Investigating the role of rare genetic variants in the aetiology of haemostasis disorders

Luca Stefanucci

Churchill College, University of Cambridge

September 2021

This thesis is submitted to the University of Cambridge for the degree of Doctor of Philosophy

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text.

It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text.

It does not exceed the prescribed word limit for the relevant Degree Committee.

The following collaborations have contributed to this work:

1. The lists of genes and variants that have been used in chapter 3 of this thesis have been compiled with the support and the supervision of Dr Elspeth Bruford and Dr Karyn Megy.
2. Dr Dragana Vuckovic has supervised the single variant analysis presented in chapter 3 of this thesis.
3. The relevant VarioPath phenotypes have been extracted in collaboration with Dr Luanluan Sun and Prof Emanuele di Angelantonio.
4. The protein-protein interaction network analyses presented in chapter 3 of this thesis have been performed by Dr Iñigo Hernando-Herraez and Dr Pedro Beltrao.
5. The BeviMed statistical analysis presented in chapter 5 of this thesis has been performed by Dr Daniel Greene and Professor Ernest Turro.
6. Dr Nick Owens has performed the CTCF motif score calculation presented in chapter 5 of this thesis.

Supervisors

Professor Willem H. Ouwehand (University of Cambridge)
Professor Mattia Frontini (University of Exeter)

Abstract

High throughput sequencing and publicly accessible genomic resources increased the diagnostic yields for inherited conditions, and, nowadays, the genetic bases for thousands of Mendelian disorders have been identified. However, providing a molecular diagnosis for these conditions remains challenging, and a considerable portion of patients with inherited conditions still lack a genetic diagnosis.

In clinical genomics, identifying the aetiological variants remains a significant hurdle because they are hidden amongst thousands of rare variants present in the genome of each person. Furthermore, historically only the coding portion of the genome, or so-called exome, has been explored to identify causal variants. It is reasonable to assume that, for a fraction of patients with unexplained inherited diseases, the answer lies in the non-coding portion of the genome. To date, the ability to interpret the functional consequences of variants in the non-coding space remains limited.

My thesis uses large-scale genomics studies and functional genomic techniques to investigate the role of genetic variants in the aetiology of haemostatic diseases. To explore the contribution of rare coding variants to the different phenotypes, I selected all the pathogenic and likely pathogenic variants from a few well-curated resources. Then, I identified carriers of these variants in the UK Biobank cohort and explored their phenotypes. This approach allowed me to estimate the effect sizes of this class of rare variants on the haemostatic phenotypes and investigate their interplay with common ones.

I then expanded my investigation to non-coding regions. I performed experiments to define the most detailed cell type-specific maps of interactions between promoters and regulatory elements for the 93 diagnostic-grade genes for haemostatic diseases. To obtain these interaction landscapes, I differentiated human induced pluripotent stem cells from the principal cell types functionally implicated in haemostasis. I also generated chromatin conformation maps for the relevant genes using a capture Hi-C approach.

Finally, I characterised the captured sequences by annotating them with cell type-specific epigenomic features, and I experimentally examined the regulatory potential of some of the defined regions. These validation experiments were based on two independent approaches: (I) reporter assays (II) perturbation of the epigenetic state for a few identified regions. Furthermore, I assessed the impact of rare genetic variants found in the NIHR BioResource Rare Disease participants, 10% of whom have haemostatic diseases.

The aim of my PhD project was: (i) to investigate the contribution of rare coding variants to different phenotypes, (ii) to improve the annotation of the non-coding space of a set of well-characterised rare diseases genes and, (iii) to improve our ability to provide an accurate molecular diagnosis for individuals with unexplained inherited haemostatic diseases.

Acknowledgements

This thesis has only been possible thanks to the support of many people.

First, I would like to thank my supervisors Professor Willem Ouwehand and Professor Mattia Frontini. They welcomed me into their laboratory and guided me into the fascinating world of genomics, which soon became an instrumental tool in my approach to scientific research. They taught me science and also the importance of sharing scientific ideas and collaborations. I couldn't have asked for better mentors, and I hope we will maintain lifelong relations, if not scientifically, for dinner from time to time. To you, I would like to express my deepest and sincere gratitude.

Second, Dr Elspeth Bruford, Dr Karine Mégy, Profesor Ernest Turro, Dr Dragana Vuckovic and all the collaborators who provided invaluable support in the experiments and analysis of this thesis. I would also like to thank all the colleagues who, daily, helped and accompanied me on this four-year-long journey: Ms Frances Burden, Dr Janine Collins, Ms Samantha Farrow, Dr Nick Gleadall, Dr Luigi Grassi, Dr John Lambourne, Dr Louisa Mayer, Dr Isabel Rosa, Dr Denis Seyres, Dr Matt Sims, Mr Jonathan Stephens, Dr Ana Rita Tomé and all the other members of Willem's and Mattia's groups in Cambridge and Exeter. Every one of you contributed to my scientific growth.

Third, my family, either in Italy or the one I built here. My mother, Annabella, is always the first to believe in me, despite not having a clear idea of what I am doing. My sister, Agnese, who, with her large calisthenic shoulders, took on double responsibilities to cover the absence of a crazy brother. My little brother, Emanuele, to whom I am deeply attached, and I feel like I owe him an apology because I missed all his childhood experiences. Last but not least, my partner, Anna Maria, who shared with me the rollercoaster of emotions I went through, supported my insecurities and has always been the positive side of my negative days.

To conclude, I would like to express my gratitude to the British Heart Foundation, the NHS Blood and Transplant, and Churchill College, which sponsored the work I performed during my PhD and allowed me to present it at national, and international conferences.

Abbreviation table

Abbreviations/Acronyms	Full-form
A	Adenine
AD	Autosomal dominant
AF	Allele frequencies
APTT	activated partial thromboplastin time
AR	Autosomal recessive
AT	antithrombin
bp	Base pairs
BSS	Bernard–Soulier syndrome
BTPD	Bleeding, thrombotic and platelet disorders
C	Cytosine
CAMT	Congenital Amegakaryocytic Thrombocytopenia
CBC	Complete blood count
CLP	Common lymphoid progenitor
CMP	Common myeloid progenitor
DNA	Deoxyribonucleic acid
DVT	Deep venous thrombosis
G	Guanine
GoF	Gain of function
GPCR	G-protein-coupled receptors
GPS	Grey platelet syndrome
GWAS	Genome-wide association studies
EC(s)	Endothelial cell(s)
H3	Histone 3
HEP(s)	Hepatocyte(s)
HES	Hospital episode statistics
HSC	hematopoietic stem cell

HTS	High-throughput sequencing
ICD	International Classification of Diseases
INDEL	Insertion and deletion
Kb	Kilobase(s); i.e. 1,000 bp
LoF	Loss of function
LV	Lentivirus
Mb	Megabase(s); i.e. 1,000,000 bp
MK(s)	Megakaryocyte(s)
MPV	Mean platelet volume
OR	Odds ratio
F3	Tissue factor
F7	Factor VII
pct	Plateletcrit
PE	Pulmonary embolism
PEI	Polyethylenimine
pdw	Platelet distribution width
plt	Platelet count
P/LP	pathogenic and likely pathogenic
pcHi-C	Promoter Capture Hi-C
PPI	Protein-protein interaction
PRS	Polygenic risk score
PT	Prothrombin time
RNA	Ribonucleic acid
RT	Room temperature
SCCS	surface connected canalicular system
SCF	stem cell factor
sd	Standard deviation
SV	Structural variants
SVM	support-vector machine

T	Thymine
TAR	thrombocytopenia with absent radii
TFBS	Transcription factor binding site
TFPI	Tissue factors pathway inhibitors
TFR	Tissue Factor
TG	ThromboGenomics
TPO	Thrombopoietin
UKB	UK Biobank
VEGF	Vascular endothelial grow factor
VEP	Variant effect predictor
VTE	Venous thromboembolism
vWF	von Willebrand Factor
WGS	Whole genome sequencing
WHO	World Health Organization

Table of contents

Table of figures and codes

List of tables

Chapter 1 Introduction

1.1 Blood

Blood is one of the most important tissues in the human body. This has been clear for a long time in human history. Indeed, the suffix *haema*-, which still today describes blood and all its related conditions, comes from the homonym ancient Greek word (Meletis and Goratsa 2002). Without knowing much about its biology, Greek culture associated blood with life, diseases and death ("On the Pathology of the Blood" 1831; Meletis and Konstantopoulos 2010). Indeed, Hippocrates and Galen described blood as one of the four components of humour, which balance with phlegm, black bile, and yellow bile to maintain health in humans (Cos and Hippocrates of Cos 1931; Jouanna 2012).

The humoral theory was abandoned while scientific discoveries in the 19th century shed light on the actual role of blood ("On the Pathology of the Blood" 1831; Blundell 1818). Nowadays, it is clear that blood plays a crucial role in biophysics (e.g. volume, pressure and temperature), chemistry (e.g. pH and osmolarity) and biology of our bodies (Hoffbrand et al. 2016). Because of its central role in human physiology and its relatively easy accessibility, blood has been extensively studied and, therefore, is one of the best-characterised human tissues.

Whole blood is exceptionally heterogeneous. It is composed of a liquid and a cellular component. The former is referred to as plasma. Plasma has several roles: (i) it regulates the osmotic pressure through circulating proteins, such as albumin; (ii) it maintains the correct pH via electrolytes; and (iii) it acts as a carrier for cell nutrients, fatty acids and hormones (Hoffbrand et al. 2008). The blood cellular component comprises several cell types, which based on their ontology can be grouped into two broad lineages: lymphoid and myeloid. The lymphoid lineage is mainly composed of T cells, B cells and natural killer cells. The myeloid lineage is mainly composed of monocytes (and derived macrophages), granulocytes, erythrocytes, and platelets (Fig. 1.1). The main functions of these lineages are: (i) monocytes, granulocytes and lymphoid cells are the main effectors of the host-invasion response; (ii) erythrocytes (or red blood cells) carry oxygen through the circulation; (iii) platelets are responsible for haemostasis by maintaining vascular integrity (see chapter 1.1.2.4).

The quality and quantity of the cells described above is critical for biological processes, such as host-invasion response or wound healing, and the balance between the

liquid phase and the cellular components is maintained during the cell production process, the haematopoiesis.

1.1.1 Haematopoiesis

Haematopoiesis is the differentiation process from which all the mature blood cell types originate. This process is tightly regulated to keep the cells in the correct number and ratio, which is relevant to maintain haemostasis and the other functions. Adult haematopoiesis occurs almost entirely in the bone marrow of the axial skeleton, although instances of extramedullary haematopoiesis have been reported (Lefrançais et al. 2017). Recently, thrombopoiesis, the branch of the haematopoiesis that gives rise to platelets, has been observed in the lungs, making these a novel site with haematopoietic potential (Lefrançais et al. 2017).

All the mature cell types circulating in the blood are derived from a common cell type, referred to as hematopoietic stem cell (HSC; Laurenti and Göttgens 2018). The HSC is a rare quiescent cell type that lies in a specific niche in the bone marrow and undergoes self-renewal upon differentiation (Cheshier et al. 1999; Wilson et al. 2008; Foudi et al. 2009; Sun et al. 2014; Baryawno, Severe, and Scadden 2017). The cell types that are metabolically active and contribute the most to the circulating cells, in terms of numbers, are the early myeloid and lymphoid progenitors (Laurenti and Göttgens 2018). Indeed, these progenitor cells completely restructure their epigenetic state and transcription pattern to accelerate their metabolism, division time and to differentiate towards different lineages in response to external stimuli (Fig. 1.1; Ji et al. 2010; Laurenti et al. 2013; Corces et al. 2016).

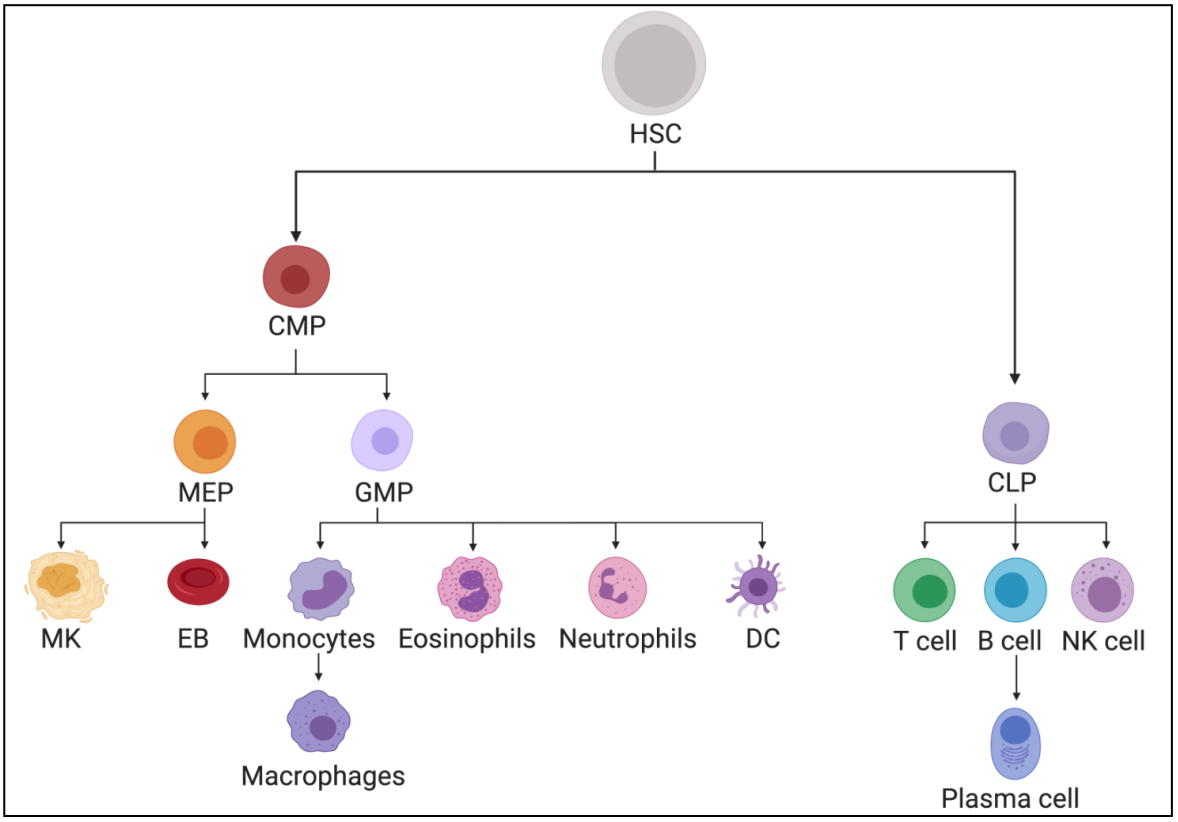


Fig. 1.1 | Schematic representation of a classic haematopoietic tree. *Haematopoietic stem cell (HSC); common myeloid progenitor (CMP); megakaryocyte-erythrocyte progenitor (MEP); granulocyte–monocyte progenitor (GMP); common lymphoid progenitor (CLP).*

Haematopoiesis is often depicted as a lineage tree, with HSCs on top, that progressively differentiate into mature cell types, following the branches of the tree, in a stepwise manner through distinct oligopotent progenitors (Kondo, Weissman, and Akashi 1997; Akashi et al. 2000; Doulatov et al. 2010). In the classic differentiation tree, the first division of an HCS would give rise to myeloid and lymphoid progenitors, respectively common myeloid progenitor (CMP) and common lymphoid progenitor (CLP; Kondo, Weissman, and Akashi 1997; Akashi et al. 2000). The former will contribute mainly to erythrocytes, polymorphonuclear white cells (i.e. granulocytes) and platelets, while the latter to the mononuclear white cells (Laurenti and Göttgens 2018).

