linear regression, $P = 5.4 \times 10^{-3}$ to 1.8×10^{-2}) (Figure 3B), although the rare variant rs555647251 was not polymorphic in this data set. Furthermore, using CRISPR/Cas9, we deleted the 4 putative causal variants in MOLM13 acute myeloid leukemia cells (supplemental Table 4; supplemental Figures 13 and 14), yielding *CXCR4* downregulation (Figure 3C). These data suggest that the 2q22 associations are caused by genetic variation in genomic regions required for *CXCR4* expression in HSPCs.

Association with *PPM1H*

Our most significant association maps to *PPM1H* (protein phosphatase, Mg^{2+}/Mn^{2+} dependent 1H), encoding an evolutionarily conserved serine/threonine phosphatase.[37,38] The few studies that have been performed suggest that *PPM1H* could be involved in cell signaling,[39-41] trastuzumab resistance,[42] lupus,[43] and colon cancer.[44] Notably, our *cis*-eQTL analysis revealed an anticorrelation between *PPM1H* expression and $CD34^+$ cell levels (Figure 2B).

The *PPM1H* credible set contains 32 variants in intron 1 (Figure 4A; supplemental Table 9). The associated region has looping interactions, both with the standard promoter and an internal promoter (Figure 4A; supplemental Figure 8B). By integrating ATAC- and RNA-seq data for sorted blood cell types,[45] a positive correlation was found between *PPM1H* expression and chromatin accessibility in an ~500-bp-long segment within the associated region (one-sided Pearson correlation, $P = 4.2 \times 10^{-9}$; false discovery rate $= 2.5 \times 10^{-5}$) (Figure 4A). Consistent with the gene expression data analysis (Figure 2C-E), this segment is accessible in HSCs, MPPs, MEPs, and common myeloid progenitors (CMPs) (Figure 4B).

The identified segment overlaps 4 credible set variants (rs772555, rs772556, rs772557, and rs772559). For functional fine-mapping, we conducted luciferase experiments with constructs representing their reference and alternative alleles in K562 erythroleukemia cells. Significantly higher luciferase activity was observed with the rs772557-G construct than with the rs772557-A construct (one-sided Student *t* test, $P = 9.8 \times 10^{-3}$) (Figure 4C), consistent with the *cis*-eQTL direction (rs772557[A>G] is in LD with rs669585[T>G]; $r^2 = 0.98$). We also found higher accessibility at rs772557-G than at rs772557-A in DNase-sequencing data[46] for heterozygous adult $CD34^+$ cells (125 vs 47 reads; binomial test, $P = 2.0 \times 10^{-8}$).

Motif analysis predicted that the rs772557-G allele contains an MYB transcription factor–binding site that is abrogated by the rs772557-A allele (Figure 4D; supplemental Table 14). Consistent with this finding, chromatin immunoprecipitation sequencing for MYB in Jurkat cells (heterozygous for rs772557) showed exclusive pull-down of rs772557-G (Figure 4E; supplemental Figure 15). In addition, luciferase experiments with cotransfected MYB siRNA showed selective attenuation of signal from rs772557-G construct but had no impact on signal from rs772557-A construct (one-sided Student *t* test, $P = 2.7 \times 10^{-4}$ for MYB siRNA-treated cells vs controls) (Figure 4F).

For further functional validation, we perturbed rs772557 by CRISPR/Cas9 in K562 cells, which are triploid-heterozygous for rs772557, having 2 copies of the rs772557-G allele and 1 copy of the rs772557-A allele. Serendipitously, we identified sgRNA sequences that overlap rs772557 and enable allele-specific disruption by cutting DNA only 1 bp upstream of rs772557 (supplemental Figure 16; supplemental Table 4). We observed *PPM1H* downregulation with rs772557-G sgRNA (one-sided Student *t* test, $P = 2.3 \times 10^{-3}$ for rs772577-G-sgRNA–treated vs sgRNA-control–treated cells; $P = 7.1 \times 10^{-3}$ for rs772557-G-sgRNA–treated vs rs772557-A-sgRNA-treated cells) (Figure 4G). In addition, *MYB* and *PPM1H* are coexpressed across blood cell types (Figure 4H), and, in our $CD34^+$ cell RNA-seq data for blood donors, stronger positive correlation was observed between *MYB* and *PPM1H* expression among 27 rs772557-G homozygotes (Pearson correlation $r = 0.74$; $P = 1.2 \times 10^{-5}$) (Figure 4I) than among 37 rs772557-A homozygotes (Pearson correlation $r = 0.51$; $P = 1.1 \times 10^{-3}$) (Figure 4J), with a borderline-significant difference between the 2 correlation coefficients (one-sided Fisher *Z* test, $P = .08$). Consistent with an additive allele dose effect, an intermediate correlation was observed among 85 rs772557-A/G heterozygotes (Pearson correlation $r = 0.56$; $P = 2.3 \times 10^{-8}$). Finally, shRNA-knockdown of *PPM1H* in primary $CD34^+$ umbilical cord blood cells resulted in higher proportions of $CD34^+$ and $CD34^+90^+$ cells relative to control (supplemental Figures 17 and 18). These data identify rs772557 as a causal variant and further support a functional anticorrelation between *PPM1H* expression and blood $CD34^+$ cell levels.

Additional associations

Among the remaining significant associations, 1p36/*ENO1-RERE* and 3p22/*ITGA9* exhibit *cis*-eQTLs in $CD34^+$ cells (Figure 2B). *ENO1* encodes the glycolytic enzyme α-enolase 1; a shorter isoform binds the *MYC* promoter as a tumor suppressor.[47] *RERE* encodes a transcription factor that binds the retinoic acid receptor.[48] The credible set contains a candidate causal regulatory variant between *ENO1* and *RERE* that interacts with both promoters (supplemental Figure 8C; supplemental Table 9). *ITGA9* encodes integrin α-9, which is expressed in HSPCs,[49] but its role remains unclear. The identified association maps to *ITGA9* introns 3 and 4, with four regulatory variants in intron 3 looping to the promoter (supplemental Figure 8D; supplemental Table 9). Finally, the 19p13 signal tags a missense variant in *ARHGAP45*, encoding minor histocompatibility protein HA-1, a regulator of T- and B-cell migration.[50] Interestingly, recent mouse data suggest that *Arhgap45* also regulates HSPC engraftment in bone marrow.[51]

Among the suggestive associations, the 8q24 signal maps to *CCDC26*, but PCHi-C analysis revealed a long-distance looping interaction with the *MYC* promoter, 1.86 Mb away (supplemental Figure 8H). Interestingly, the identified signal corresponds to the Blood ENhancer Cluster (BENC), an evolutionarily conserved "super-enhancer" required for *Myc* expression in mouse HSPCs.[52] The BENC comprises 8 enhancer modules (denoted A to H). Our signal spans module D. Deletion of this module in mice affects *Myc* expression in HSCs and MPPs. Finally, the second suggestive signal maps to *TERT*. Mutations in *TERT* cause dyskeratosis congenita, in which impaired telomere maintenance leads to problems with HSPC regeneration and increased risk of myelodysplastic syndrome.[53] Somatic mutations in *TERT* and *MYC* have also been reported in blood malignancies.[33]

Discussion

We report the first large-scale GWAS on blood CD34$^+$ cell levels. We identify 9 significant association signals explaining approximately one-third of the total estimated SNP heritability (4.6% of 12.7%). This study represents the first successful GWAS on a stem cell trait, and the results provide proof-of-principle for the idea that HSPC regulators can be identified in vivo in humans through genetic variation.