Thanks to the development of better assays (i.e. *in vitro* clonal assay and transplant-based tracking assays; Osawa et al. 1996) and the advent of single-cell technologies (Wilson et al. 2015; Grover et al. 2016), the haematopoietic tree has been widely reshaped and debated in the scientific community, with several alternative trees proposed. Nowadays, the consensus is that haematopoiesis is a continuous process and not a differentiation tree with discrete steps and branches (Nestorowa et al. 2016; Notta et al. 2016; Velten et al. 2017; Laurenti and Göttgens 2018; Cheng, Zheng, and Cheng 2020). It

has been proposed that different HSCs have biases (or are primed) towards one or more blood cell fates (Velten et al. 2017; Laurenti and Göttgens 2018; Popescu et al. 2019). For instance, some stem cells are enriched in the protein to differentiate towards the megakaryocyte–platelet lineage component (Sanjuan-Pla et al. 2013), possibly for their central role in human haemostasis. However, if triggered, all the HSC still retain the capacity to differentiate towards several lineages (Velten et al. 2017; Laurenti and Göttgens 2018).

Despite not being linear and unbiased as early models suggested, the haematopoietic differentiation is tightly regulated to ensure that the correct cells are produced in the required amount, for physiological processes and even in response to external stimuli (Paul et al. 2016). In fact, each HSC generates ~10^{11} new cells per day in order to keep the number of blood cells constant and replace dead cells (Orkin and Zon 2008). The half-life of blood cells varies greatly, because of their biology and external stimuli. Neutrophils have one of the shortest lifespans (5.4 days), whilst T cells have one of the longest (years after immunization; Sallusto et al. 1999; Pillay et al. 2010).

1.1.2 Haemostasis

Haemostasis is the physiological process that aims to keep blood flowing through the veins and arteries while maintaining vascular integrity. The process can be summarised in four main tasks: (i) maintaining the blood vessel unobstructed; (ii) repairing damages where they occur; (iii) maintaining blood in a liquid state in the regions surrounding the injury; (iv) removing clots from the wound sites when they are no longer needed (Versteeg et al. 2013). To work properly, haemostasis requires the harmonic effort of cells (i.e. platelets, vascular endothelium and hepatocytes; see paragraph 1.1.2.1) and proteins (i.e. procoagulant and anticoagulant factors; see paragraph 1.1.2.2; Marcdante and Kliegman 2018).

Because of the intricate network of interactions, the redundancy and the large number of factors that are involved, perturbations affecting a haemostatic player do not produce a linear and obvious outcome in the haemostatic process (Hurwitz et al., 2014, Gomez and McVey, 2015; Link et al., 2020). Haemostasis is divided into two components: primary and secondary. Primary haemostasis refers to all the processes that are involved in platelet clot formation (see chapter 1.1.2.4). Secondary haemostasis refers to all the enzymatic digestions and zymogen activations that occur in the coagulation cascade (i.e. intrinsic, extrinsic and common cascade, see chapter 1.1.2.3; Gale 2011; Chaudhry, Usama, and Babiker 2020).

1.1.2.1 Cells involved in haemostasis

There are many cell types involved in haemostasis. These have different embryonic origins and derive from the mesoderm and endoderm (Oberlin et al. 2002; Pansky 1982; Oberlin et al. 2002). The mesoderm-derived cells grow and differentiate into a common progenitor (i.e. hemangioblast) and generate the vascular endothelium and the hematopoietic portion of the bone marrow (Hoffbrand et al. 2016). Endothelial cells are the most relevant haemostatic cell type of the vascular endothelium, whilst the haematopoietic bone marrow contributes to the haemostasis through platelet production (Fig. 1.1). The endoderm-derived cells form many organs including the liver, which is responsible for the production of the majority of coagulation factors (Hoffbrand et al. 2016).

Endothelial cells

Endothelial cells (ECs) inlay the entire vascular system and are involved in several biological processes, directly or indirectly related to haemostasis. For instance, they regulate vascular tone and inflammation and mitigate the communications between the bloodstream and the connective tissues below the vascular endothelium (Chesterman 1988; Krüger-Genge et al. 2019). The vascular endothelium is a monolayer of cells, these are 25-50 μm long and 10-15 μm wide, and their morphology is commonly referred to as cobblestone (Haudenschild 1984). In spite of their morphological similarities, ECs are a very heterogeneous population (Khan et al. 2019; Nakato et al. 2019) and their identity is conferred by blood mechano- and chemical-stimuli, namely shear stress and cytokines (Furie and Furie 2008).

ECs mechanically contribute to the haemostasis via the interplay with the vascular smooth muscle to control vasodilation and vasoconstriction and, therefore, regulate the local blood pressure (Michiels 2003). Two molecules are the main vasodilators: nitric oxide and prostacyclin (Palmer, Ferrige, and Moncada 1987). They are both constitutively synthesized in the endothelium and released upon EC stimulation via angiotensin, histamine, thrombin (*F2*) and others (Fleming, Bauersachs, and Busse 1996; Moncada, Higgs, and Vane 1977). These molecules prevent platelet adhesion to ECs, an event that is one of the triggers of coagulation and inflammation (Moncada et al. 1977; Gibbins 2004). The opposite mechanism, vasoconstriction, is controlled by other molecules, such as angiotensin II or prostaglandin H2 (Chien 2008; Garland and Dora 2017). In the case of blood vessel injuries,

ECs reduce vascular lumen to limit the haemorrhage (Ho-Tin-Noé, Boulaftali, and Camerer 2018; Doherty and Kelley 2020).

In normal conditions, ECs promote the activity of several anticoagulants, such as protein C and protein S (see chapter 1.1.2.3; Sadler 1997). In response to vascular damages, on the other hand, ECs release proteins and molecules that trigger the activation of platelets and initiate the wound repair mechanisms (Krüger-Genge et al. 2019). This interplay between bloodstream and ECs acquires the distinct definition of thromboregulation (Marcus et al. 2001). Examples of procoagulant proteins are von Willebrand Factor *(VWF) and* P-selectin *(SELP)*. Amongst others, these proteins are synthesized by EC and stored in the Weibel-Palade bodies, granules specific to the EC that are released in response to blood vessel damages (Lichtman and Williams, 2006).

Thromboregulatory reactions are crucial to balance haemostasis (Table 1.1): on the one hand, ECs have to orchestrate clot formation and platelets adhesion, and on the other hand, platelets inhibition and fibrinolysis (Lichtman and Williams, 2006). Indeed, one of the primary purposes of a healthy vascular endothelium is preventing unwanted thrombus formation.

Anti-Thrombotic events or reactions	**Prothrombotic events or reactions**
Biosynthesis of prostacyclin	Expression of P-selectin
Secretion of Nitric oxide	Degranulation of Weibel-Palade bodies
Inhibition of platelets by CD39	Release of von Willebrand Factor
Synthesis of endothelin	
Production of fibrinolytic proteins (e.g. Protein C)	

Table 1.1 | List of procoagulants and anticoagulants events carried out by vascular endothelium. Adapted from Williams Hematology 8th edition.

Platelets

Platelets are a unique cell type. Evolutionarily, they may be derived from a shared progenitor with white blood cells, as it is still the case in lower organisms. For instance, hemocytes in the horseshoe crab have a shared role in immunity and coagulation (Levin 1977; Morrell et al. 2014; Hoffbrand et al. 2016). In humans, platelets have roles in inflammation, atherosclerosis, tumour growth, metastasis and angiogenesis, but their primary role is in haemostasis (Bertozzi, Hess, and Kahn 2010; Watson, Herbert, and Pollitt 2010; Feng, Madajka, and Kerr 2011; Nurden 2011). They are highly specialised cells that

participate in wound healing, mainly in two ways: by forming platelet plugs and by enhancing prothrombin activation (de Gaetano 2001).

Platelets are produced via a tightly regulated process (mainly by thrombopoietin, *THPO*) by megakaryocytes (Fig. 1.1; Wendling et al. 1994). Megakaryocytes produce thousands of proplatelets released into the bloodstream as platelets (2,000-3,000 platelets per megakaryocyte; the normal range of platelet count 150-450x10^9 cells/L; Hoffbrand et al. 2016). They are ten times smaller than an average cell and have a discoid shape, with a radius of ~1.5 μm and a mean volume (mpv) of ~9 fL (Hoffbrand et al. 2016). Experimental evidence correlates mpv to platelet granule release capacity (Karpatkin 1978; Haver and Gear 1981; Martin et al. 1983; Mangalpally et al. 2010; Gieger et al. 2011). Under normal circumstances, the platelet count (plt) and mpv are inversely correlated, with a correlation coefficient of -0.5 (Levin and Bessman 1983; Astle et al. 2016). When the platelet count decreases, the bone marrow tries to compensate and releases new platelets, which are still reticulated and have, on average, larger volumes (Levin and Bessman 1983).

Platelet internal structure is fundamental to their functions and it is the subject of extensive studies. The most obvious structural feature is that platelets, which are megakaryocyte cytoplasmic fragments, do not have nuclei (Melchinger et al. 2019). This characteristic allows them to pass through the narrowest capillaries alongside red cells (Hoffbrand et al. 2016). However, the lack of a nucleus limits their ability to transcribe new mRNA (see chapter 1.3 for mRNA definition) and, consequently, constraints platelet half-life to 5-8 days (Levin and Bessman 1983; Italiano et al. 2021). This does not mean that platelets are molecularly inactive. In fact, compelling evidence shows significant RNA-splicing and translation activities in platelets that have been activated by agonists like *F2* (Weyrich et al. 1998; Schwertz et al. 2006). Platelets also have an extensive network of surface membrane invaginations, known as the canalicular system (Escolar and White 1991), linked to the cytoskeleton and required to expand their volume during activation. The main components of the platelet cytoskeleton are tubulin (*TUBB1*), actin filaments (*ACTN1*) and spectrin (*SPTBN1*) proteins (Raslova et al. 2007; Rendu 2011).

Platelet cytoplasm is also characterised by secretory granules, namely lysosomes, dense granules (~7 granules per platelet) and α granules (~80 granules per platelet; Italiano et al. 2021). The main role of the latter two types of granules is to regulate thrombus formation, however, they are also involved in other biological processes, like wound healing and immunity (Koupenova and Freedman 2020). Dense granules contain haemostatically active small molecules, such as serotonin, ADP, ATP and Calcium (as Ca^{2+} cation;

Youssefian and Cramer 2000). α granules contain hundreds of proteins that are mainly synthesised by megakaryocytes, such the cytokine platelet factor 4 (*PF4*), thrombospondin 1 (*THBS1*), vascular endothelial cell growth factor (*VEGF*), *VWF* and fibrinogen (*FGA-FGB-FGG;* Heijnen et al. 1998). These granules originate as protrusion of the trans-Golgi network and have multivesicular bodies as intermediate stages (Podolak-Dawidziak et al. 1995; Heijnen et al. 1998; Bariana et al. 2017).

Platelets play an important and fundamental role in the aetiology of cardiovascular diseases (CVD), such as myocardial infarction, thrombotic stroke and peripheral artery disease (Willoughby et al. 2002). Hence the production of these cells by megakaryocytes has been extensively studied at the biophysics and cell biology levels. plt is the main parameter that is being used clinically; the other platelet physical features, such as mpv, platelet distribution width (pdw) and plateletcrit (pct), are typically limited to research studies (Italiano et al. 2021; Sivapalaratnam et al. 2021). pdw is an indicator of the platelet volume variability and pct is the percentage of blood volume occupied by platelets (Budak et al. 2016).

Platelets circulate in the bloodstream in an inactive state and are maintained so via nitric oxide and prostacyclin released by the ECs (see chapter 1.1.2.1). Indeed, platelets should only be activated in response to traumatic vascular injury or a ruptured atherosclerotic plaque (Badimon et al. 2012). However, some circumstances (e.g. progressive atherosclerosis) may activate platelets erroneously at sites of disturbed arterial blood flow, like arterial bifurcation (Ruggeri 2009; Massai et al. 2012).

Hepatocytes

Liver is the largest gland of the human body, and it has more than 500 different roles, amongst which there is detoxification, digestion metabolism and immunity (Boyer et al. 2012). Hepatocytes (HEPs) constitute about 80% of the liver mass (Zhou et al. 2016) and are the most important parenchymal cells (Vekemans and Braet 2005; Stanger 2015). The remaining 20% volume is composed mainly of biliary epithelial cells and a few others, such as Kupffer cells, which play a central role in the immune response (Stanger 2015).

HEPs are large polygonal-shaped cells, with a diameter of ~25 μm and a volume of ~5,000 μm^3. In homeostatic conditions, HEPs are quiescent cells. Generally, less than 2% of them go through mitosis (MacDonald 1961), however, upon injuries, they have the ability to re-enter the cell cycle and renew the damaged tissue (Sawada and Ishikawa 1988). If

needed, each HEP is able to undergo hundreds of cellular divisions (Overturf et al. 1997). These cells are tightly adherent to the surrounding ones and their cellular membranes are highly polarised (Burt, Ferrell, and Hubscher 2017). Cell polarisation and adhesion are fundamental to create liver structures such as sinusoids and lobules (Sellaro 2007; Brunt et al. 2014).

One of the main roles of HEPs is the synthesis of the proteins that circulate in the bloodstream (Feldmann et al. 1972; Roberts, Patel, and Arya 2010; Zhou et al. 2016). Indeed, HEPs have a central role in primary and secondary haemostasis and produce the majority of the procoagulant and anticoagulant factors (Table 1.2; Roberts, Patel, and Arya 2010). Moreover, HEPs not only synthesise most of the coagulation factors but are also the main site for the synthesis of TPO, the key growth factor for differentiating haematopoietic stem cells into megakaryocytes (Jelkmann 2001; Rios et al. 2005). The binding of old platelets to the Ashwell-Morell receptor induces the transcription of the *THPO* gene, thereby providing a direct regulatory mechanism to maintain a stable platelet count (Grozovsky et al. 2015).

Pro-coagulant proteins	**Anticoagulant proteins**
Fibrinogen	Antithrombin
FII,FV, FVII, FVIII, FIX, FX, FXI, FXII	Tissue factors pathway inhibitors
PAI-1	Plasminogen
α2-antiplasmin	Protein C
	Protein S

Table 1.2 | List of proteins synthesised in the hepatocytes that have a direct role in haemostasis. Adapted from Roberts, Patel, and Arya 2010.

Notwithstanding the liver's exceptional self-regeneration capacity, several conditions can compromise HEPs function (Michalopoulos and Bhushan 2021). Chronic hepatitis and cirrhosis are often linked to liver failure and the cause of impaired haemostasis (Hillman et al. 2005; Tripodi and Mannucci 2011). Particularly, advanced alcoholic liver cirrhosis can lead to combinations of thrombocytopenia (i.e. low platelet count) and reduced levels of clotting factors (Mammen et al. 1992; Trotter et al. 2006), with portal hypertension and bronchial varices resulting in a severely increased risk of haemoptysis (Youssef et al. 1994).

1.1.2.2 Proteins involved in haemostasis

Haemostasis is a complex and intricate process, in which not only cells have to maintain their function in tight control, but also proteins need to conserve the correct stoichiometry across all the elements of the process. Haemostasis is controlled by (i) proteins that start the coagulation, (ii) proteins that accelerate the activation cascade and (iii) proteins that inhibit coagulation and remove the platelet plug. Interestingly, some of these proteins and their functions might have evolved from a few ancestral proteins. Coagulation protein families show a high degree of similarity, sharing several domains. For instance, the serine protease domain is present in several coagulation factors, namely *F2*, factor VIII, IX, X, XI (*F8*, *F9*, *F10*, *F11* respectively), protein C (*PROC*), prekallikrein (*KLKB1*), prourokinase (*PLAU*) and plasminogen (*PLG*; Hoffbrand et al. 2016). A list of the major circulating coagulation proteins is reported in Table 1.3 and their function is described in the following paragraphs.

Protein name	**Gene name**	**Cell**	**Main role**	**pLI**
Prothrombin	*F2*	HEP	Create fibrinogens clots	0
Tissue Factor	*F3*	HEP	Cofactor for active factor VII	0
Factor V	*F5*	HEP	Cofactor in tenase	0
Factor VII	*F7*	HEP	Activates factors IX and X	0
Factor VIII	*F8*	HEP, MK	Cofactor for factor IX	1
Factor IX	*F9*	HEP	Activates factor X	1
Factor X	*F10*	HEP	Activates prothrombin	0
Factor XI	*F11*	HEP	Activates factor IX	0
Prekallikrein	*KLKB1*	HEP	Start intrinsic pathways	0
Fibrinogen (α - chain)	*FGA*	HEP	Stabilise the clot	0
Fibrinogen (β - chain)	*FGB*	HEP	Stabilise the clot	0.57
Fibrinogen (γ - chain)	*FGG*	HEP	Stabilise the clot	0.05
Von Willebrand factor	*VWF*	EC/MK	Platelet adhesion	0
Thrombomodulin	*THBD*	EC/MK	Cofactor in protein C activation	0.05
Protein C	*PROC*	HEP	Inactivation of factor V and VIII	0
Protein S	*PROS1*	HEP/MK/EC	Inactivation of factor V and VIII	0
Tissue factor pathway inhibitor	*TFPI*	HEP/MK/EC	Stop coagulation initiation	0.25

Antithrombin	*SERPINC1*	HEP	Cleave pro-coagulant proteins	1
Plasminogen	*PLG*	HEP	Digest clot after wound repair	0.01
Tissue plasminogen activator	*PLAT*	EC	Activate plasminogen	0
Prourokinase	*PLAU*	EC	Activate plasminogen	0
α_2-Antiplasmin	*SERPINF2*	HEP	Inhibits plasmin	0.01

Table 1.3 | The major circulating coagulation proteins. The gene symbols are reported according to their HGNC name. Cell expression is limited to HEP, MK, EC and derived from the expression profiles of the BLUEPRINT and GTEx (Chen et al. 2016; GTEx Consortium 2020). pLI (probability of loss of function intolerance; see chapter 1.3.6) is a measure of conservation of the tolerance to loss of function variants (0 = tolerated; 1 = not tolerated).

1.1.2.3 Coagulation processes, the secondary haemostasis mechanisms

The coagulation process, also referred to as the coagulation cascade, describes all the enzymatic reactions that happen in the secondary haemostasis (for a description of the primary haemostasis see chapter 1.1.2.4). These catalytic activations are required to convert blood from a liquid to a gel state in order to contain bleeds.

Two independent laboratories devised the landmark models of human coagulation in the 1960s (Macfarlane 1964; Davie and Ratnoff 1964). These models are still commonly cited; however, throughout the years they have been complemented with further details and more complex models have been proposed (Hoffman and Monroe 2001; Monroe and Hoffman 2006; Mackman 2009). Refinements to the models try to reflect more accurately what happens *in vivo,* redefining the function of some enzymes and acknowledging the interplay between extrinsic and intrinsic pathways (described below; Repke et al. 1990; Gailani and Broze 1991).

Extrinsic pathway

The extrinsic pathway corresponds to the initiation phase of the coagulation process (Fig. 1.2). When the vascular endothelium is damaged, by mechanical or chemical injuries, adventitial cells - i.e. smooth muscle cells in the tunica media (for veins and arteries) or fibroblasts (for capillaries) - become exposed to the bloodstream and reveal the tissue factor (*F3*) to the bloodstream. *F3* is one of the most important proteins in the coagulation cascade and it is ubiquitously expressed in all the cell types of the human body, with the exception of blood cells and ECs surrounding the vascular epithelium (e.g. vascular adventitia; Grover

and Mackman 2018). This widespread expression ensures that damages to the vascular endothelium are immediately detected and repaired.

Once exposed, *F3* binds to the circulating zymogen factor VII (*F7*) and converts it to its active form. The *F3-F7* complex is then able to activate *F10*, which in response can cleave and activate a minimum amount of prothrombin (the inactive form of *F2*; Fig. 1.2). The small amount of active *F2* cuts and activates factor V (*F5*), *F8* and *F9* and these begin a positive feedback loop, commonly referred to as the amplification phase. The increase of the *F2* activation, required to the cleavage of the fibrinogen and therefore the formation of the fibrin clot, happens after the formation of the *F5-F10* complex, also known as tenase (Gailani and Renné 2007; Butenas, Orfeo, and Mann 2009; Grover and Mackman 2018). Clinical tests, such as prothrombin time (PT), are based on the measure of the activity of *F2*, *F7*, and *F10*, and assess the function of the extrinsic pathway (Chaudhry, Usama, and Babiker 2020).

Intrinsic pathway

The role of the intrinsic pathway was thought to be only that of amplifying the effect of the extrinsic coagulation cascade. This theory was supported by the fact that the concentration of the coagulation factors involved increases descending the coagulation cascade (Fig. 1.2) thus suggesting a ‘domino effect’ amplification (Baird, Clancy, and McVicar 2005; Tormoen et al. 2013; Chaudhry, Usama, and Babiker 2020). However, experiments in mouse models have altered the idea that the intrinsic pathway is only secondary, as they have identified physiological triggers for its direct activation which include collagen, polyphosphates, and neutrophil extracellular traps (NETs; Versteeg et al. 2013). Briefly, in response to stimuli such as the release of PRCP in the bloodstream, factor 12 (*F12*) and *KLKB1* initiate a series of enzymatic reactions that cleave and activate themselves and also digest high-molecular-weight kininogen (HMWK, Fig. 1.2; Waldmann et al. 1975; Hillman et al. 2005; Ivanov et al. 2020). The products of these reactions activate *F11* which in response cleaves *F9* and *F8* (Ngo et al. 2008; Gardner and Davies 2009). Finally, *F8* cleaves *F10* and generates the intrinsic-pathway tenase that cleaves prothrombin (Fig. 1.2; Hoffbrand, Catovsky, and Tuddenham 2008). The intrinsic cascade initiation is quite redundant, for instance, *F12* deficiency does not compromise the initiation of the cascade (Chaudhry, Usama, and Babiker 2020). The intrinsic pathway activity is clinically measured through the activated partial thromboplastin time (APTT) test (Chaudhry, Usama, and Babiker 2020).

Common pathway

Both coagulation cascade pathways (i.e. extrinsic and intrinsic) terminate with a tenase complex, which cleaves and activates *F10* (Fig. 1.2). The extrinsic pathway tenase complex is composed of *F7-F3* while the intrinsic pathway tenase complex is *F8-F9* (Undas, Brummel-Ziedins, and Mann 2005). This redundant and convergent evolution of the pathways stresses the central role of *F10* in secondary haemostasis. Calcium, in its bivalent cationic form (i.e. Ca^{2+}), is required for all the catalytic steps in the intrinsic and extrinsic pathways. Also, phospholipids are important for both tenase complexes and they are offered by ECs and activated platelets (Hoffbrand et al. 2016; Chaudhry, Usama, and Babiker 2020). Phospholipid surfaces host the tenase complexes and this physical constraint is necessary to limit the coagulation cascade to the relevant site of injury (Hoffbrand et al. 2016).

Once activated, *F10* binds to its cofactor, *F5*, in order to cleave prothrombin into its active form, *F2. F2*, in response, cleaves soluble fibrinogen into fibrin (*FGA-FGB-FGG*). *FGA-FGB-FGG* is an insoluble fibrous protein, which polymerises and creates a mechanical frame to clots at the wound site, closing the blood vessel injury (Lichtman and Williams, 2006). Fundamental for wound closure is *F8*, which crosslinks fibrin fibres to make them more stable (Hoffbrand et al. 2011). The major players of the common pathway are *FGA-FGB-FGG*, *F2*, *F5*, *F8* and *F10* (Chaudhry, Usama, and Babiker 2020).

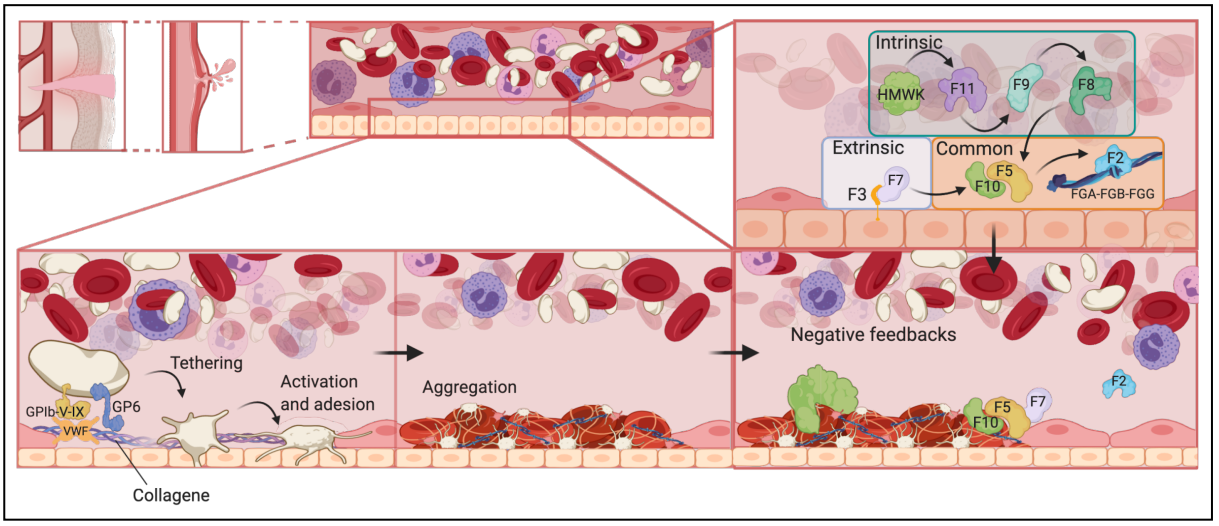


Fig. 1.2 | Schematic representation of the haemostasis. *The top right panel have a brief representation of the coagulation cascade. Intrinsic, extrinsic and common. The three bottom panels depict the primary haemostasis and all the steps involved in the platelet plug formation.*

Inhibition of the pro-coagulant stimuli

Blood needs to be kept in its liquid form aside from the injury sites. The process of thrombus formation needs to remain localised and resolved in a timely manner once

vascular integrity has been restored. For this purpose, the negative feedback of haemostasis (also named thrombolysis) is essential to prevent the unnecessary formation of thrombi and to limit the excessive growth of a thrombus. The control and termination of the coagulation cascade have two main mechanisms: the inhibition of the serine-proteases via their plasma-residing inhibitors and the lysis of activated coagulation factors, like fibrinogen.

There are 2 types of inhibitors involved in the resolution of the coagulation cascade: serine-protease inhibitors (serpins) and Kunitz-type ones (Neuenschwander 2006; Rein et al. 2011). Serpins are a family of at least seven irreversible inhibitors that circulate in the plasma. They function as substrates for the serine-proteases and, once cleaved, they irreversibly bind to the serine-protease inactivating that molecule. Antithrombin (*SERPINC1*) is the best-characterised serpin and it modulates the coagulation complex with *F3* and *F10* and its main role is the inhibition of *F2* (Lomas et al. 1992; Huntington et al. 2000; Quinsey et al. 2004; Patnaik and Moll 2008). Some serine proteases involved in the coagulation cascade are also able to cleave cell receptors involved in haemostasis. As an example, the Protease-Activated Receptors (PARs) are a family of proteins that are expressed on platelets and vascular ECs. The role of PARs is to restore the correct physiology after clot formation. For instance, to repristinate the vascular endothelium (i.e. angiogenesis) through the growth of ECs (Versteeg et al. 2013). The existence of PARs is one example that underlines that not only circulating coagulation factors but also the interplay with cells is crucially important for maintaining haemostasis (see chapter 1.1.2.4).

Kunitz-type inhibitors are quenching coagulation cascades via reversible binding to the coagulation factors (Antuch et al. 1994). The most important serine-proteases Kunitz-type is the Tissue Factor Pathway Inhibitor (*TFPI*), which is present mainly on EC membranes and within platelet α granules (Broze and Girard 2012; Mast 2016). *TFPI* mainly inhibits *F10* and *F3-F7*, forming a quaternary structure and blocking further activation of the coagulation cascade (Hoffbrand, Catovsky, and Tuddenham 2008).

F2 not only activates pro-coagulant proteins (e.g. fibrinogen), but it also starts the inhibition of the coagulation (Dahlbäck and Villoutreix 2005). For instance, its binding to *PROC* initiates the protein C pathway. *PROC* is a vitamin K-dependent serine protease that circulates as a zymogen in the plasma and uses protein S (*PROS*) as a cofactor. When activated, *PROC* digests *F5* and *F8* reducing the pro-coagulant stimuli (Dahlbäck and Villoutreix 2005).

Fibrinolysis

The last step in the haemostasis is known as fibrinolysis and it consists of the elimination of the fibrin clots by several different proteases (Chapin and Hajjar 2015). When fibrin binds to the damaged tissues, it activates tPA (*PLAT*) and uPA (*PLAU*; Chapin and Hajjar 2015). These two proteins compete with the circulating *PLG* activator inhibitor (PAI)-1 and, eventually, lead to the activation of *PLG* (Forsgren et al. 1987). *PLG* breaks the fibrin clots into soluble units which are normally referred to as D-dimers (Khalafallah et al. 2014). The level of D-dimers is of considerable diagnostic value in determining whether there is inappropriate thrombus formation (see chapter 1.2.2; Hulle et al. 2013). Ultimately, also *PLG* has a negative regulation that is mainly carried out by α2-antiplasmin (Chapin and Hajjar 2015).

1.1.2.4 Primary haemostasis

Thrombus formation is the result of platelet adhesion and aggregation. This process is triggered by molecules such as collagen and *F2* (Furie and Furie 2008) and has slightly different pathways in response to the blood shear stress (Hoffbrand et al. 2016). Vasculature damages at medium-high flow rates (i.e. 1000-5000/s; arterial pressure) trigger platelet activation mainly via the *VWF*-collagen complex which is bound by the GPIb–IX–V complex. Whilst, at lower flow rates (i.e. < 500/s; venous blood flow), platelets can adhere directly to the subendothelial matrix, without the intermediated help of *VWF* and collagen (Hoffbrand et al. 2016).

After initial adhesion, platelets roll over the damaged area in an inactive state (Vanhoorelbeke et al. 2003). This process, which is still reversible, is normally referred to as tethering, because of the tethers formed by the adhering platelet membranes. The initial anchoring, driven by *VWF*-GPIb-IX-V complex, is strengthened by GPIIb-IIIa and *GP6* binding to collagen and *F2*, with the cooperation of other integrins as well (López 1994; Nieswandt and Watson 2003). The events following the tethering steps commit the platelets to their activation, inducing changes in morphology and releasing their granules (Blair and Flaumenhaft 2009; Hoffbrand et al. 2016).

Indeed, after activation, platelets undergo an internal cytoskeleton rearrangement, mainly because of actin polymerization, which allows the spreading of the cell membrane and the formation of filopodia and lamellipodia (Fox 1996; Albuschies and Vogel 2013). These morphological rearrangements are necessary for platelet function and facilitate the

adhesion of other cells on the injury site (Bearer, Prakash, and Li 2002). $Ca2^{+}$ is a key molecule in the activation of the platelets because it helps the formation of actin fibres and it also works as a cofactor for a series of proteins that induce granules release (Bearer, Prakash, and Li 2002). The release of granules is mainly mediated by tyrosine kinases and G-protein-coupled receptors (GPCRs; Shattil, Kashiwagi, and Pampori 1998). These proteins induce a cascade of signals that transduce in α granules and dense granules release, which helps in orchestrating the haemostasis.

α granules contain proteins that regulate thrombus formation (e.g. *FGA-FGB-FGG*, *VWF*, *TFPI*, *PF4* and *PLG*), sustain vascular endothelium integrity (e.g. *VEGF*) and stimulate inflammation (e.g. β-thromboglobulin; King and Reed 2002; Kaushansky et al. 2015). Some α granules proteins have a dual role that links haemostasis to immune response. For instance, *SELP*, which is transferred from the granules to the platelet membrane upon activation, is able to interact with endothelial cells and leukocytes (Michelson et al. 2019).

Dense granules take their name because of how they appear in electron microscopy images, being highly osmophilic due to the high concentration of adenosine triphosphate (ATP) and adenosine diphosphate (ADP; McNicol and Israels 1999). The ADP, released after initial platelet activation, acts as a positive amplification signal to induce the activation of more platelets. The positive feedback is mediated by binding to GPCRs, namely *P2RY1* and *P2RY12* (Michelson et al. 2019).

Ultimately, the final process in thrombus formation is plug stabilisation (Hawiger 1987). In order to resist the blood shear stress, *F8* crosslinks fibrin fibres to the just-formed platelet plug, stressing one last time the interplay between primary and secondary haemostasis (Hawiger 1987).