We identified associations with known regulators of blood CD34$^+$ cell levels, as well as with 5 novel regulators (*PPM1H*, *ENO1*, *RERE*, *ITGA9*, and *ARHGAP45*). Understanding the mechanisms and molecular pathways that underlie these associations poses intriguing challenges. The clinical efficacy of CXCR4 antagonists in the context of stem cell mobilization, and our data (Figure 3), suggest that the rate of egression of CD34$^+$ cells is the key variable in the case of the *CXCR4* associations. At the same time, variation in blood CD34$^+$ cell levels can reflect variation in several physiological variables, including egression rate, time spent in the bloodstream, and stem cell pool size. Potentially, the identified variants and genes could affect either, or several, of these variables.

Today, the most common method to harvest stem cells for transplantation is by leukapheresis, requiring mobilization of HSPCs into the blood. Current mobilization regimens include cyclophosphamide, granulocyte colony-stimulating factor, and CXCR4 inhibitors. Intriguingly, for the most significant locus, we discovered an anticorrelation between *PPM1H* expression and CD34$^+$ levels, supported by *cis*-eQTL, functional fine-mapping, and shRNA-knockdown data (Figures 2B and 4). This finding suggests that PPM1H could be used as an inhibition target to facilitate stem cell harvesting by leukapheresis, although further studies are needed to understand the biology of PPM1H and to identify pharmacologic inhibitors. Regarding potential side effects of such inhibitors, *PPM1H* has not been associated with any severe disease according to GWAS Catalog and UK Biobank data.[27,54] Although the gene is conserved against loss-of-function variants relative to synonymous coding variants (observed/expected score 0.18 vs 0.98 in the Genome Aggregation Database, gnomAD[55]), the UK Biobank database contains 32 individuals with loss-of-function *PPM1H* variants (https://www.ukbiobank.ac.uk and https://azphewas.com), all of whom were aged >40 years at recruitment, with >10 years of follow-up.[26] Moreover, the effects of any inhibitor, when used in the context of stem cell mobilization, will likely be short term. Other notable findings include ENO1 as a regulatory enzyme with links to the MYC pathway, RERE as a transcription factor with links to retinoic acid signaling, and ITGA9 as a cell surface marker with a functional role in the regulation of blood CD34$^+$ cell levels. In all, we report the first large-scale analysis of the genetic architecture of blood CD34$^+$ cell levels, with potential implications for stem cell harvesting and transplantation.

Acknowledgments

The authors are indebted to Ellinor Jonsson and the personnel at the Clinical Chemistry and Clinical Immunology and Transfusion Medicine in Region Skåne for their assistance, and the patients and blood donors who participated in the study.

This work was supported by grants from the European Research Council (CoG-770992), the Knut and Alice Wallenberg Foundation (2014.0071 and 2017.0436), the Swedish Research Council (2017-02023 and 2018-00424), the Swedish Cancer Society (2017/265 and 20.0694), the Swedish Children's Cancer Fund (PR2018-0118 and TJ2017-0042), and the Inga Britt & Arne Lundberg Research Foundation (2017-0055).

Authorship

Contribution: A.L.d.L.P., U.T., J.L., K.S., and B.N. conceived the study; A.L.d.L.P., L.E., C.C., Z.A., M.P., K.Ž., G.N., U.T., J.L., K.S., and B.N. designed experiments; A.L.d.L.P., C.C., and N.M. developed flow cytometry protocols; L.E. and A.L.A. developed AliGater; A.L.d.L.P., L.E., C.C., Z.A., N.M., N.U.D., and D.B. performed the phenotyping; A.L.d.L.P., C.C., G.T., M.P., U.T., and K.S. performed genetic analyses; C.C., Z.A. N.M., and K.Ž. conducted functional experiments; A.L.d.L.P., L.E., C.C., Z.A., N.M., G.T., K.Ž., A.L.A., M.T., M.P., P.D., E.B., A.N., V.G.S., G.K., and B.N. performed data analyses; and A.L.d.L.P., L.E., C.C., Z.A., and B.N. drafted the manuscript. All authors contributed to the final manuscript.

Conflict-of-interest disclosure: G.T., G.N., U.T., and K.S. are employed by deCODE genetics/Amgen Inc. V.G.S. serves as an advisor to and/or has equity in Branch Biosciences, Ensoma, Novartis, Forma, and Cellarity, all unrelated to this work. The remaining authors declare no competing financial interests.

ORCID profiles: L.E., 0000-0002-4257-7284; K.Ž., 0000-0002-4098-0184; M.T., 0000-0001-6244-3841; P.D., 0000-0002-8070-7238; D.B., 0000-0003-3178-1586; N.U.D., 0000-0003-4938-3973; V.G.S., 0000-0003-0044-443X.

Correspondence: Björn Nilsson, Hematology and Transfusion Medicine, Department of Laboratory Medicine, BMC B13, 221 84 Lund, Sweden; e-mail: bjorn.nilsson@med.lu.se.

Footnotes

Submitted 6 July 2021; accepted 11 December 2021; prepublished online on *Blood* First Edition 10 January 2022. DOI 10.1182/blood.2021013220.

*L.E., C.C., and Z.A. contributed equally to this study and are joint second authors.

The sequencing data and counts of fragments per kilobase of transcript per million mapped reads are available from the European Genome Archive repository (accession numbers EGAS00001005655 and EGAD00001008194).

The online version of this article contains a data supplement.

The publication costs of this article were defrayed in part by page charge payment. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

REFERENCES

1. Notta F, Zandi S, Takayama N, et al. Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science*. 2016;351(6269):aab2116.

2. Barnett D, Janossy G, Lubenko A, Matutes E, Newland A, Reilly JT; General Haematology Task Force of the British Committee for Standards in Haematology. Guideline for the flow cytometric enumeration of CD34+ haematopoietic stem cells. Prepared by the CD34+ Haematopoietic Stem Cell Working Party. *Clin Lab Haematol*. 1999;21(5):301-308.

3. Cohen KS, Cheng S, Larson MG, et al. Circulating CD34(+) progenitor cell frequency is associated with clinical and genetic factors. *Blood*. 2013;121(8):e50-e56.

4. Eidenschink L, Dizerega G, Rodgers K, et al. Basal levels of CD34 positive cells in peripheral blood differ between individuals and are stable for 18 months. *Cytometry B Clin Cytom*. 2012;82(1):18-25.

5. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst*. 1959; 22(4):719-748.

6. Sveinbjornsson G, Albrechtsen A, Zink F, et al. Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat Genet*. 2016;48(3):314-317.

7. Maller JB, McVean G, Byrnes J, et al; Wellcome Trust Case Control Consortium. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet*. 2012;44(12):1294-1301.

8. Ulirsch JC, Lareau CA, Bao EL, et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat Genet*. 2019;51(4):683-693.

9. Corces MR, Buenrostro JD, Wu B, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet*. 2016; 48(10):1193-1203.

10. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol*. 2015;109:21.29.1-21.29.9.

11. Meuleman W, Muratov A, Rynes E, et al. Index and biological spectrum of human DNase I hypersensitive sites. *Nature*. 2020; 584(7820):244-251.

12. Bulik-Sullivan BK, Loh PR, Finucane HK, et al; Schizophrenia Working Group of the Psychiatric Genomics Consortium. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet*. 2015;47(3):291-295.

13. Finucane HK, Bulik-Sullivan B, Gusev A, et al; RACI Consortium. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet*. 2015;47(11):1228-1235.

14. Mifsud B, Tavares-Cadete F, Young AN, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet*. 2015;47(6):598-606.

15. Mansour MR, Abraham BJ, Anders L, et al. Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science*. 2014;346(6215):1373-1377.

16. Gioia L, Siddique A, Head SR, Salomon DR, Su AI. A genome-wide survey of mutations in the Jurkat cell line. *BMC Genomics*. 2018;19(1):334.

17. Granja JM, Klemm S, McGinnis LM, et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat Biotechnol*. 2019;37(12): 1458-1465.

18. Sommarin MNE, Dhapola P, Safi F, et al. Single-cell multiomics reveals distinct cell states at the top of human hematopoietic hierarchy. *bioRxiv*. Preprint posted online 2 April 2021. doi: https://doi.org/10.1101/2021.04.01.437998.

19. Becht E, McInnes L, Healy J, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol*. 2018;37(1):38-47.

20. van Dijk D, Sharma R, Nainys J, et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*. 2018; 174(3):716-729.e27.

21. Auton A, Brooks LD, Durbin RM, et al; 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74.

22. Schreiber TD, Steinl C, Essl M, et al. The integrin α9β1 on hematopoietic stem and progenitor cells: involvement in cell adhesion, proliferation and differentiation. *Haematologica*. 2009;94(11):1493-1501.

23. Staley JR, Blackshaw J, Kamat MA, et al. PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics*. 2016;32(20):3207-3209.

24. Kamat MA, Blackshaw JA, Young R, et al. PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics*. 2019;35(22):4851-4853.

25. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*. 2015;31(21):3555-3557.

26. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726): 203-209.

27. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019;47(D1):D1005-D1012.

28. Milanesi S, Locati M, Borroni EM. Aberrant CXCR4 signaling at crossroad of WHIM syndrome and Waldenstrom's macroglobulinemia. *Int J Mol Sci*. 2020; 21(16):1-15.

29. Heusinkveld LE, Majumdar S, Gao JL, McDermott DH, Murphy PM. WHIM syndrome: from pathogenesis towards personalized medicine and cure. *J Clin Immunol*. 2019;39(6):532-556.

30. Smith ML, Cavenagh JD, Lister TA, Fitzgibbon J. Mutation of CEBPA in familial acute myeloid leukemia. *N Engl J Med*. 2004;351(23):2403-2407.

31. Kaiser LM, Hunter ZR, Treon SP, Buske C. CXCR4 in Waldenström's macroglobulinema: chances and challenges. *Leukemia*. 2021;35(2):333-345.

32. Panuzzo C, Signorino E, Calabrese C, et al. Landscape of tumor suppressor mutations in acute myeloid leukemia. *J Clin Med*. 2020; 9(3):802.

33. Mitelman F, Johansson B and Mertens F (Eds.) Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer (2022). https://mitelmandatabase.isb-cgc.org.

34. Karpova D, Bonig H. Concise review: CXCR4/CXCL12 signaling in immature hematopoiesis—lessons from pharmacological and genetic models. *Stem Cells*. 2015;33(8):2391-2399.

35. Hernandez PA, Gorlin RJ, Lukens JN, et al. Mutations in the chemokine receptor gene CXCR4 are associated with WHIM syndrome, a combined immunodeficiency disease. *Nat Genet*. 2003;34(1):70-74.

36. Liles WC, Broxmeyer HE, Rodger E, et al. Mobilization of hematopoietic progenitor cells in healthy volunteers by AMD3100, a CXCR4 antagonist. *Blood*. 2003;102(8): 2728-2730.

37. Madeira F, Park YM, Lee J, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res*. 2019;47(W1): W636-W641.

38. Toft N, Birgens H, Abrahamsson J, et al. Results of NOPHO ALL2008 treatment for patients aged 1-45 years with acute lymphoblastic leukemia. *Leukemia*. 2018;32(3): 606-615.

39. Sugiura T, Noguchi Y. Substrate-dependent metal preference of PPM1H, a cancer-associated protein phosphatase 2C: comparison with other family members. *Biometals*. 2009;22(3):469-477.

40. Chen MJ, Dixon JE, Manning G. Genomics and evolution of protein phosphatases. *Sci Signal*. 2017;10(474):eaag1796.

41. Berndsen K, Lis P, Yeshaw WM, et al. PPM1H phosphatase counteracts LRRK2 signaling by selectively dephosphorylating Rab proteins. *eLife*. 2019;8:1-37.

42. Lee-Hoeflich ST, Pham TQ, Dowbenko D, et al. PPM1H is a p27 phosphatase implicated in trastuzumab resistance. *Cancer Discov*. 2011;1(4):326-337.

43. Ghodke-Puranik Y, Imgruet M, Dorschner JM, et al. Novel genetic associations with interferon in systemic lupus erythematosus identified by replication and fine-mapping of trait-stratified genome-wide screen. *Cytokine*. 2020;132:154631.

44. Sugiura T, Noguchi Y, Sakurai K, Hattori C. Protein phosphatase 1H, overexpressed in

colon adenocarcinoma, is associated with CSE1L. *Cancer Biol Ther.* 2008;7(2):285-292.

45. Ulirsch JC, Nandakumar SK, Wang L, et al. Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell.* 2016;165(6):1530-1545.

46. Davis CA, Hitz BC, Sloan CA, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* 2018; 46(D1):D794-D801.

47. Ghosh AK, Steele R, Ray RB. Functional domains of c-*myc* promoter binding protein 1 involved in transcriptional repression and cell growth regulation. *Mol Cell Biol.* 1999; 19(4):2880-2886.