1.2 Bleeding, thrombotic and platelet disorders

Inherited defects of haemostasis comprise a heterogeneous group of rare diseases, referred to, in this thesis, as inherited bleeding, thrombotic and platelet disorders or BTPDs in short. The BTPD symptomatology ranges from severe to mild and in some cases symptoms may be entirely absent, limited only to altered FBC (e.g. mild thrombocytopenia). For the purpose of this thesis, BTPDs are defined as a group of rare inherited conditions caused by pathogenic variants in a set of 93 diagnostic-grade genes, which have been curated by the Scientific and Standardization Committee for Genomics in Thrombosis and

Haemostasis for the International Society on Thrombosis and Haemostasis (Table 1.4; Megy et al. 2019).

Inherited BTPDs tend to have earlier manifestations in life and, most of the time, there is a family history (Lentaigne at al. 2016; Sivapalaratnam et al. 2017), although sometimes these can arise due to *de novo* mutations. For instance, 30% of haemophilia A cases have *de novo* variants (Castaldo et al. 2007; Rodriquez-Merchan and Lee 2008). The aetiology of inherited BTPDs resides in germline mutations in one, or more, genetic regions that are involved in the hemostasis process (see also chapter 1.3.7; Westbury et al., 2015; Simeoni et al., 2016; Downes et al., 2019). BTPDs can affect only one tissue (i.e. non-syndromic) or several tissues (i.e. syndromic), such as the case of *MYH9*-related diseases which are characterised by macrothrombocytopaenia, kidney insufficiency and hearing loss (Althaus and Greinacher 2009).

Accurate estimates on the prevalence of rare inherited BTPDs are lacking, but it is estimated to range between 1 in 5,000 and 1 in 2,000,000 (Peyvandi et al. 2009; Castaman and Linari 2017; UKHCDO annual report 2018). Taken together, all the inherited forms of BTPDs affect 1 in 3,500 people (Simeoni et al. 2016) and nearly 25,000 individuals are registered on the United Kingdom Haemophilia Centre Doctors' Organisation (UKHCDO) database. Given their broad range of aetiologies and different manifestations, it is difficult to make a reliable estimate of their frequencies. Indeed, milder forms do not require medical attention and therefore they are very hard to recognise. Notwithstanding the lack of accurate estimates for their prevalence, it is important to reach a molecular diagnosis early in life because it informs the choice of treatment, aids prognostication and can be used in family planning (Sivapalaratnam et al. 2021).

The World Health Organization (WHO) has standardised the disease reports and health conditions with unique identifiers, namely the International Classification of Disease (ICD) codes. This categorisation is helpful to estimate the global burden of disease and inform the delivery of integrated health care (see Table 1.5 for the ICD codes relative to BTPDs, Table 2.2 for the BTPD symptoms and chapter 1.5.2 for a more detailed description of ICD codes). However, the ICD codes are not granular enough to encode the wide range of symptoms observed in patients with inherited BTPDs. An alternative ontology, named the Human Phenotype Ontology (HPO) has been introduced to support studies in patients with inherited disorders (see chapter 1.5).

1.2.1 Bleeding disorders

The absence or lack of activity of procoagulant proteins can cause pathological bleeding (Table 1.4; Megy et al. 2019). Amongst the several types of inherited bleeding disorders, haemophilia A, haemophilia B and von Willebrand disease (VWD) are the most common and well characterised. According to the UKHCDO annual report for 2018, the UK has 8,159 and 1,795 haemophilia A and B patients respectively (occurrence 1 in 8,000 and 1 in 37,000 in the whole UK population) and 10,798 VWD patients (occurrence 1 in 6,000). Rarer bleeding disorders, for example, *FVII* and *FXIII* deficiencies have frequencies between 1 in 500,000 for and 1 in 2 million, respectively (see Table 1.4; Palla et al. 2015; Hoffbrand et al. 2016; UKHCDO annual report 2018; Peyvandi et al. 2009; Castaman and Linari 2017). There is a diagnostic bias towards females because of the medical attention needed to treat iron-deficiency anaemia caused by heavy menstrual bleeding (Hoffbrand et al. 2016).

Bleeding phenotypes caused by secondary haemostasis problems vary greatly in their severity, from almost undetectable to life-threatening. The most common symptoms of coagulation factor deficiency are mucosal bleeding, post-trauma or post-surgery bleeding and central nervous system bleeding. Death caused by bleeding disorder *per se* is rare, occurring in 1 in 10,000 patients, mainly because of intracranial haemorrhage (Darby et al. 2007; Cavazza et al. 2016). More frequent are deaths due to HIV and Hepatitis C comorbidities caused by the past use of blood-derived factors, which increases the death rate to 1 in 100 patients per year (Diamondstone et al. 2002; Darby et al. 2004; *UKHCDO Annual Report 2019, accessed in September 2021*).

1.2.2 Thrombosis

Inherited conditions with an increased risk of blood clots in the venous circulation are referred to as thrombosis or thrombophilia (Sexton and Smyth 2014). The two main conditions are deep vein thrombosis (DVT), usually occurring in the lower limbs, and pulmonary embolism (PE). These two are often grouped together as venous thromboembolism (VTE; Goldhaber and Morrison 2002). According to the Global Burden of Diseases, Injuries, and Risk Factors Study 2010, thrombosis causes 25% of all deaths worldwide and VTE itself caused 500,000 deaths in Europe in 2004 (Raskob and ISTH Steering Committee for World Thrombosis Day report 2014; Thrombosis UK, accessed in September 2021). Approximately 60% of the VTE events occur during or after hospitalization (Thrombosis UK, accessed in September 2021) and females are more susceptible because

of the use of oral contraceptives containing oestrogen and pregnancy, which increase the risk of VTE of 3-6 and 5-10 fold respectively (Gialeraki et al. 2018).

Thrombophilia is a complex disorder (see chapter 1.3.1 for a definition of complex disorders) that presents late-onset, in which environment, genetics and lifestyle act synergistically. However, inherited thrombophilia are a group of early-onset disorders with loss-of-function (LoF; see chapter 1.3.6) variants in anticoagulant and antithrombin proteins (e.g. *PROC, PROS1, SERPINC1*) and gain-of-function (GoF) variants in procoagulant factors (e.g. *F5* and *F2*; Franco et al. 2001; Hoffbrand et al. 2016). Possibly, the best characterised genetic risk factor is Factor V Leiden variant (FV.R506Q), which is a relatively common allele that in most studies increases the risk of VTE up to three folds (Klarin et al. 2019).

1.2.3 Platelet disorders

Inherited platelet disorders are rare, with occurrences that are between 1 in 200,000 and 1 in 1,000,000 newborns (Peyvandi et al. 2012). They can be quantitative or qualitative and they can have a series of causes, from bone marrow failure (Fanconi anaemia) to autoimmune reactions (Key et al. 2017). Quantitative platelet disorders affect platelet counts, reducing or increasing their concentration in the bloodstream. Qualitative platelet defects affect their ability to correctly respond to stimuli and contribute to primary haemostasis.

The standard reference parameter for platelet count is 150-450x10^3/µl, with differences in the distribution between genders, and females having on average a higher platelet count (Collins et al. 2021). However, many people with a platelet count below 150x10^3/µl do not have any bleeding events unless seriously challenged (Key et al. 2017; Izak and Bussel 2014). Only If the platelet concentration goes below 20x10^3/µl, bleeding may occur without any injury and a platelet transfusion is required (Key et al. 2017; Izak and Bussel 2014). Low platelet count (i.e. thrombocytopenia) can occur alone or in combination with alteration of other platelet features (e.g. size and volume). An example of reduced platelet count with a normal size is the thrombocytopenia with absent radius syndrome (TAR; mutated gene *RBM8A*), a recessive disease linked to hemorrhagic manifestations that can also be fatal. Wiskott-Aldrich syndrome (WAS; mutated gene *WAS*) is an example of platelet disease where size and number are both reduced; while *MYH9*-related and Bernard-Soulier syndromes (BSS; mutated genes *GP9*, *GP1BA* and *GP1BB*) are both linked to reduced platelet count but increased platelet volume (i.e. macrothrombocytopenia). On the other side of the spectrum, if too many platelets are produced in the bone marrow, there is an even

rarer condition called thrombocythemia, which has a broad range of symptoms ranging from blood clots to excessive bleeding. Most of the time, this is caused by somatic variants, but variants in *MPL* or *THPO* have been reported to cause inherited forms of familial thrombocythemia (Tefferi et al. 2007; Wilkins et al. 2008).

Thrombocytopathy is the general definition for qualitative platelet defects; these defects can be in adhesion, signal transduction or aggregation. BSS is also characterised by reduced platelet agglutination, because of defects in the binding of VWF and delayed response to *F2* (Hoffbrand et al. 2016). Problems in the signal transduction have a less well-characterised pathophysiology, but usually are grouped under an umbrella term, platelet signalling disorder, because of their nature, most of the time due to defects in one or more of the receptors coupled to G-proteins. Lastly, Glanzmann's thrombasthenia (GT) is an example of reduced aggregation (Hoffbrand et al. 2016). People affected by GT have a normal platelet count, but these platelets do not respond to aggregation stimuli such as ADP, collagen or *F2*.

With the exception of severe cases that require platelet transfusions, most of these conditions do not require extensive medical attention and an adapted lifestyle can prevent most of the symptoms. For example, trauma avoidance and appropriate dental care to prevent gingival bleeding. However, it is convenient to know of their existence because they may require medical attention if coupled with other conditions or habits. For instance, pregnancy reduces platelet count (Reese et al. 2018), while smoking is discouraged in people that already have high platelet count (Levine 1973; Mundal et al. 1998; Ghahremanfard et al. 2015).

HGNC symbol	**Associated disorder(s)**	**Inheritance**	**Disease domain**
F10	Factor X deficiency	AR; AD	Bleeding/coagulation
F11	Factor XI deficiency	AR; AD	Bleeding/coagulation
F12	Factor XII deficiency Angioedema	AR; AD	Coagulation Angioedema
F13A1	Factor XIII deficiency	AR	Bleeding/coagulation
F13B	Factor XIII deficiency	AR	Bleeding/coagulation
F2	Prothrombin deficiency Thrombophilia resulting from thrombin defect	AR; AD	Bleeding/coagulation Thrombosis
F5	Factor V deficiency Thrombophilia resulting from activated protein C resistance	AR; AD	Bleeding/coagulation Thrombosis
F7	Factor VII deficiency	AR; AD	Bleeding/coagulation
F8	Hemophilia A	XLR	Bleeding/coagulation

F9	Hemophilia B	XLR	Bleeding/coagulation
FGA	Fibrinogen deficiency	AR; AD	Bleeding
FGB	Fibrinogen deficiency	AR; AD	Bleeding
FGG	Fibrinogen deficiency	AR; AD	Bleeding
GGCX	Vitamin K-dependent clotting factors deficiency 1	AR	Bleeding/coagulation
KNG1	Kininogen deficiency	AR	Coagulation
LMAN1	Combined factor V and VIII deficiency	AR	Bleeding/coagulation
MCFD2	Combined factor V and VIII deficiency	AR	Bleeding/coagulation
SERPINE1	Plasminogen activator inhibitor 1 deficiency	AR; AD	Bleeding
SERPINF2	Alpha 2 antiplasmin deficiency	AR	Bleeding
VKORC1	Vitamin K-dependent clotting factors deficiency 2	AR	Bleeding/coagulation
VWF	VWD	AR; AD	Bleeding/Platelet
ADAMTS13	Thrombotic thrombocytopenic purpura	AR	Thrombosis
HRG	Histidine-rich glycoprotein deficiency	AD	Thrombosis
PLG	Plasminogen deficiency	AR	Thrombosis
PROC	Protein C deficiency	AR; AD	Thrombosis
PROS1	Protein S deficiency	AR; AD	Thrombosis
SERPINC1	Antithrombin deficiency	AR; AD	Thrombosis
SERPIND1	Heparin cofactor 2 deficiency	AD	Thrombosis
THBD	Thrombomodulin deficiency; Bleeding resulting from high soluble thrombomodulin	AD	Thrombosis/Bleeding
ABCG5	Sitosterolemia with macrothrombocytopenia	AR	Platelet
ABCG8	Sitosterolemia with macrothrombocytopenia	AR	Platelet
ACTB	Baraitser-Winter syndrome 1 with macrothrombocytopenia	AD	Platelet
ACTN1	Macrothrombocytopenia	AD	Platelet
ANKRD26	AD thrombocytopenia 2	AD	Platelet
ANO6	Scott syndrome	AR	Platelet
AP3B1	HPS	AR	Platelet
AP3D1	HPS	AR	Platelet
ARPC1B	Platelet abnormalities with eosinophilia and immune-mediated inflammatory disease	AR	Platelet
BLOC1S3	HPS	AR	Platelet
BLOC1S6	HPS	AR	Platelet
CDC42	Takenouchi-Kosaki syndrome with thrombocytopenia	AD	Platelet
CYCS	AD thrombocytopenia 4	AD	Platelet
DIAPH1	Macrothrombocytopenia and sensorineural hearing loss	AD	Platelet
DTNBP1	HPS	AR	Platelet
ETV6	Thrombocytopenia and susceptibility to cancer	AD	Platelet

FERMT3	Leukocyte integrin adhesion deficiency, type 3	AR	Platelet
FLI1	Paris-Trousseau and Jacobson syndrome	AR; AD	Platelet
FLNA	Syndrome with macrothrombocytopenia	XLD; XLR	Platelet
FYB1	Thrombocytopenia 3	AR	Platelet
GATA1	X-linked thrombocytopenia with dyserythropoiesis	XR	Platelet
GFI1B	Platelet-type bleeding disorder 17	AD; AR	Platelet
GNE	Myopathy associated with Thrombocytopenia	AR	Platelet
GP1BA	BSS Mild macrothrombocytopenia Platelet-type VWD	AR; AD	Platelet
GP1BB	BSS Mild macroTP	AR; AD	Platelet
GP6	Bleeding diathesis resulting from glycoprotein VI deficiency	AR	Platelet
GP9	BSS	AR	Platelet
HOXA11	Amegakaryocytic thrombocytopenia with radioulnar synostosis	AD	Platelet
HPS1	HPS	AR	Platelet
HPS3	HPS	AR	Platelet
HPS4	HPS	AR	Platelet
HPS5	HPS	AR	Platelet
HPS6	HPS	AR	Platelet
ITGA2B	GT Platelet-type bleeding disorder 16	AR; AD	Platelet
ITGB3	GT Platelet-type bleeding disorder 16	AR; AD	Platelet
KDSR	Thrombocytopenia and erythrokeratoderma	AR	Platelet
LYST	Chediak-Higashi syndrome	AR	Platelet
MECOM	Amegakaryocytic thrombocytopenia with radioulnar synostosis 2	AD	Platelet
MPIG6B	Thrombocytopenia, anaemia, and myelofibrosis	AR	Platelet
MPL	Congenital amegakaryocytic thrombocytopenia	AR	Platelet
MYH9	MYH9-related disorders	AD	Platelet
NBEA	Autism with platelet dense granule defect	AD	Platelet
NBEAL2	Grey platelet syndrome	AR	Platelet
P2RY12	ADP receptor defect	AR	Platelet
PLA2G4A	Deficiency of phospholipase A2, group IV A	AR	Platelet
PLAU	Quebec platelet disorder	AD	Platelet
RASGRP2	Platelet-type bleeding disorder 18	AR	Platelet
RBM8A	Thrombocytopenia-absent radius syndrome	AR	Platelet
RNU4ATAC	Roifman syndrome	AR	Platelet
RUNX1	Familial platelet disorder with predisposition to AML	AD	Platelet

SLFN14	Platelet-type bleeding disorder 20	AD	Platelet
SRC	Thrombocytopenia 6	AD	Platelet
STIM1	Stormorken syndrome (York platelet syndrome)	AD	Platelet
STXBP2	Familial hemophagocytic lymphohistiocytosis type 5	AR	Platelet
TBXA2R	Thromboxane A2 receptor defect	AR; AD	Platelet
TBXAS1	Ghosal syndrome	AR	Platelet
THPO	Thrombocytopenia progressing to trilineage bone marrow failure	AR	Platelet
TUBB1	Macrothrombocytopenia	AD	Platelet
VIPAS39	Arthrogryposis, renal dysfunction, and cholestasis 1	AR	Platelet
VPS33B	Arthrogryposis, renal dysfunction, and cholestasis 2	AR	Platelet
WAS	Wiskott-Aldrich syndrome	XLR	Platelet

Table 1.4 | Lists of BTPDs. Adapted from Megy et al. 2019.