48. Vilhais-Neto GC, Maruhashi M, Smith KT, et al. Rere controls retinoic acid signalling and somite bilateral symmetry. *Nature.* 2010;463(7283):953-957.

49. Peng Y, Wu D, Li F, Zhang P, Feng Y, He A. Identification of key biomarkers associated with cell adhesion in multiple myeloma by integrated bioinformatics analysis. *Cancer Cell Int.* 2020;20(1):262.

50. Bleakley M, Riddell SR. Exploiting T cells specific for human minor histocompatibility antigens for therapy of leukemia. *Immunol Cell Biol.* 2011;89(3):396-407.

51. He L, Valignat M-P, Zhang L, et al. ARHGAP45 controls naïve T- and B-cell entry into lymph nodes and T-cell progenitor thymus seeding. *EMBO Rep.* 2021;22(4): e52196.

52. Bahr C, von Paleske L, Uslu VV, et al. A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies [published correction appears in *Nature*. 2018;558(7711):E4]. *Nature.* 2018; 553(7689):515-520.

53. Vulliamy T, Marrone A, Goldman F, et al. The RNA component of telomerase is mutated in autosomal dominant dyskeratosis congenita. *Nature.* 2001;413(6854):432-435.

54. Karczewski KJ, Francioli LC, Tiao G, et al; Genome Aggregation Database Consortium. The mutational constraint spectrum quantified from variation in 141,456 humans [published correction appears in *Nature.* 2021;590(7846):E53]. *Nature.* 2020; 581(7809):434-443.

55. Wang Q, Dhindsa RS, Carss K, et al; AstraZeneca Genomics Initiative. Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature.* 2021; 597(7877):527-532.

© 2022 by The American Society of Hematology

Update

Blood
Volume 141, Issue 8, 23 February 2023, Page 962

DOI: https://doi.org/10.1182/blood.2022019221

blood®

Errata

Markewitz RDH, Falk KK. Platelet satellitism in immune thrombocytopenic purpura. *Blood.* 2022;140(19):2091.

In the second paragraph, "glycoprotein IIa/IIIb complex" should read "glycoprotein IIb/IIIa complex."

https://doi.org/10.1182/blood.2022019086

© 2023 by The American Society of Hematology

Lopez de Lapuente Portilla A, Ekdahl L, Cafaro C, et al. Genome-wide association study on 13 167 individuals identifies regulators of blood CD34$^+$ cell levels. *Blood.* 2022;139(11):1659-1669.

In "Acknowledgments" on page 1667, funding from the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie grant agreements 754299 and 847583 should have been cited.

https://doi.org/10.1182/blood.2022019221

© 2023 by The American Society of Hematology

Paper II

scientific reports

OPEN

Functional dissection of the gene regulatory mechanism underlying variation in blood CD34⁺ cell levels at 1p36.23/*ENO1*

Zain Ali[1,2], Caterina Cafaro[1,2], Ludvig Ekdahl[1,2], Antton Lamarca[1,2], Aitzkoa Lopez de Lapuente Portilla[1,2] & Björn Nilsson[1,2,3,4]✉

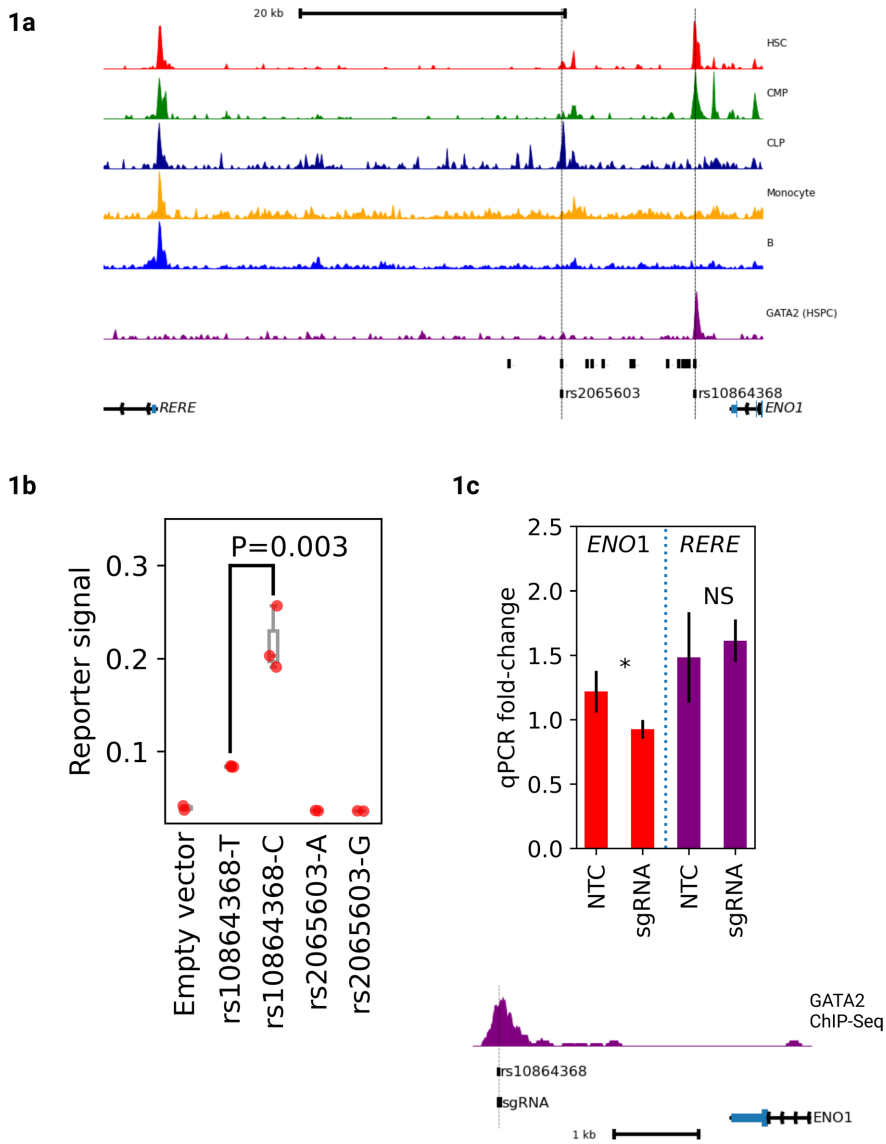
Understanding the genetic regulation of blood CD34⁺ cell levels could lead to better methods for stem cell mobilization and transplantation. Recently, in a genome-wide association study (GWAS), we identified 11 genetic associations with blood CD34⁺ cell levels. One of the most significant maps to an intergenic region between the *ENO1* and *RERE* genes at chromosome 1p36.23. However, the underlying mechanism remains unknown. Here, using functional fine-mapping, we identify rs10864368:C>T as a causal non-coding variant. Specifically, the rs10864368-T allele abrogates a binding site for the GATA2 transcription factor, leading to downregulation of *ENO1* and *RERE* and lower CD34⁺ cell levels. Our data link a specific GATA2 target to the regulation of CD34⁺ cell levels, expanding on the role of GATA2 in human hematopoieisis.

Keywords Hematopoietic stem and progenitor cells, Stem cell transplantation, Gata-2

Circulating CD34⁺ hematopoietic stem and progenitor cells (HPSCs) are routinely used for stem cell transplantation[1]. These cells are harvested through leukapheresis of peripheral blood. Understanding the genetic regulation of blood CD34⁺ cell levels could expose new drug targets for stem cell mobilization and new insight into stem cell biology in humans. As baseline levels of CD34⁺ cells in peripheral blood are an important predictor of stem cell mobilization, better genetic understanding of these cell populations could be clinically relevant[2].

In a genome-wide association study encompassing 13,167 individuals from Sweden, we recently identified 11 independent genetic associations with blood CD34⁺ cell levels[3]. One of the most significant associations mapped to an intergenic region between the genes *ENO1* (Enolase-1) and *RERE* (Arginine-Glutamic Acid Dipeptide Repeats) on chromosome 1p36.23 (lead variant rs2047094; minor allele frequency = 0.49, $P = 1.3 \times 10^{-15}$, $\beta = -0.102$ with blood CD34⁺ cell levels). *ENO1* catalyzes the conversion of 2-phosphoglycerate to phosphoenolpyruvate, the penultimate step in glycolysis. *RERE* encodes a transcriptional coactivator for the retinoic acid receptor[4,5]. Through expression quantitative locus (eQTL) analysis, we found that the rs2047094-A allele (which confers higher blood CD34⁺ cell levels) is associated with higher expression of *ENO1* and, to a lesser extent, *RERE* in primary blood CD34⁺ cells. Consistent with this, promoter capture Hi-C data for blood CD34⁺ cells[6] revealed chromatin looping interactions with the *ENO1* and *RERE* promoters. Our previous work thus indicates that the genetic association at 1p36.23 with CD34⁺ cell levels is driven by cell-autonomous changes in *ENO1* and/or *RERE* expression in HSPCs. However, the 1p36.23 association is represented by 17 variants in high linkage disequilibrium (LD) spanning a ~ 14 kB region (Fig. 1a) and the variant underlying the changes in gene expression remains unknown. To address this, we conducted a functional fine-mapping study aimed at pinpointing the causal variants and their mechanisms of action.

[1]Division of Hematology and Transfusion Medicine, Department of Laboratory Medicine, BMC B13, Lund University, Lund SE-221 84, Sweden. [2]Lund Stem Cell Center, Lund University, Lund SE-221 84, Sweden. [3]Division of Hematology, Helsingborg Hospital, Charlotte Yhléns gata 10, 251 87 Helsingborg, Sweden. [4]Broad Institute, 415 Main Street, Cambridge, MA 02142, USA. ✉email: bjorn.nilsson@med.lu.se



Methods

Cell culture

All assays were performed using K-562 erythroleukemia cells (DSMZ, #ACC 10) which are routinely used to study HSPC function and are naturally heterozygous for rs10864368. K-562 cells were cultured in RPMI 1640 supplemented with 10% heat-inactivated FCS (Gibco) at 37 °C with 5% CO2 and maintained at a density of 0.5–1.0.5.0 million cells/mL.

ATAC-Seq and ChIP-Seq data analysis

Chromatin accessibility data for primary blood cell types was accessed using GEO (GSE74912)[7]. GATA2 and SMAD1 ChIP-Seq data for primary HSPCs were accessed using GEO (GSE60793, GSE104574)[8,9]. Chromatin accessibility and ChIP-Sequencing data for K-562 cells were accessed using ENCODE for allelic imbalance

◀**Fig. 1**. rs10864368 is the causal variant for the 1p36.23 association (**a**) *Top*: ATAC-Sequencing data[7] from 5 different primary cell types and ChIP-Sequencing data[6] for GATA2 in primary HSPCs. HSC: Hematopoietic stem cell, CMP: common myeloid progenitor, CLP: common lymphoid progenitor. Y-axis represents normalized read depth averaged over 500 bp windows. *Middle*: Unlabelled black bars represent 17 high LD variants corresponding to the 1p36.23 association. Dotted vertical lines show 2 variants directly intersecting chromatin accessibility peaks in primary HSPCs. *Bottom*: Gene diagrams showing locations of *ENO1* and *RERE*. (**b**) Luciferase reporter data comparing alleles of rs2065603 and rs10864368 in K-562 cells (3 biological replicates). Y-axis shows mean luciferase/renilla signal ratios and error bars show standard deviation across biological replicates. rs10864368-C shows significantly higher reporter signal than rs10864368-T (two-sided *t*-test $P = 0.003$). Neither allele of rs2065603 shows any expression above background. (**c**) CRISPR-Cas9 deletion of rs10864368 in K-562 cells (3 biological replicates). *Top*: qPCR data comparing sgRNA deleted cells to non-targeting controls (NTCs) for *ENO1* and *RERE*. Y-axis shows mean qPCR fold-change across biological replicates and error bars show standard deviation. Statistics shown are for one-sided *t*-test: $^*P \leq 0.05$, NS not significant. *Bottom*: Genome plot showing location of sgRNA relative to rs10864368 and GATA2 binding peak.

analysis[10]. Raw FASTQ files were aligned to the hg38 human genome assembly using HISAT2. BAM files were analyzed using IGV to identify heterozygous samples and extract read counts at rs10864368. In cases of allelic imbalance, counts of both alleles were used to evaluate the statistical significance of imbalance using a binomial test.

Luciferase assay

We performed luciferase assays for two candidate variants (rs2065603 and rs10864368). For rs2065603, we constructed 900-bp DNA sequences centered on the variant (Integrated DNA Technologies). For rs10864368, due to synthesis problems, we instead adopted a PCR strategy and amplified an 864-bp sequence containing rs10864368 from genomic DNA extracted from K-562 cells, which are naturally heterozygous for this variant (see Supplementary Table 1 for full reporter sequences). We screened PCR clones using Sanger sequencing until both alleles were recovered. Both sequences were subsequently cloned into the pGL3-Basic vector (E1751, Promega). Reporter plasmids were electroporated into K-562 cells using the Neon electroporation system (Thermo FisherScientific). Luciferase and renilla activities were measured 24 h after electroporation using DualGlo Luciferase (Promega, #E1960) on a GLOMAX 20/20 Luminometer (Promega).