ICD-10 Code category	Condition	Disease name
D82.0	Platelet disease	Wiskott-Aldrich syndrome
D68.5	Platelet disease	Primary thrombophilia
D69.6	Platelet disease	Thrombocytopenia unspecified
D69.1	Platelet disease	Qualitative platelet defects
D69.4	Platelet disease	Other primary thrombocytopenia
D69.3	Platelet disease	Idiopathic thrombocytopenic purpura
D68.0	Bleeding	Von Willebrand disease
D68.1	Bleeding	Hereditary factor XI deficiency
D68.2	Bleeding	Hereditary deficiency of other clotting factors
D68	Bleeding	Other coagulation defects
D67	Bleeding	Hereditary factor IX deficiency
D66	Bleeding	Hereditary factor VIII deficiency
D65	Thrombosis	Disseminated intravascular coagulation
I26	Thrombosis	Pulmonary embolism
I801	Thrombosis	Phlebitis and thrombophlebitis of femoral vein
I802	Thrombosis	Phlebitis and thrombophlebitis of other deep vessels
I803	Thrombosis	Phlebitis and thrombophlebitis of lower extremities

I676	Thrombosis	Nonpyogenic thrombosis of the intracranial venous system
O225	Thrombosis	Cerebral venous thrombosis in pregnancy
I81	Thrombosis	Portal vein thrombosis
I820	Thrombosis	Budd-Chiari syndrome
I822	Thrombosis	Embolism and thrombosis of vena cava
I823	Thrombosis	Embolism and thrombosis of renal vein
O223	Thrombosis	Deep phlebothrombosis in pregnancy
O871	Thrombosis	Deep phlebothrombosis in the puerperium

Table 1.5 | Lists of BTPDs phenotypes described according to the world health organization ICD-10 nomenclature (see chapter 1.5.2; https://icd.who.int/browse10/2010/en).

1.3 Genetics, epigenetics and genome function

1.3.1 Genetics

Heritability is the proportion of variation, for a defined phenotype, that can be explained by genetic inheritance (Visscher, Hill, and Wray 2008). Mendel was the first biologist to formalise trait heritability in 1866 (Mendel 1866) and, not long after the English translation of his findings, Garrod adopted Mendel's research on his alkaptonuria patients (Garrod 1902), whilst Fisher and Wright followed with the mathematical conceptualisation of trait heritability (Fisher 1919; Wright 1920). These scientists pioneered genetics and its application to the study of trait heritability much earlier than the concept of genes even existed (see chapter 1.3.4 for the definition of genes).

Indeed, the experiment that proved that the molecule encoding for the genetic material is the DNA, starting the field of molecular genetics as we know it today, happened in the 1950s (Hershey and Chase 1952). Today, we know that DNA contains the information required to pass on life, relying just on 4 nucleotides: Adenine (A), Thymine (T), Cytosine (C) and Guanine (G; Watson and Crick 1953).

There are two main modes of trait heritability, Mendelian traits and complex traits (Antonarakis and Beckmann 2006). In the case of a Mendelian trait, the heritability is explained almost entirely by one locus (Antonarakis and Beckmann 2006; NIH definition of Mendelian inheritance, accessed in September 2021). However, the proportion of phenotype

variability explained solely by one locus is rarely 100%, even for Mendelian trait inheritance (Visscher, Hill, and Wray 2008; Hill, Goddard, and Visscher 2008). For instance, in the case of cystic fibrosis, which severely compromises lung function, several genes contribute to the variability of the phenotype influencing the final lung function (Cutting 2010).

In the case of a complex trait, multiple genes and/or multiple loci contribute to the manifestation of the same phenotypes (Lander and Schork 1994; Antonarakis and Beckmann 2006). Different models have been created to explain how several genes orchestrate to obtain the final phenotype (Marchini, Donnelly, and Cardon 2005), but they all mainly rely on two basic concepts: additive effect or epistatic effect. Additive effect models assume that every gene gives an independent contribution to the phenotypic manifestation, even if minimal (Robertson and Lerner 1949; Dempster and Lerner 1950; Yang et al. 2011). More complex epistatic effect models take into account gene-gene interplay and epistatic influences of one gene onto the effect of another (Mackay and Moore 2015), but, due to their complexity, these are less commonly applied to the study of human heritability (Yang et al. 2011; Sheppard et al. 2021).

All the different approaches used to study genetic inheritance do not explain heritability and trait variabilities fully (Kong et al. 2018; Young 2019). Exposure to environmental stimuli influences gene expression and the phenotype we observe; giving rise to the nature versus nurture dichotomy and the rising epigenetic field (Kong et al. 2018).

1.3.2 Epigenetics

The idea that the DNA molecule *per se* does not contain all the relevant information to cell function in multicellular organisms was postulated, in the 1950s (*The Physical Foundation of Biology* 1958), possibly because earlier a DNA modification, namely 5-methylcytosine, was observed in mammalian DNA (Hotchkiss 1948). The role of methylation in gene regulation was reported in the 1980s (Razin and Riggs 1980) and soon after associated with human diseases, cancer in particular (Riggs and Jones 1983). In 1984, an elegant experiment by McGrath and Solter showed that the sole DNA sequence is not sufficient to support the correct mouse embryogenesis and therefore other information needed to be passed (McGrath and Solter 1984), starting a new field of research.

The interpretation of this additional code of information is the subject of study of epigenetics, a term introduced in its modern meaning by Moss in 1981 (Moss 1981). Nowadays, it is accepted that on top of the DNA sequence, eukaryotes have further layers of

complexity resulting from (i) nucleotide modifications (e.g. 5-methylcytosine), (ii) histones post-translational modifications (e.g. acetylation) and (iii) chromatin conformation (e.g. loop; Kelsey, Stegle, and Reik 2017). Importantly, none of the above-mentioned epigenetic modifications affects the DNA sequence. These additional layers of complexity are required to achieve the correct regulation of the gene expression in embryonic development, cell differentiation and ageing (Kelsey, Stegle, and Reik 2017). Interestingly, at least in part, this information can be passed to the offspring, priming behaviour or predisposing to certain traits (Juengst et al. 2014).

There are several nucleotide modifications, such as adenine or cytosine methylation (Kumar, Chinnusamy, and Mohapatra 2018; Fang et al. 2012). DNA methylation is the addition of a methyl group to DNA nucleotides. The most common modification present in the human genome is 5-methylcytosine (i.e 5mC; Ehrlich et al. 1982; Matsuda et al. 2018). 5mC is found particularly at the promoter of genes, defining the CpG islands and its role is to repress gene expression (Deaton and Bird 2011). Possibly, the best examples of genetic imprinting, mediated by DNA methylation, come from two human diseases, the Angelman and Prader-Willi syndromes (Buiting et al. 1995). The group of genes that causes Prader-Willi syndrome is expressed only from the paternal chromosome 15, whilst maternal genes are methylated and therefore silenced, hence the loss of the paternal copy of these genes causes the homonym syndrome. Similarly, Angelman syndrome has the uniparental expression of the maternal alleles, because of the methylation of the paternal ones, therefore leading to a similar disease mechanism if the maternal copies of the genes are non-functional (Buiting et al. 1995; Stelzer et al. 2014).

Eukaryotic DNA is wrapped around proteins, i.e. histones, that compact and organise the DNA structure (Alberts 2014). On top of this mechanical contribution, histones control the state and function of the DNA via chemical modifications (Alberts 2014; Kelsey, Stegle, and Reik 2017). Arguably, the most well-characterised histone modification is the acetylation of the twenty-seventh lysine of the histone H3 (H3K27Ac), which labels genomic regions that are actively transcribing (i.e. promoters and enhancers, see chapter 1.3.4 and 1.3.5; Rosenfeld et al. 2009). A list of the most commonly found histone modifications and their biological functions is in Table 1.6 (Wang et al. 2008; Rosenfeld et al. 2009; Calo and Wysocka 2013; Huang, Litt, and Blakey 2015).

The combination of DNA sequence, DNA modification and histone modifications shape the 3D structure of the genome. This higher structure information is a fundamental

part of gene regulation, epigenetics and cell identity (see chapter 1.3.5; Dekker, Marti-Renom, and Mirny 2013; Schoenfelder and Fraser 2019).

Histone	Residue	Modification	Biological functions
H2A	Lysin 5	Acetylation	Transcription activation
H2A	Lysine 9	Acetylation	Highly expressed genes
H2B	Lysine 12	Acetylation	Highly expressed genes
H2B	Lysine 20	Acetylation	Enhancers
H2B	Lysine 120	Acetylation	Low gene expression,
H3	Lysine 9	Acetylation	Active promoters
H3	Lysine 14	Acetylation	Low gene expression, subtelomeric regions
H3	Lysine 9	Di-methylation	Non-genic regions, pericentromeric regions
H3	Lysine 9	Tri-methylation	Heterochromatin, silenced genes
H3	Lysine 27	Tri-methylation	Silenced genes
H3	Lysine 4	Mono-methylation	Poised enhancers (when with H3K27me3)
H3	Lysine 4	Tri-methylation	Active promoters
H3	Lysine 27	Acetylation	Highly expressed genes, active regulatory regions (promoter and enhancers)
H3	Lysine 79	Di-methylation	Active gene bodies
H4	Lysine 5	Acetylation	Mark TSS and gene bodies
H4	Lysine 8	Acetylation	Active promoters and transcribed regions
H4	Lysine 20	Mono-methylation	Cell cycle regulation

Table 1.6 | List and function of the most common histone modifications (Zhibin Wang et al. 2008; Rosenfeld et al. 2009; Calo and Wysocka 2013; Huang, Litt, and Blakey 2015).

1.3.3 Human genome

The human genome has a size of 3.2 billion base pairs (“GRCh38.p13 - Genome - Assembly - NCBI”). The first complete human reference sequence has been the outcome of the ‘Human Genome Project’, which took about 20 years from its conceptualisation in 1984 to its completion in 2003 (Bentley 2000; McPherson et al. 2001; Lander et al. 2001; Collins et al. 2003). Despite almost 40 years of international collaboration, the human genome reference sequence has still some unresolved regions and consortia, such as “Telomere-to-Telomere”, are actively working to fill in the gaps that are still persisting (Miga et al. 2020; GIS). Indeed, some relevant parts of the human genome, such as highly

repetitive regions or transposon elements, are particularly difficult to sequence and map on the reference genome (Miga et al. 2020). The human genomic sequence has been key to improve our understanding of basic biological processes, such as chromosomal organisation or mitosis, but it has also been at the basis of the genomic studies, which allowed the unravelling of the genetic architecture of common and rare diseases (Miga et al. 2020; Rood and Regev 2021).

During the intervening years of molecular genetic research, the role of the coding genome, i.e. genes, has been understood relatively well. On the other hand, the role of the non-coding part of the genome started to be evident and broadly studied more recently. Indeed, what was initially referred to as “junk” DNA (Ehret and de Haller 1963; Ohno 1972) is now known to have several functions, from regulating gene expression to achieving the correct structural organisation of the DNA within the cell nucleus and therefore driving the evolution (Biémont and Vieira 2006; Palazzo and Gregory 2014; Ecker et al. 2012).

1.3.4 The coding genome

With the term coding genome, I refer to that part of the genome (~4%) that encodes information for the synthesis of other molecules, namely protein and non-coding RNA (Palazzo and Lee 2015). The coding genome led molecular biology studies since the late ‘50s and brought to the postulation of the “central dogma of molecular biology” (Crick 1958; Crick 1970). Today, a revised and expanded vision to the “central dogma of molecular biology” includes a few more instances, indeed the product of a gene can be either coding or non-coding for proteins: the former has its RNA molecule spliced into mRNA and translated into protein, the latter use RNA molecules as its final product.

There are about 20,000 to 25,000 protein-coding genes in the human genome. The number of possible transcripts also increases dramatically when considering alternative splicing. Indeed it has been calculated that at least 95% of the genes use alternative splicing (Pan et al. 2008) for a total of more than 100,000 alternative transcripts (Nilsen and Graveley 2010). While the number of non-coding genes is yet to be determined, it is thought to be around the same number (Palazzo and Lee 2015); well-known examples of non-coding genes are the tRNA and rRNA genes (International Human Genome Sequencing Consortium 2004; Clamp et al. 2007; Pertea et al. 2018). Regardless of being coding or not, most of the genes have a shared structure that is based on three functional units: promoter, introns and exons.

Promoters

Promoters are genetic sequences that allow the expression of a gene and, usually, they are localised upstream the gene transcription starting site (TSS), which is the region where the RNA polymerase begins to transcribe (Alberts 2014). Generally, promoters are located in the 2Kb regions surrounding the TSS (Alberts 2014). The presence of tissue-specific transcription factor binding sites leads to the alternative use of the promoter regions that are relevant for cell-type specific transcription (Chen et al. 2016). There are different types of promoters, defined by their different nucleotides ratio (Gagniuc and Ionescu-Tirgoviste 2012). For instance, human promoters are characterised by their high guanine and cytosine (GC) concentration, which is forming the hypomethylated sequences usually called ‘CpG’ islands (see Chapter 1.3.2; Gagniuc and Ionescu-Tirgoviste 2012). Promoters act together with enhancers to recruit the molecular machinery required for the transcription initiation and extension (see chapter 1.3.5 for a description of the proteins involved in the transcription and its regulation).

Introns

Introns are the regions of a gene that usually do not code for proteins (Alberts 2014). With few exceptions, introns are usually spliced out, during transcription, from the mature mRNA, and rarely are retained to form alternative mRNAs (Herzel et al. 2017; Zheng et al. 2020). The biological evolution of introns is widely discussed and not resolved yet. They are absent in prokaryotes and It is not clear whether they have always been present and lost in this domain or acquired after prokaryotes and eukaryotes evolution diverged (Jeffares, Mourier, and Penny 2006). Introns are actively involved in the trafficking of the pre-mRNA from the nucleus to the cytoplasm and they regulate mRNA splicing and its stability (Jeffares, Mourier, and Penny 2006; Cenik et al. 2011; Bicknell et al. 2012).

Exons

The exons mainly contain the nucleotides sequences that are retained into the mature mRNA after splicing and that during the translation process are converted to amino acids forming the proteins (Alberts 2014). The coding information is encoded and organised in consecutive triplets of nucleotides, called codons, each containing the information for one amino acid (Crick 1966a). There are 64 possible different combinations of the 4 nucleotides (i.e. A, T, C and G) and they encode the information for the translation of 21 amino acids (Alberts 2014). These redundancies in the triplets encoding for amino acids are referred to

as “amino acid code degeneration”. This implies that translation is more tolerant to single nucleotide variations (Lagerkvist 1978; Crick 1966b; Barrell, Bankier, and Drouin 1979). This tolerance is vital to confer some resilience against the naturally occurring variants, leading to synonymous variants. In spite of the code degeneration, variants may still affect translation and protein structure, especially if involving multiple nucleotides (see chapter 1.3.6 for the description of variants). The reading frame (i.e. consecutive non-overlapping triplets of nucleotides) of the codons needs to be respected during the transcription, splicing and translation, otherwise frameshift variants can disrupt the correct reading frame and compromise protein function (see chapter 1.3.6).