CRISPR/Cas9 deletion of variant region

To delete the rs10864368-harboring region, we used a single guide RNA (sgRNA) directly spanning the variant (Supplementary Table 2). The sgRNA was cloned into SpCas9(BB)−2 A-GFP PX458 vector (Addgene, #48138) and electroporated into K-562 cells using the Neon electroporation system (Thermo FisherScientific). GFP-positive cells were sorted 24 h after transfection. RNA was extracted from the same cells, reverse-transcribed, and quantified by Taqman assays (Hs00201558_m1 for *RERE*; Hs00361415_m1 for ENO1; Hs01060665_g1 for ACTB) using the StepOnePlus qPCR instrument (Applied Biosystems).

Motif analysis

For motif analysis we used PERFECTOS-APE[11] to scan the HOCOMOCO database of transcription factor binding motifs to predict what TF motifs would be most likely to be disrupted by rs10864368. As a complementary approach, we calculated binding efficiencies following a recently described method[12]. We obtained the GATA2 position weight matrix (PWM) from JASPAR database (MA0036.3). For each allele, we extracted a 51 bp genomic sequence centered on the variant position (hg38 chr1:8858254) and scanned all possible 11 bp windows overlapping the variant on both forward and reverse strands. The PWM score for each sequence was calculated as the sum of position-specific log-odds ratios: Score(s) = sum from i = 1 to n of log2(M_si, i/b_si), where s is the DNA sequence of length n (11 bp for GATA2), M_si, i is the frequency of nucleotide si at position i in the PWM, and b_si is the background frequency (0.25 for uniform nucleotide distribution). Binding efficiency was calculated as 2^(PWM score), representing the relative binding affinity. The fold change in binding efficiency between alternate and reference alleles was computed as the ratio of their binding efficiencies. The sequence with the highest PWM score overlapping the variant position on either strand was selected as the best match for each allele.

Results

Our initial GWAS findings identified a LD block of 17 linked variants but could not determine which of these is directly causal for altered HSPC levels. As non-coding variants often act by altering transcription factor binding, we prioritized candidate causal variants that intersect with accessible chromatin in HSPCs. Using ATAC-sequencing data for 18 cell types derived from blood and bone marrow[7], we found that two of the 17 associated variants (rs10864368 and rs2065603; $r^2 = 0.996$ and 0.9034 with rs2047094, respectively) map to accessible chromatin of primary $CD34^+$ cell types, including hematopoietic stem cells (HSC), common myeloid progenitors (CMP), and common lymphoid progenitors (CLP) (Fig. 1a). Particularly, the rs10864368-harboring region is highly accessible in HSCs. To assess the transcriptional activity of the two candidate variants, we performed luciferase reporter assays in K-562 leukemia cells. Luciferase constructs harboring rs10864368-C, which confers higher *ENO1* and *RERE* expression and higher blood $CD34^+$ cell levels, showed higher activity than rs10864368-T constructs (two-sided Student's *t*-test, $P = 0.003$).

At the same time, we could not see a difference between rs2065603-A and rs2065603-G constructs (Fig. 1b). To obtain further support for a regulatory relationship, we disrupted the rs10864368-harboring region in K-562 cells by use of CRISPR/Cas9 with a single guide RNA (sgRNA) directly overlapping the variant. We previously observed a positive correlation between chromatin accessibility at rs10864368 and *ENO1* expression and so tested the hypothesis that this regulatory element acts as a transcriptional activator. Confirming our hypothesis, CRISPR deletion resulted in lowered *ENO1* expression (one-sided Student's *t*-test, $P = 0.042$) but, consistent with our previous study showing a weaker eQTL on *RERE* than for *ENO1*, we observed no significant change in *RERE* expression (Fig. 1c). Thus, our data primarily point to ENO1 being the main candidate gene, though we cannot rule out a role for *RERE* without further functional work. To characterize the effects in an endogenous chromosomal context, we examined allelic imbalance in chromatin accessibility at rs10864368 in ENCODE DNase-Seq data from primary blood $CD34^+$ cells from seven heterozygous donors, observing a significant bias towards rs10864368-C (Fig. 2a; two-sided binomial test $P = 2.3 \times 10^{-10}$). Altogether, these data identify rs10864368 as a likely causal variant, with rs10864368-C associating with increased *ENO1* expression, local chromatin accessibility and blood $CD34^+$ cell levels.

Next, we searched for transcription factors mediating the effects of rs10864386 on *ENO1* expression. Using motif analysis, we found that the most significantly perturbed motif was for GATA2 (Fig. 2b), with rs10864368-T disrupting a canonical GATA motif (*P*-value fold-change: 291.01)[11]. To further validate this prediction, we used a complementary approach[12] to calculate binding affinities for both alleles based on the GATA2 position weight matrix (JASPAR MA0036.3) and find that the C allele demonstrated a 2,012-fold increase in binding efficiency compared to the T allele. GATA2 has an extensively characterized role in stem and progenitor cell biology. In mice, conditional deletion of *Gata2* in hematopoietic cells leads to complete loss of adult HSPCs[13]. In humans, germline mutations in *GATA2* lead to GATA2 Deficiency, a heritable disease that combines immunodeficiency with myelodysplastic syndrome, and somatic mutations are observed in 1–4% of patients with myeloid malignancies[14,15]. ChIP-Seq data[8] show GATA2 binding directly to rs10864368 in primary $CD34^+$ HSPCs (Fig. 1a).

To investigate the effect of rs10864368 on GATA2 binding in an endogenous genomic context, we took advantage of the fact that the K-562 cell line is naturally heterozygous for this variant. We examined allelic imbalance for GATA2 binding at this site using public ENCODE ChIP-Seq data[10]. We observed a significant bias towards rs10864368-C (Fig. 2c, percent C reads: 95%, H_0: $p = 0.5$, two-sided binomial test $P = 5.8 \times 10^{-27}$), further supporting that rs10864368-T disrupts GATA2 binding in the native genomic context. Although K-562 cells are triploid at many loci, whole-genome sequencing data from ENCODE indicate that the region containing rs10864368 is tetraploid. Furthermore, the degree of imbalance is so severe that the binomial test would still be significant in either triploid heterozygous genotype (H_0: $p = 0.33$, two-sided binomial test $P = 1.3 \times 10^{-13}$; H_0: $p = 0.66$, two-sided binomial test $P = 1.2 \times 10^{-45}$). We conclude that rs10864368-T abrogates a GATA2 binding site and reduces local chromatin accessibility in primary HSPCs.

We investigated the developmental dynamics of GATA2 binding at rs10864368 in time-course ChIP-Seq data from primary HSPCs differentiating to red blood cells[9]. GATA2 disappears from the site over the course of erythroid differentiation, but is not displaced by GATA1 as the cell commits to an erythroid fate (Fig. 2d). The signalling TF SMAD1 also transiently binds this site, indicating that the regulatory element integrates BMP signalling with lineage-determining factors like GATA2. Regulatory elements of this kind that integrate signalling and lineage-determining TFs have been shown to be enriched for GWAS variants[9].

Finally, we searched for all other significant phenotypes associated with rs10864368 in GWAS data[16] and find that rs10864368-C is associated with increased frequency of myeloid lineage cells such as neutrophils, and with reduced lymphocyte percentage (Table 1)[17–20]. This suggests that increased GATA2 binding at rs10864368 leads to expansion of myeloid lineage HSPCs such as CMPs, leading to an increase in circulating $CD34^+$ cells and biasing hematopoiesis towards the myeloid lineage.

Discussion

Understanding the genetic regulation of blood $CD34^+$ cell levels could have potential clinical implications for stem cell harvesting and transplantation and provides insights into HSPC physiology in humans. Our data support that the 1p36.23 association with blood $CD34^+$ cell levels is the result of rs10864368-T decreasing GATA2 binding and local chromatin accessibility, leading to downregulation of *ENO1* and *RERE* in primary human $CD34^+$ cells, though the downstream mechanisms of these transcriptional changes remain to be elucidated.

We note that our work has technical limitations that caution against the conclusion that rs10864368 is the only causal variant for this association. Cancer cell lines such as K-562 are imperfect surrogates for primary HSPCs and future work should evaluate GWAS variants in primary cells. Additionally, the CRISPR editing strategy we employed is suboptimal as it employs a single cut site without directly assessing editing efficiency through sequencing. While our experiments provide limited in vivo evidence linking the regulatory element to *ENO1* expression, precise editing through homology directed repair or base editing would provide stronger evidentiary support for the functional impact of rs10864368 in its native genomic context. As many eQTL associations are driven by multiple causal variants, without more comprehensive genome editing experiments in primary HSPCs we cannot completely rule out that other linked variants are responsible for downstream changes in gene expression and HSPC frequency.

Intriguingly, our PheWAS analysis revealed that rs10864368-C is also associated with reduced intrinsic epigenetic age acceleration (Table 1), a measure of epigenetic aging derived from DNA methylation data. As HSPC function has been directly linked to epigenetic age acceleration, the mechanistic link between altered *ENO1* or *RERE* expression and stem cell aging could be an interesting avenue for future research[21,22].

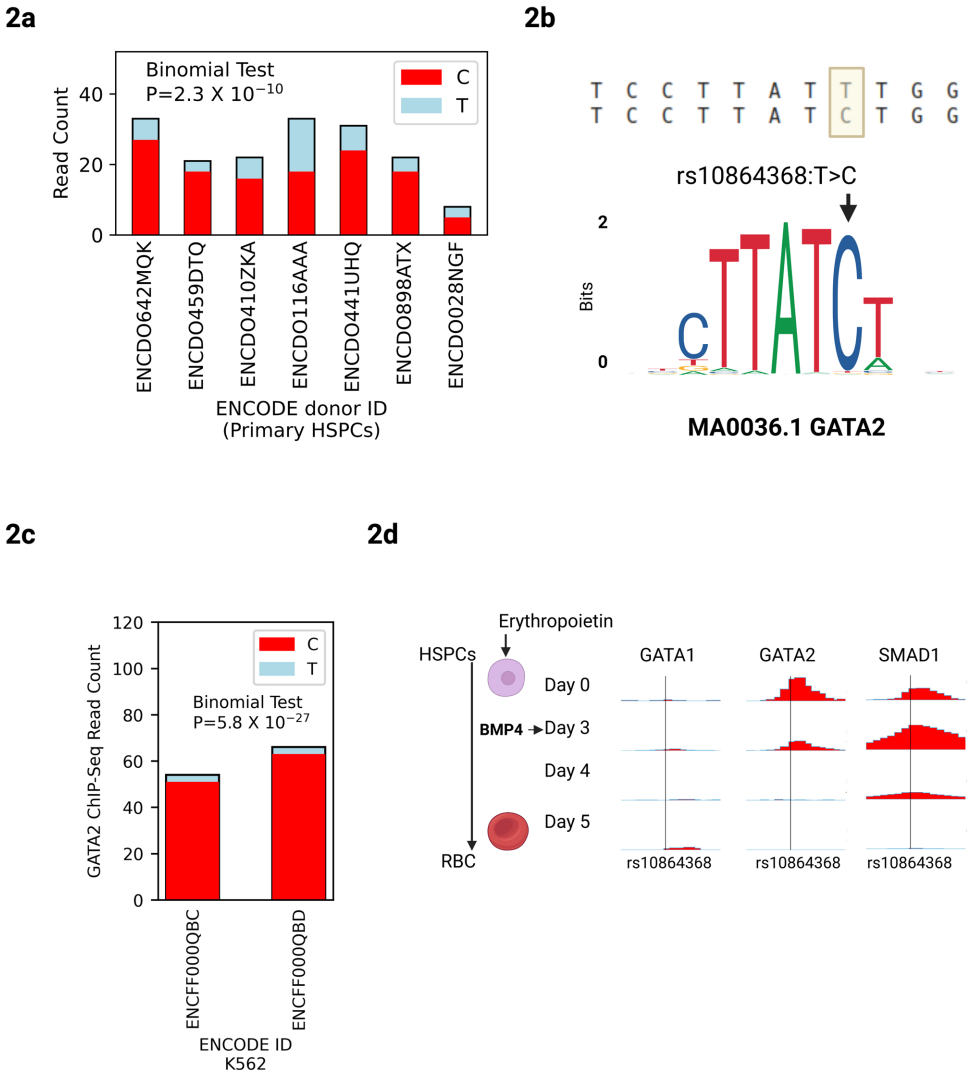


Fig. 2. rs10864368-T reduces chromatin accessibility and GATA2 binding. (**a**) ENCODE DNase-Sequencing data in primary CD34^{+} HSPCS from 7 heterozygous donors[7], showing allelic read counts for rs10864368 as stacked bar plots, showing significant allelic imbalance towards rs10864368-C across all donors. (**b**) *Top*: Local sequence context around rs10864368 highlighting the location of the variant. *Bottom*: JASPAR binding motif for GATA2 showing position affected by rs10864368. (**c**) ChIP-Sequencing data for GATA2 in K-562 cells, showing read counts for rs10864368 as stacked bar plots, for 2 technical replicates. K-562 cells are naturally heterozygous for rs10864368 and show significant imbalance towards rs10864368-C. (**d**) Time-course ChIP-Sequencing data showing temporal dynamics of GATA1, GATA2 and SMAD1 binding at rs10864368 during ex vivo differentiation of primary HSPCs to red blood cells. Y-axes show normalized read depth averaged across 50 bp windows.