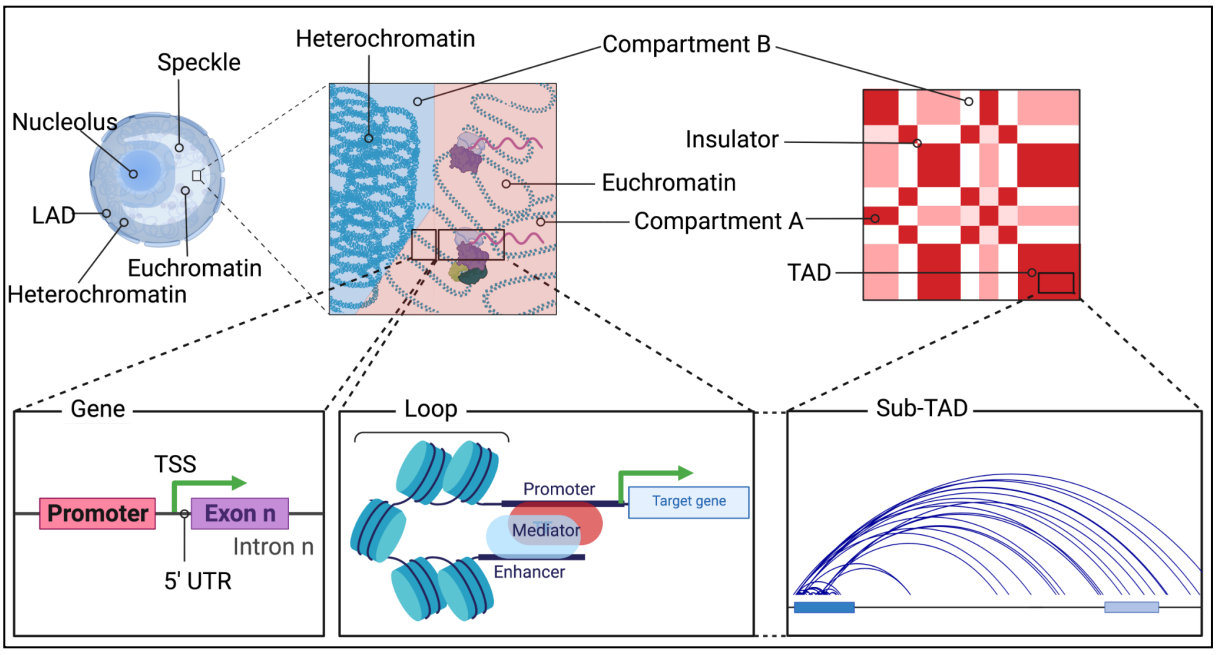


Fig. 1.3 | Cartoon representing the genetic and epigenetics structure of the human genome. *TAD = Topologically associated domain. LAD = lamina associated domain. TSS = transcription starting site. UTR = untranslated region.*

1.3.5 The non-coding portion of the genome

Non-coding regions are referred to as the regions of the genome that are not part of genes. These regions are more tolerant to variation and, for this reason, poorly conserved across evolutionarily distant animals. Non-coding variants that do not influence fitness are not under evolutionary pressure. However, roughly 5% of the genome, in the case of humans, makes exceptions and deviates from neutral evolution (see chapter 1.3.6; Bejerano et al. 2004; Boffelli, Nobrega, and Rubin 2004).

The role of these conserved regions is still partially under examination, however, it has been shown that some of these are crucial for transcription (Visel et al. 2008, 2013; Dickel et al. 2018; Byeon et al. 2021) or post-transcriptional regulation (Byeon et al. 2021). Non-coding regions harbour functional elements that are required for gene regulation, especially to confer the correct identity to the different cell types (Young 2011; Whyte et al. 2013), and regulate the basic biological processes such as in the case of mitosis (i.e. centromeres, Fachinetti et al. 2015).

Focusing the description on interphase chromatin, the main human genomic non-coding features are enhancers, TAD, LAD, compartments, euchromatin and heterochromatin.

Enhancers

Enhancers are eukaryotic transcriptional modulators and they can be located anywhere with respect to their cognate genes, upstream, downstream or even within its cognate or other genes (Alberts 2014). Enhancers contribute to determining cell identity during development and differentiation (Freire-Pritchett et al. 2017) and they are estimated to be in the order of hundreds of thousands in all human cell types (Pennacchio et al. 2013).

There are several mechanisms used by enhancers to regulate gene expression; experimental evidence points to long-range direct physical interactions called “loops” (Deng et al. 2012; de Wit et al. 2015; Schoenfelder and Fraser 2019; Bonev et al. 2017). Most likely, mechanisms other than direct contact loops exist and they depend on the distance and type of interaction. For instance, linking is a mechanism that connects enhancers to the promoter with a chain of proteins that does not require looping and physical contact (Gupta et al. 2017; McCord, Kaplan, and Giorgetti 2020; Furlong and Levine 2018). Indeed, Benabdallah and colleagues characterised and reported an example in which when the enhancer moves away from the promoter of SOX1, it increases the expression of this gene (Benabdallah et al. 2019). Also, more complex scenarios exist, for instance, *UMLILO*, a lncRNA, regulates cognate genes sharing a regulatory circuit with enhancers, therefore a combined regulation of DNA and RNA (Fanucchi et al. 2019).

Chen and colleagues generated experimental evidence of the distance effect on the enhancer-promoter interaction with microscopy and also investigated the dynamics of these interactions in *Drosophila* (Chen et al. 2018). A stable and persistent association is required for transcription in the regulatory loop they investigated (Chen et al. 2018). However, the

dynamics and duration of the contacts look less stable in studies that investigated these at single-cell topology (Bintu et al. 2018). Promoter-enhancers are dynamic interactions that stabilise and increase the residence time of the RNA polymerase II machinery, at the population level, on the promoter of the gene they regulate, leading to an increase in the number of RNA molecules produced (Bartman et al. 2016).

Enhancers contain motif sequences for constitutive and tissue-specific transcription factors that are required to regulate the cognate genes and cofactors are required to link enhancers to the cognate genes and promote gene expression (Whyte et al. 2013; Schoenfelder and Fraser 2019). Fundamental proteins that are exclusively involved in the enhancer-promoter loops are Mediator, p300, YY1, LDB1. Other proteins (i.e. CTCF and cohesin) are involved in enhancer-promoter loops, but probably more for structural reasons.

The most likely molecular topology that links enhancers and promoters are loops (Eeftens and Dekker 2017). Part of the proteins required to have functional regulatory loops have yet to be identified, but there are a few known players (de Wit et al. 2015; Nora et al. 2017; Hansen et al. 2017; Bintu et al. 2018). The ‘loop extrusion model’ sees CTCF and cohesin binding the DNA on CTCF binding sites. Once bound to the DNA, CTCF works as an anchor and cohesin pulls the string of DNA through the complex creating a loop, hence the model name. The extrusion stops when the complex encounters another CTCF protein anchored to the DNA, but oriented in the opposite direction. During the process of extrusion, the enhancers come in close proximity to the regulated genes and cause the gene to be transcribed. The CTCF loop extrusion has two roles: regulating enhancer-promoter interactions and defining the higher structure of genomic architecture, indeed it is often anchored at the edges of topologically associated domains (TAD) with its companion protein, cohesin (Schoenfelder and Fraser 2019).

TADs

Enhancers can interact with promoters up to 1 megabase (Mb) apart, but these interactions are usually contained within the same TAD (Lupiáñez et al. 2015; Nora et al. 2012; Schoenfelder and Fraser 2019). Indeed, TADs are thought to constrain the “movement” of enhancers and, by doing so, increase the probability that they find the correct target genes (Schoenfelder and Fraser 2019); or, on the contrary, prevent the activation of genes that should not be active (Lupiáñez et al. 2015). Confining the spreading of chromatin

state (i.e. euchromatin versus heterochromatin) has also been hypothesised as one of the other functions of TADs in eukaryotes (Narendra et al. 2015).

CTCF loops usually connect and define the insulators, boundaries that are at the edges of a TAD and constrain the DNA movement. Insulators are genetic features composed of several proteins that constrain enhancers and chromatin modification in discrete regions of the genome (West, Gaszner, and Felsenfeld 2002; Gaszner and Felsenfeld 2006). There are thousands of TAD insulators in the human genome (Hnisz, Day, and Young 2016), often in the sequences surrounding enhancers (Schoenfelder and Fraser 2019). Despite the existence of insulators, about a third of the interactions of each viewpoint span across TAD boundaries (Javierre et al. 2016). This incomplete insulation may be a reflection of a subgroup of cells having different TAD conformations or of the fact that TADs are more dynamic structures than one would have imagined (Bintu et al. 2018).

The role of TADs on transcription regulation needs more investigation. Indeed, when TAD boundaries are reshaped, either by the depletion of CTCF or cohesin, gene expression does not always reflect expectations (Zuin et al. 2014; Nora et al. 2017; Schwarzer et al. 2017). At the mouse *Shh* locus, the inversion of the CTCF binding sites compromised TAD structure and altered *Shh* expression resulting in monodactyly (Symmons et al. 2016). This observation is in contrast with the one from the mouse *Sox9-Kcnj2* locus. In this region, the ablation of the TAD insulator did not alter gene expression, suggesting that the enhancer-promoter loop was maintained independently of TAD absence (Despang et al. 2019). This expression independent from the TAD structure suggests that transcription may also play a role in the organisation of the chromatin, especially for the interactions happening at the sub-TAD level (Dixon et al. 2012; van Steensel and Furlong 2019).

TADs tend to be conserved across cell types and species (Dixon et al. 2015; Battulin et al. 2015), however, the smaller domains within TADs, named sub-TADs, are highly dynamic and enriched in regulatory loops that may also contribute to the 3D structure (Bintu et al. 2018). These sub-TAD interactions undergo the most dramatic changes in cell differentiation (Schoenfelder and Fraser 2019), however, some interactions pre-exist and become functional just after differentiation (Stadhouders, Filion, and Graf 2019).

LADs

Lamina associated domains (LADs) are large regions of DNA (from 10Kb to 10Mb) that during interphase are associated with the nuclear lamina (ENCODE Project Consortium

2012; van Steensel and Belmont 2017). LADs are characterised by repressed chromatin and inactive transcription, physically separating them from the active chromatin (van Steensel and Belmont 2017). LADs can be constitutive, if always found to be associated with the nuclear lamina, or facultative, if associated with the nuclei periphery only in certain cell types (Meuleman et al. 2013). Interestingly, while the LAD positions in the chromosomes are evolutionarily conserved, their sequence is not. This observation indicates that further investigations are needed to understand the mechanisms of association of the nuclear lamina to the DNA (Meuleman et al. 2013). LAD separation from the active chromatin occurs because of structural proteins, such as CTCF, or active promoters that block the heterochromatic state expansion (Guelen et al. 2008). The chromatin state is repressed as shown by the histone marks (see Table 1.6).

Compartments

Compartments are large chromatin domains, larger than TAD, that have been discovered with the use of high-throughput technologies to study chromosomal structure (Lieberman-Aiden et al. 2009). They are more dynamic than TADs and show great rearrangement during cell differentiation (Schoenfelder and Fraser 2019) thanks to transcription factors and chromatin remodelers, such as YY1 and CTCF (Therizols et al. 2014; Harr et al. 2015; Wijchers et al. 2016; van Steensel and Belmont 2017).

Compartments are spatially segregated within the nucleus and are commonly referred to as “compartment A” and “compartment B” (Lieberman-Aiden et al. 2009). During the interphase, compartment A is positioned in the centre of active chromatin, while compartment B is in the proximity of the nuclear lamina (LAD) and nucleolus (Wijchers et al. 2015; Vieux-Rochas et al. 2015; van Steensel and Belmont 2017). Compartment As are transcriptionally active regions and are characterised by genomic features such as H3K27Ac (Guelen et al. 2008; Giorgetti et al. 2016). On the other hand, compartment Bs are transcriptionally inactive and histone marks, such as H3K9me3 are widespread in these regions (van Steensel and Belmont 2017). The two compartments almost entirely overlap with euchromatin and heterochromatin, respectively compartment A and compartment B (Lieberman-Aiden et al. 2009).

Euchromatin and heterochromatin

At the beginning of the 20th century, Emil Heitz noticed some differences in the nuclei response to cytological staining. The DNA within the nuclei is not homogeneously distributed

and Heitz called euchromatin the regions less dense and heterochromatin the denser ones (Heitz 1928).

Nowadays, these differences have been investigated further and linked to their molecular state and biological meaning. Euchromatin is the chromatin that is actively transcribed and it is characterised by H3K4me1, H3K4me3, H3K27Ac and a few other modifications (see Table 1.6). There are other structures that are present exclusively in the euchromatin, namely speckles. Speckles are regions of high transcriptional activity and are thought to be formed thanks to phase-phase separation mechanisms (Misteli, Cáceres, and Spector 1997; Lamond and Spector 2003; Kim et al. 2019). Heterochromatin is linked to the inactive state of transcription and is associated with histone markers such as H3K9me2, H3K9me3, H3K27me3.

Events that compromise the structure or the function of any of the genomic features described above have been linked to inherited disorders. The main cause of aberration of these biological structures is genetic variants.

1.3.6 DNA variants: types and impact

Genetic variants occur naturally in the genome of every living organism and are the main source of evolution (Barrett and Schluter 2008; Futuyma and Kirkpatrick 2017). DNA can mutate because of chemical (e.g. reactive oxygen species or ethidium bromide), physical (e.g. ultraviolet or radioactivity decay), and biological events (e.g. DNA break during replication, transposons and viruses; Kozmin et al. 2005; McKinnon and Caldecott 2007; Halliwell and Gutteridge 2015). Interestingly, each source of variation leads to a particular signature mutation in the DNA, as for example, ultraviolet light mutations are enriched in the transition C→T (Alexandrov et al. 2013; Maura et al. 2019; Zou et al. 2021). Genetic variants create different versions of a given genomic locus and these different versions of the same region are called alleles (Elston, Satagopan, and Sun 2012). When multiple variants co-occur together more often than expected from random recombination, these clusters of variants are called haplotypes (Hartl and Clark 1997; Elston, Satagopan, and Sun 2012).

Variation in the genome tends to occur randomly, with a few exceptions of mutational hotspots (Nesta, Tafur, and Beck 2020). The effect of variants on fitness (i.e. ability to pass the genetic material of an organism to its progeny) is in part determined by the effect of the variation on the molecule function and in part determined by the environment and natural selection. The balance of these two components determines if a variant is fixed in a

population or lost (see below the example on malaria; Smith and Smith 1989; Hartl and Clark 1997; Eyre-Walker and Keightley 2007). Indeed, if a variant confers an advantage then it will be positively selected, or if the variant decreases the fitness it will be negatively selected (Eyre-Walker and Keightley 2007). There is also the possibility that some variants are neutral to any function, in this case, their allele frequencies are determined by genetic drift and therefore independent from natural selection (Eyre-Walker and Keightley 2007). Natural selection or genetic drift determines the frequency of an allele in a population (i.e. allele frequency, AF). The AF of a variant is the number of alleles with that variant present in a given population over the total number of alleles for that region present in the population (Elston, Satagopan, and Sun 2012).

The population frequency of a variant over time has been mathematically described in the Hardy-Weinberg equation (Hardy 1908). In a locus that is occupied by 2 alleles, the equation is: $(p + q)^2 = 1$, where *p* and *q* are the AF of the two alleles. This means that the allele frequency stays constant over generations. However, this equilibrium is valid only if the population respects some assumptions (e.g. organisms are diploid, the population is large and matings are random; Edwards 2008). It is important to note that genetic mutations are one of the causes that induce deviation from the AF equilibrium (Edwards 2008).

Variants drove the speciation of the *Homo sapiens* from hominids and, nowadays, we know that the human DNA differs, from the closest living relative, of about 1.23% of their genomes, with 35 million SNPs and approximately 5 million structural variants (Wienberg 2005; Chimpanzee Sequencing and Analysis Consortium 2005; Khaitovich et al. 2006). From that branching point onwards, the human genome kept acquiring new variants and alleles, some are shared across all humans, whilst others appeared later and remained confined to certain ancestries due to geographical or cultural differences (Relethford 2008). For instance, sub-Saharan Africa inhabitants have the highest levels of genetic diversity, in line with the theory of the origin of the human species in this continent (Relethford 2008). Statistical approaches, such as principal component analysis (PCA), can determine and discriminate amongst different ethnic groups based on the differences in the common variants present in each ancestry (Patterson, Price, and Reich 2006; Alexander, Novembre, and Lange 2009). Similar approaches are used to study admixture and reconstruct human migration (Choudhury et al. 2020).

The presence and AF of a given variant are also due to environmental selective pressures. For example, the variants in haemoglobins that cause sickle cell disease are

present in individuals from sub-Saharan Africa with a relatively common AF. In fact, being a carrier of those variants is protective against malaria, an endemic disease in certain African countries. In homozygosity, these alleles cause severe forms of sickle cell disease which is a life-threatening condition as they compromise red cells function. Therefore, the AF of haemoglobin variants is a tight balance between sickle cell disease and the protective effect that, in heterozygosity, these variants confer against the malaria parasite.