Trait	P	Beta	PMID
Plateletcrit	6.59-58	0.0337624	32888494
Platelet count	6.45E-36	0.022872	32888493
Neutrophil count	4.01-35	0.023683	32888493
White blood cell count	1.15E-31	0.021532	32888493
Neutrophil percentage of white cells	7.8E-14	0.0160352	32888494
High light scatter reticulocyte count	8.1E-13	0.0131213	34226706
Lymphocyte percentage of white cells	1E-11	-0.0145583	32888494
Granulocyte count	6.8E-11	0.0233531	27863252
Intrinsic epigenetic age acceleration	1.09E-10	-0.201	34187551
Sum basophil neutrophil counts	1.16E-10	0.0230463	27863252
Sum neutrophil eosinophil counts	1.23E-10	0.0230015	27863252
Myeloid white cell count	3.78E-10	0.0224726	0.0224726
Basophil count	5.59E-10	0.012657	32888493

Table 1. Known GWAS trait associations with rs10864368

While our gene expression data and CRISPR-Cas9 results primarily point to a role for *ENO1*, both *ENO1* and *RERE* remain plausible candidate genes based on eQTL data. As a glycolytic enzyme, *ENO1* expression could influence the balance between glycolysis and mitochondrial respiration. The regulation of energy metabolism is known to play a key role in HSPC function[23]. *RERE* is a co-regulator of retinoic acid signaling[4,5] and animal models have shown that reduced retinoic signalling leads to reduced HSPC numbers[24].

Of note, GATA2 is a central regulator of hematopoietic stem and progenitor cell function and a causal factor in several blood disorders. Here, we link a specific GATA2 target variant to the regulation of blood $CD34^+$ cell levels, which has previously not been reported. In conclusion, our work expands on the role of GATA2 in human hematopoiesis.

Data availability

Data is provided within the manuscript or supplementary information files. All data are contained within the paper.

Received: 31 March 2025; Accepted: 20 November 2025
Published online: 29 November 2025

References

1. Panch, S. R., Szymanski, J., Savani, B. N. & Stroncek, D. F. Sources of hematopoietic stem and progenitor cells and methods to optimize yields for clinical cell therapy. *Biol. Blood Marrow Transplant.* **23**, 1241–1249 (2017).
2. Martino, M. et al. Basal CD34 + cell count predicts peripheral blood stem cell mobilization in healthy donors after administration of granulocyte Colony-Stimulating factor: A Longitudinal, Prospective, Observational, Single-Center, cohort study. *Biol. Blood Marrow Transpl.* **23**, 1215–1220 (2017).
3. Lopez, L. et al. Genome-wide association study on 13 167 individuals identifies regulators of blood CD34 + cell levels. *Blood* **139**, 1659–1669 (2022).
4. Vilhais-Neto, G. C. et al. The WHHERE coactivator complex is required for retinoic acid-dependent regulation of embryonic symmetry. *Nat. Commun.* **8** (1), 8–1 (2017). (2017).
5. Vilhais-Neto, G. C. et al. Rere controls retinoic acid signalling and somite bilateral symmetry. *Nature 2010 463:7283* 463, 953–957 (2010).
6. Mifsud, B. et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics 2015 47:6* 47, 598–606 (2015).
7. Ulirsch, J. C. et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat. Genet.* **51**, 683 (2019).
8. Shearstone, J. R. et al. Chemical Inhibition of histone deacetylases 1 and 2 induces fetal hemoglobin through activation of GATA2. *PLoS One* **11**, (2016).
9. Choudhuri, A. et al. Common variants in signaling transcription-factor-binding sites drive phenotypic variability in red blood cell traits. *Nature Genetics 2020 52:12* 52, 1333–1345 (2020).
10. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).
11. Vorontsov, I. E., Kulakovskiy, I. V., Khimulya, G., Nikolaeva, D. D. & Makeev, V. J. PERFECTOS-APE: Predicting regulatory functional effect of SNPs by approximate P-value estimation. *BIOINFORMATICS –6th International Conference on Bioinformatics Models, Methods and Algorithms, Proceedings; Part of 8th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2015* 102–108 (2015) 102–108 (2015) (2015). https://doi.org/10.5220/0005189301020108
12. Carrasco Pro, S. et al. Widespread perturbation of ETS factor binding sites in cancer. *Nat. Commun.* **14**, 913 (2023).
13. Menendez-Gonzalez, J. B. et al. Gata2 as a crucial regulator of stem cells in adult hematopoiesis and acute myeloid leukemia. *Stem Cell. Rep.* **13**, 291 (2019).
14. Nanaa, A. et al. Clinical and biological characteristics and prognostic impact of somatic GATA2 mutations in myeloid malignancies: a single institution experience. *Blood Cancer Journal 2021 11:6* 11, 1–4 (2021).
15. Calvo, K. R. & Hickstein, D. D. The spectrum of GATA2 deficiency syndrome. *Blood* **141**, 1524–1532 (2023).
16. Elsworth, B. et al. The MRC IEU OpenGWAS data infrastructure. *BioRxiv* **2020.08.10.244293** https://doi.org/10.1101/2020.08.10.244293 (2020).
17. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429e19 (2016).

18. Barton, A. R., Sherman, M. A., Mukamel, R. E. & Loh, P. R. Whole-exome imputation within UK biobank powers rare coding variant association and fine-mapping analyses. *Nat. Genet.* **53**, 1260–1269 (2021).
19. Chen, M. H. et al. Trans-ethnic and Ancestry-Specific Blood-Cell genetics in 746,667 individuals from 5 global populations. *Cell* **182**, 1198–1213e14 (2020).
20. Vuckovic, D. et al. The polygenic and Monogenic basis of blood traits and diseases. *Cell* **182**, 1214–1231e11 (2020).
21. Nachun, D. et al. Clonal hematopoiesis associated with epigenetic aging and clinical outcomes. *Aging Cell.* **20**, e13366 (2021).
22. Ursi, M. et al. Epigenetic age acceleration in hematopoietic stem cell transplantation. *Haematologica* **110**, 481 (2024).
23. Nakamura-Ishizu, A. I. et al. Hematopoietic stem cell metabolism during development and aging. *Dev. Cell.* **54**, 239–255 (2025).
24. Cabezas-Wallscheid, N. et al. Vitamin A-Retinoic acid signaling regulates hematopoietic stem cell dormancy. *Cell* **169**, 807–823e19 (2017).

Acknowledgements
This work was supported by grants from European Research Council (ERC) consolidator grant (CoG-770992 BloodVariome and EU-MSCA-COFUND 754299 CanFaster), the Swedish Research Council (2018 – 00424 and 2017–02023), Swedish Cancer Fund (23-2851), Region Skåne ALF (2022 – 0363), Swedish Children's Cancer Foundation (PR2023-0067), ARMEC Lindberg's Foundation, and Inga-Britt and Arne Lundberg's Foundation.

Author contributions
Z.A., B.N., A.L.d.L.P. designed the study. Z.A. and C.C. performed the experiments. Z.A., L.E., A.L. analyzed the data. Z.A. and B.N. drafted the manuscript. All authors contributed to the final manuscript.

Funding
Open access funding provided by Lund University.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-30068-4.

Correspondence and requests for materials should be addressed to B.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2025