If variants decrease the function of a protein or alter the regulatory capacity of a DNA region, up to completely inhibiting its functions, they are generally called a LoF variant. Whilst, variants conferring extra activities to the DNA molecule or protein are named GoF. The intrinsic probability of a gene to be LoF intolerant (pLI) has been recently calculated thanks to the sequencing aggregation efforts of human genetic consortia, namely exAC and gnomAD (see chapter 1.5.1; Lek et al. 2016; Minikel et al. 2020). pLI is a reflection of how many times LoF variants (e.g. premature stop codons) are observed in a gene, normalised by the size of the gene and the number of variants that would be observed if that gene was not under selective pressure. If a gene is observed to be less mutated than expected it means that this gene is probably not tolerant to LoF variants and reflected with a pLI score closer to 1 (pLI score goes from 0 to 1; Lek et al. 2016).

Genetic variants are classified mainly according to the number of nucleotides affected by the change. The main groups of genetic variants are listed below.

Single nucleotide polymorphisms

Single nucleotide polymorphisms (SNPs), often also called single nucleotide variants (SNVs) are the most tolerated variants in both coding and non-coding regions, they are substitutions of one nucleotide for another (Alberts 2014). They are called transitions if the purine and pyrimidine chemical structures are conserved in the mutations (i.e. A $\rightleftarrows$ T or C $\rightleftarrows$ G) or transversions if a purine flips to pyrimidine or *vice versa* (i.e. A $\rightleftarrows$ C; A $\rightleftarrows$ G; T $\rightleftarrows$ C; T $\rightleftarrows$ G).

If a SNP occurs in a transcription factor binding site or in a region that regulates other epigenetic features, it can quantitatively alter protein expression and therefore influence biological pathways. A very well-known example of a phenotype caused by non-coding SNV is Hemingway's cat and its descendants, which have polydactyly because of a point mutation in the *ZRS* regulatory region (Schoenfelder and Fraser 2019). If SNPs occur within exons some of the variants will be synonymous, i.e. a variant in the DNA will not reflect on

the protein sequence and therefore function, because of the degenerated nature of the genetic code. These variants tend to be functionally neutral and are not subject to selection during evolution.

Non-synonymous SNPs cause a change in the amino acid sequence or introduce a premature stop codon, and they are usually referred to as missense or nonsense variants respectively. Depending on their effect on the proteins, missense variants can be functionally neutral, LoF or GoF. The missense AFs undergo natural selection or genetic drift and eventually their AF is fixed in or erased from the population. There are well-known cases of polymorphic loci, such as the ABO blood group, where at least 3 main haplotypes coexist and encode for the presence of the A or B transferase, or in the case of group O a nonsense SNP leads to a premature stop codon and the absence of either transferase (Seltsam et al. 2003). The increased number of genome sequencing studies at population scale and in populations of different ethnicities is incrementally improving the catalogue of polymorphic loci, but saturation has not been achieved (Karczewski et al. 2020). Indeed, each new genome sequence from individuals of European ancestry identifies approximately 1,000 DNA variants previously unobserved (Turro et al. 2020).

Another category of coding SNPs may ablate the transcription start or stop codons. Losing the initiation or termination codon has a high impact on the protein structure with consequences on the protein function, generally leading to a LoF (Alberts 2014). Alternatively, SNPs can affect splice-sites and alter the protein structure by changing the splicing events. Alternative splicing can be deleterious for the protein function, because it may result in a different reading frame or because the protein may miss an important domain (Alberts 2014).

Insertions and deletions

INDEL is the univerbation of the words insertion and deletion, this category of variants refers to the addition or removal of a string of nucleotides. There is not a defined consensus on the size of INDELs, but generally, they are between 50 and 10,000 (or 10 Kb) nucleotides in length (Mullaney et al. 2010). The number of INDELs in a human genome is far less than SNPs, with estimates of one INDEL for every 10 SNPs (Beyter et al. 2021).

In coding regions, if variants involve more than one nucleotide then the mature mRNA sequence might result in a frameshift. If the variant is a multiple of the genetic codons (i.e. 3 nucleotides) and it is in frame, then the new reading frame may be better tolerated

because it returns to the correct amino acid sequence just after the INDEL. If the variant puts the mRNA out-of-frame then the impact on the protein tends to be larger and more deleterious, especially because out of frame mRNA will often result in a premature stop codon and a truncated form of the protein that will be degraded via the nonsense-mediated decay ubiquitin-proteasome systems (Amm, Sommer, and Wolf 2014; Lykke-Andersen and Jensen 2015).

Because of their size, INDELs may remove a big portion of a gene or a non-coding functional region. The biological result of these aberrations depends on the region affected: if the gene or genes involved are haplosufficient, then the effect of the variants may be tolerated, whilst if the genes are haploinsufficient, then the INDELs may result in a compromised biological pathway.

Structural variants

Structural variants (SVs) are all variants larger than 10 Kb, they are complex rearrangement events such as duplication, inversion, deletion, translocation or repositioning of mobile elements (Collins et al. 2020). SVs are the most difficult variation events to detect using the results from high-throughput sequencing technologies. Most genomic studies rely on short-read sequencing (e.g. Illumina sequencing) and this technology does not directly observe SVs but infers them *post hoc* via the aberrations generated in the alignment steps (e.g. depth of coverage or pair-read distances; Coster, De Coster, and Van Broeckhoven 2019). DNA properties, such as GC content, or technical aspects of the protocols, such as PCR amplification, can introduce biases and affect the SV calling (Coster, De Coster, and Van Broeckhoven 2019). However, progress has been made in the biology of SVs with the use of the 3rd generation long-read nucleic acid sequencers, such as Oxford Nanopore or PacBio (Mahmoud et al. 2019; Sedlazeck, Rescheneder, et al. 2018; Chaisson et al. 2015; Audano et al. 2019). This technology, in spite of having a higher nucleotide call error rate, has much higher mappability for SVs because it directly sequences across breakpoints and repetitive regions (Treangen and Salzberg 2011; Sedlazeck, Lee, et al. 2018).

One of the earliest and most famous examples of non-coding SVs affecting a phenotype came from the laboratory of Douglas Higgs. In a manuscript published in 1990, Hatton and colleagues identified a 62Kb deletion upstream the haemoglobin locus and causally associated this SV to α-thalassemia (Hatton et al. 1990).

1.3.7 Role of genetics to study disease aetiologies, improve diagnostics and design new drugs

Variants that alter the function of any of the genetic features described in the previous paragraphs can contribute to the heritability of diseases (Schoenfelder and Fraser 2019; Claussnitzer et al. 2020). These can be complex or Mendelian disorders, depending on the genetic architecture that causes the phenotype (see chapter 1.3.1).

Mendelian disorders cumulatively affect 1 in 20 people (Boycott et al. 2017), and scientists have been discovering variants and genes associated with this group of diseases for decades, preceding the introduction of next-generation sequencing technologies (Claussnitzer et al. 2020). Before the advent of these new technologies resolving the genetic cause of rare diseases relied on biochemical studies combined with genetic investigations in large informative pedigrees allowing the use of linkage analysis to identify regions for targeted investigation (Jarvik 1998). These endeavours have populated gene-to-phenotype-association databases with thousands of genes and variants known to be causally linked to rare inherited diseases (see chapter 1.5.2).

Autosomal recessive diseases are caused by two pathogenic variants co-occurring on the two alleles of a locus, either in homozygosity (the same variant on both alleles) or compound heterozygosity (two different variants affecting the function of the same locus). This mode of inheritance implies that one functioning copy of the gene is sufficient to correctly perform the relevant biological functions (Genetic Alliance, 2010). The probability that a person inherits both pathogenic alleles from parents that are carriers of deleterious variants (i.e. having the pathogenic variants but in heterozygosity), is 0.25, as postulated in Mendel's laws. The AF for recessive variants is expected to be lower than 1 in 1,000 (Turro et al. 2020a) but needs adjustment based on the prevalence of the disease in consideration. Indeed, the AF of pathogenic variants spans from singletons (i.e. private variants present in a single person) to relatively common ones (AF > 0.001). For example, chr12:6034812:C>T (GRCh38) has an AF of about 1 in 300 and it is the most common cause of VWD in Europeans (Goodeve 2010; Downes et al. 2019). Another example of an autosomal recessive condition, from the BTPD disease domain, is the grey platelet syndrome (GPS). GPS is a rare autosomal disorder caused by LoF variants in *NBEAL2* and is characterised by macrothrombocytopaenia, bleeding diathesis, absence of platelet α-granules (Albers et al. 2011; Sims et al. 2020; Mayer et al. 2018). A recent cohort study also revealed extensive

immune dysregulation, resulting in the presence of autoantibodies and autoimmune diseases in many of the GPS patients (Sims et al. 2020).

In contrast, inherited diseases with an autosomal dominant mode of inheritance are caused by a single pathogenic variant. This consequence indicates that, in the case of LoF variants, a single copy of the gene cannot maintain normal biological function, or that, in the case of GoF variants, the excess gene activity leads to pathology. The reference AF that is typically used for variants causing dominant conditions is 1 in 10,000 (Turro et al. 2020) and the molecular mechanisms leading to dominant conditions are haploinsufficiency, a GoF effect or dominant-negative effect (Herskowitz 1987; Veitia 2002). The dominant-negative refers mainly to proteins that multimerise and negatively affect also the wild-type copy of the gene (Veitia 2002). The probability that a person inherits the pathogenic variant from a heterozygous and affected parent is 50% (Genetic Alliance, 2010). For example, a single rare variant of the *RUNX1* gene may cause an autosomal dominant condition that initially presents as thrombocytopenia and at a later age may be complicated by acute myeloid leukaemia (Song et al. 1999).

The mode of inheritance described in the paragraph above refers just to autosomal chromosomes (i.e. not sexual chromosomes, from 1 to 22 in humans). Indeed, genes that lie in the sexual chromosomes (i.e. chromosomes X and Y in humans) follow a slightly different mode of inheritance, because males are hemizygous for the X chromosome and the only carriers of the Y chromosomes. In this case, diseases that would normally be recessive, tend to be dominant in the male offspring (Genetic Alliance 2010). Haemophilia A is an example of X-linked recessive disease. Males that inherit pathogenic variants from their mothers manifest directly the phenotype, while it is very rare for females to be affected because they need both the X chromosomes to carry a pathogenic variant. Variants in *F8* are the cause of haemophilia A, there are thousands of documented variants (Giansily-Blaizot et al. 2020), 30% of which *de novo* (i.e. spontaneous mutation absent in the parents; Kentsis et al. 2009) with the most common variant being the inversion of intron 22, which explains ~40% of all the haemophilia A cases (Lakich et al. 1993).

The development of high-throughput genotype technologies, namely SNP array and short-read sequencing, made possible the study of the genetic architecture for rare and common complex conditions, such as BTPDs, cancer or autoimmunity (Ng et al. 2009; Albers et al. 2012; Firth et al. 2011; Karczewski et al. 2020; Turro et al. 2020; Thaventhiran et al. 2020; Smedley et al. 2021).

Genetics has proven itself an invaluable tool in perinatal counselling, diagnostics and prognostication of inherited disorders. For this reason, there has been a growing investment in genotyping in clinical genetics, which also urged the need for guidelines in variant pathogenicity interpretation. Arguably, the most widely adopted guidelines are those of the American College Medical Genetics (ACMG; Richards et al. 2015; Kleinberger et al. 2016). These guidelines aim to consider the different evidence that can support classification, helping to allocate each variant in the correct group according to its likelihood of pathogenicity. Examples of evidence are (i) the effect of the variants on the transcript, with premature stop codon predicted to be one of the most pathogenic (ii) functional studies supporting the aetiological role of the variants (iii) an odds ratio greater than five to have the related disease (Richards et al. 2015). Also, in the UK the ACMG guidelines are widely applied, and a similar set of criteria have also been suggested by the UK Association for Clinical Genomics Science (Association for Clinical Genomic Science website, accessed in September 2021)

In summary, the identification of genes and variants causally implicated in inherited diseases has laid the foundation for clinical genetics. Defining the genetic and molecular basis is an essential first step in disease diagnosis, administration of the most effective treatment and prognostication. For example, these studies contributed to the development of treatments for rare diseases, such as gene therapy for haemophilia or allogeneic bone marrow transplant for WAS (Burroughs et al. 2020; Pasi et al. 2020). Moreover, genetics and Mendelian disorder studies have also informed drug discovery pipelines and contributed to drug target prioritisation, drug repurposing, the definition of the therapeutic windows and estimation of side effects (Plenge et al. 2013; Nelson et al. 2015; King et al. 2019). For instance, molecular observations on familial hypercholesterolemia have identified the *PCSK9* gene as the cause of this condition and further studies brought to the conceptualisation and approval of *PCSK9* inhibitor to treat high cholesterol levels (Abifadel et al. 2003; Sabatine et al. 2017).

1.4 Methods to study human genetics and epigenetics.

The healthcare advancements described at the end of the previous sub-chapter will only be possible because of the technological developments that have distinguished the beginning of the 21st century, namely high-throughput genotyping, CRISPR/Cas9 and all

their derived applications and resources (Jinek et al. 2012; Reuter, Spacek, and Snyder 2015; Slatko, Gardner, and Ausubel 2018; Nakamura et al. 2021). Genotyping and sequencing technologies allowed the scientific community to study genomic variants in hundreds of thousands of individuals, generating new insights into human biology (Gaziano et al. 2016; Di Angelantonio et al. 2017; Bycroft et al. 2018; Turro et al. 2020b; Thaventhiran et al. 2020). Meanwhile, CRISPR/Cas9 technology made possible the modification of specific genomic and epigenomic regions in an easy and scalable fashion, virtually in every cell type (Adli 2018; Nakamura et al. 2021).

1.4.1 Models systems

In biomedical sciences, the availability of human biological materials is often limited and researchers rely on the use of animals or cell lines as models. The most common animal model for human biology is the mouse which has biology relatively similar to the human one (Bedell et al. 1997). There are many successful examples where knock-out mice recapitulate well the human pathophysiology of inherited disorders. For instance, for inherited platelets disorders, knock-out models have been successful for Glanzmann thrombasthenia (Morgan et al. 2010), Bernard and Soulier syndrome (Strassel et al. 2007), and more recently for the grey platelet syndrome (Mayer et al. 2018), amongst many other inherited BTPDs.

However, there are many discordances between human and mouse biology. For instance, humans, but not mice, are haploinsufficient for many developmental genes (Bedell et al. 1997; Wilkie 2003). Moreover, the ENCODE project showed that a large portion of the regulatory elements in the human genome are evolutionary recent and cannot be identified by sequence homology between humans and mice (ENCODE Project Consortium et al. 2020).

Hence, stable cell lines are a valid alternative to study human biology. For instance, the immortalized megakaryocyte cell line from Nakamura and colleagues (Nakamura et al. 2014) allows access to an unlimited source of biological material that resembles megakaryocytes (MK). In general, cellular models are convenient for experiments that require many cells as starting material or that need a system less prone to inter-personal genetic variability. However, cell lines tend to acquire chromosomal aberration which may alter the genetics and epigenetics of the cells. For instance, translocation may compromise or rewire the physiological regulatory loops happening in the parental cell type.

A particularly interesting cell model is the human-induced pluripotent stem cells (hiPSC; Takahashi and Yamanaka 2006; Takahashi et al. 2008). Specifically for the research in MK and platelet biology, it is now feasible to generate MKs in large numbers by forward programming of human iPSCs (see chapter 2.2; Moreau et al. 2016). Ideally, one could generate hiPSC from patients and the differentiation into the relevant cell type (Laugsch et al. 2019). Initiatives such as HipSci (https://www.hipsci.org/) created an hiPSC biobank from hundreds of healthy and rare disease volunteers (Kilpinen et al. 2017). This method allows working directly on the original genomic background that is causing the disease, without the necessity of genome editing that may create some biological noise due to off-target effects of the CRISPR-Cas9 technology (see chapter 1.4.2). A successful example comes from the laboratory of Rada-Iglesias (Laugsch et al. 2019). The authors of this manuscript used iPSC derived from a patient and their *in vitro* differentiation to study the orofacial clefting aetiology and identified the cause in a SV that altered a regulatory loop (Laugsch et al. 2019).

1.4.2 Genome-wide association studies, effect size and polygenic risk scores

Genome-wide association studies (GWAS) are a hypothesis-free statistical approach to determine the association between genetic variants and one or more categorical or quantitative traits. GWAS have been widely used since 2002, as they facilitate the study of complex traits, such as blood cell count (Ozaki et al. 2002; Bush and Moore 2012; Astle et al. 2016b; Vuckovic et al. 2020; WTCCC 2007), and became one of the most adopted tools to improve biological understanding of defined traits (Lichou and Trynka 2020). Nowadays, GWAS have identified thousands of common SNPs associated with human traits, and haematology is one of the biomedical fields in which GWAS application has been most successful (Soranzo et al. 2009; van der Harst et al. 2012; Shin et al. 2014; Astle et al. 2016; Astle et al. 2016; Buniello et al. 2019; Vuckovic et al. 2020; Chen et al. 2020).

GWAS design can investigate either categorical or quantitative traits (Bush and Moore 2012). The former often refers to binary variables, because the sampled population is divided into two groups (e.g. phenotype present versus absent) and the two groups are interrogated for their genetic differences. The latter is applicable to all traits that are continuous variables, such as weight, height or blood cell counts (Bush and Moore 2012). The advantage of quantitative traits is that they do not have to be grouped, hence less prone to mitigate the differences because of the grouping strategies and therefore resulting in

increased statistical power (Bush and Moore 2012). A few factors may cause spurious SNP associations in GWAS (i.e. variants correlated with the phenotype for technical reasons), for example, consanguinity in the cohort or ethnic stratification (Zeng et al. 2015). Hence, sample size and genomic architecture, which lie at the base of statistical power, have to be well-calibrated, balancing costs and power (Bush et al. 2012; Ball et al. 2013, Astle et al 2106a).

GWAS identify the association between single genetic variants and a particular phenotype. As a consequence of linkage disequilibrium usually, a large number of associated variants is identified at a certain locus. Conditional analysis can be applied to differentiate between 'bystander associated variants' and the variants which are causally associated (Vuckovic, Cell 2016). Mapping causally associated variants to genes, protein and pathways remain one of the key challenges in the post-GWAS era (Lichou and Trynka 2020; Cano-Gamez and Trynka 2020).

This is further complicated by the fact that over 90% of associated variants are localised in the non-coding region of the genome (Astle et al. 2016; Vuckovic et al. 2020). While non-synonymous SNPs in the coding space are readily mapped to protein and pathways, this is not the case for non-coding variants (Peters et al. 2017; Cano-Gamez and Trynka 2020). One of the most used computational approaches to assign variants to genes is colocalisation (Wallace 2020), where GWAS results are co-analysed with the results of expression quantitative trait loci (eQTL) studies (Wallace 2020). An alternative approach is to use the insight gained from experiments that identify long-range interaction between regulatory elements and gene promoters (e.g. promoter capture Hi-C, see chapter 1.4.3; Javierre et al. 2016; Sati and Cavalli 2017). Ultimately, some tools combine multiple layers of evidence, both computational and experimental, for instance, the "Locus-2-Gene" tool from the Open Targets Consortium (Carvalho-Silva et al. 2019).

GWAS results also provide information on the contribution of every single associated variant to the phenotype. The contribution is expressed in the form of odds ratio (OR) for discrete traits, or beta coefficient (β) for continuous traits. These parameters are generally referred to as effect sizes. The effect size of common variants (MAF ≥ 0.05) is in the vast majority of cases moderate (OR ~ 1.2 and β ~ 0.05, Fig. 1.4; Goldstein 2009; López-Cortegano and Caballero 2019), meaning that each variant independently makes a small contribution to the phenotype (Vuckovic et al. 2020). In contrast, the effect sizes of rare

variants (MAF < 0.01) tend to have larger effect sizes (OR>1 and β>1; Bomba, Walter, and Soranzo 2017).

To increase the statistical power of GWAS, several imputation methods have been developed to infer the role and association of rare variants, which have remained unmeasured by array genotyping (Li et al. 2009; Marchini and Howie 2010; Bycroft et al. 2018). However, imputation is not reliable to infer the genotype of variants with a MAF < 0.001 (Van Hout et al. 2020).

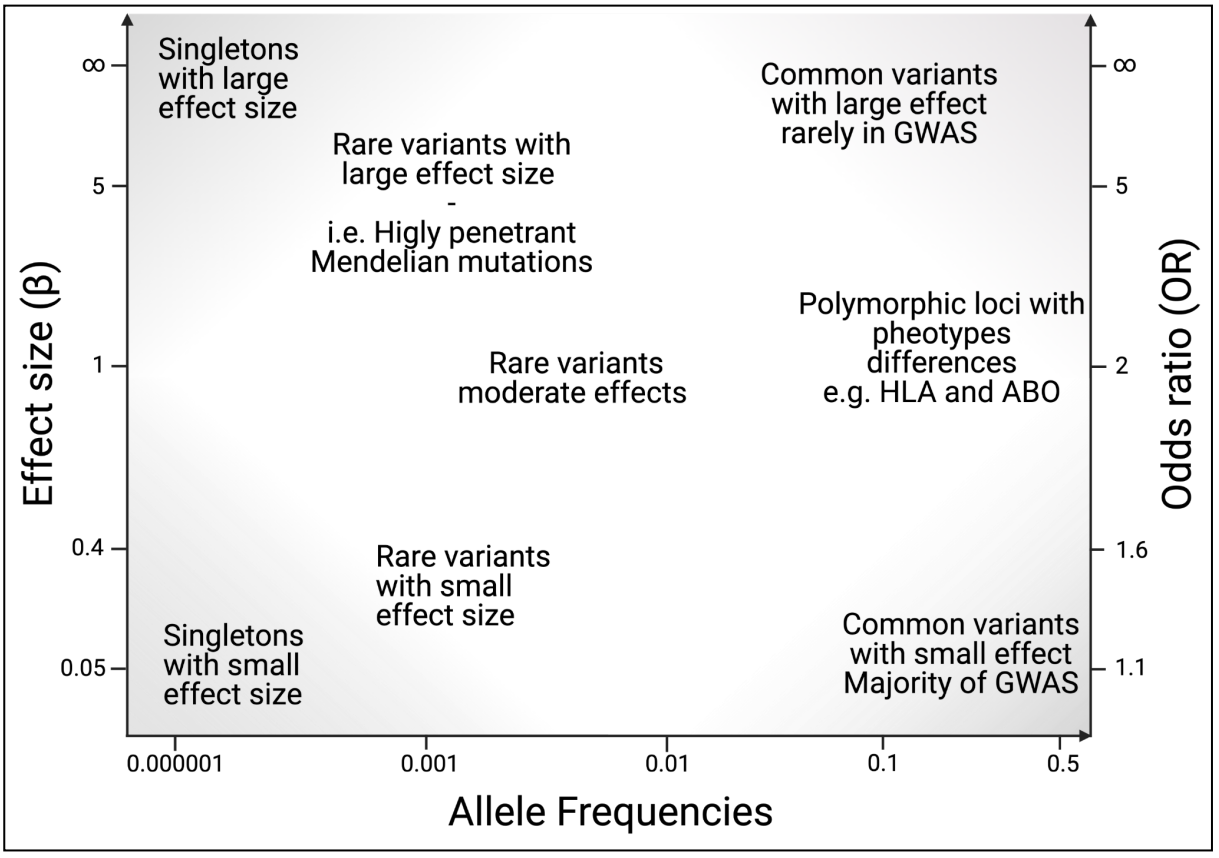


Fig. 1.4 | Cartoon illustrating the correlation between allele frequency and effect size/Odds ratio of human variants. *Adapted from Manolio et al. 2009.*

Rare and ultra-rare variants (MAF < 0.001) with large effect sizes on medically relevant traits are of particular interest in medicine, as has been demonstrated by the study of Mendelian disorders (Manolio et al. 2009; Long et al. 2017; Turro et al. 2020; Thaventhiran et al. 2020). These deleterious variants with large effect sizes (Fig. 1.4, top-left corner) are confined to rare allele frequencies because of the negative selective pressure (McCarthy et al. 2008; Manolio et al. 2009). Indeed, for FBC parameters, the effect size for variants with a MAF > 0.01 does not exceed 0.4 standard deviations (sd), whilst associated rare variants have effect sizes up to 0.86 SD (Vuckovic et al. 2020).

Leveraging the recent GWAS results from large cohort studies, biomedical scientists reliably calculated for the first time a single predictive value for a given phenotype, namely polygenic risk score (PRS). Briefly, the PRS uses the sum of all the variant effects to forecast the predisposition or protection to a certain trait (Khera et al. 2018). In some cases, the sum of the effect sizes of common variants is comparable to the effect of one single rare (e.g. Mendelian disease) variant with a large effect (Fahed et al. 2020). PRS have been calculated for many common diseases, such as VTE and cancer (Kachuri et al. 2020; Klarin et al. 2019), and for FBC traits (Vuckovic et al. 2020). To give an example of the role and use of PRS, Vuckovic and colleagues reported that about 30% of the mpv can be explained from the set of SNPs used for the calculation of the PRS (Vuckovic et al. 2020). It is hoped that, in the future, PRS will become clinically relevant for providing targeted public health measures to those who are most at risk. However, further pieces of evidence are needed to determine the clinical value of PRS (Eisenstein 2021). Collins et al. have illustrated how the PRS for platelet count modifies the effect of a Mendelian rare variant (Collins et al. 2021). Variant rs5030764 is a pathogenic variant in *GP9* and if present on both alleles is causal of BSS (see chapter 1.2.3). It was assumed that carriers of this variant do not present a clinically relevant reduction in platelet count, however several individuals carrying this variant present a platelet count < 150×10^9 platelet/L. The platelet count was correlated to PRS, people who are in the lowest decile tend to have the lowest plt. If not considering PRS, such genetic configuration may be misdiagnosed as a case of immune thrombocytopenia.

1.4.3 Sequencing technologies

Sequencing DNA molecules became readily possible in the late 1970s (Sanger, Nicklen, and Coulson 1977), since then Sanger sequencing has been the “gold standard” sequencing technology. It underwent several improvements in the fluorescent dideoxynucleotides and electrophoresis technology, (Sanger 1988; Kheterpal et al. 1996; Rosenblum et al. 1997) driving the completion of the human genome project (Collins et al. 2003; McPherson et al. 2001). However, the turning point in sequencing technologies happened in 2008, when the sequencing-by-synthesis technology by Solexa became available (Bentley et al. 2008). This scalable sequencing technology resulted critical for the efforts to decipher the whole human genome in a large number of individuals (NIH - The cost of sequencing a human genome webpage, accessed September 2021). Indeed, the latest Illumina instrument (i.e. NovaSeq 6000), with the largest flow cell (i.e. S4) allows multiplexing 48 human genomes (30x coverage), 500 exomes (100x coverage) and about

400 transcriptomes (50 million reads; Illumina website, accessed in September 2021). There is also the 3rd generation of sequencing technologies, such as Oxford Nanopore and Pacific Biosciences, which allow the sequencing of long strings of DNA (Niedringhaus et al. 2011).

This technological development led to the establishment of a series of sequencing-based approaches that are used to study the genome, epigenome and transcriptome (see chapter 1.3.1 and 1.3.2).

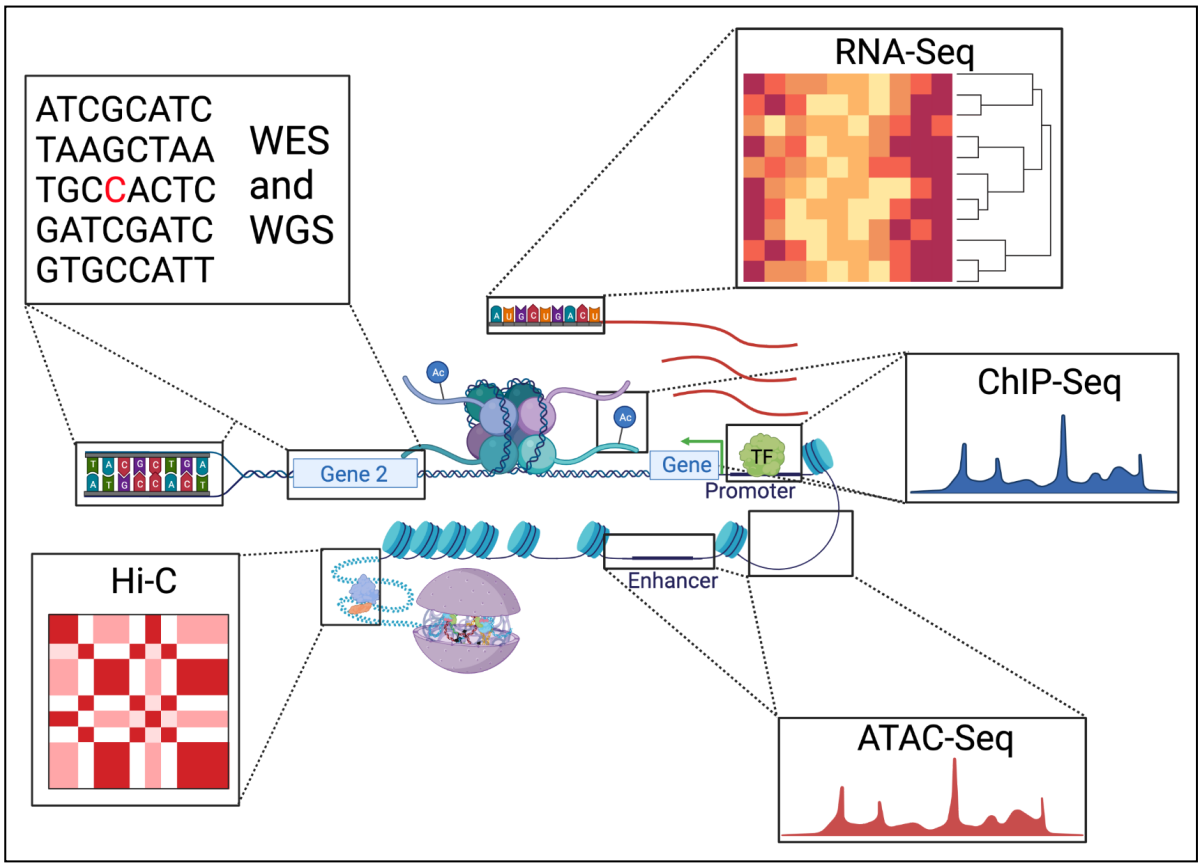


Fig. 1.5 | Representation of the technologies based on sequencing and their standard output after analysis.

High-throughput sequencing (HTS) of the genome

Human genome sequencing has two main high-throughput approaches: targeted sequencing and genome-wide. Targeted approaches allow the sequencing of defined regions of interest via the use of a set of capture probes. These probes can enrich for a group of regions (or genes) that have a high probability of being involved in one or more diseases. The ThromboGenomics (TG) HTS test was a research-diagnostic platform that used this targeted approach to sequence all the diagnostic-grade genes involved in BTPDs (Simeoni et al. 2016; Downes et al. 2019). In a more agnostic way, one could look at the

whole exome. Whole-exome sequencing (WES) covers most of the coding sequence including the mRNA untranslated regions (UTR). With WES, the amount of target DNA sequenced is approximately 2% of the entire genome (~40 Mb; Clark et al. 2011).

However, the capture efficiency varies on the hybridization efficiency of the utilised probes, which are often depending on the complexity of the DNA sequence. For instance, GC or AC rich regions tend to have worse capture efficiency (Dohm et al. 2008). This hybridization variability may introduce some bias in the downstream analyses and interpretation (Clark et al. 2011). Moreover, in spite of being a cost-effective resource for the identification of the variants in protein-coding regions, WES overlooks the genetic material of ~98% of the human genome and this limitation is magnified in approaches that focus on smaller panels. The reduction in the sequencing cost and the increased sequencing depth (30x coverage is sufficient to recall rare variants with good confidence; Turro et al. 2020) are shifting the approach of choice to whole-genome sequencing (WGS), in order to achieve the greatest amount of information per sequencing run, including non-coding regions (see chapter 1.3).

Transcriptome sequencing

Similarly to DNA, also the RNA within a cell (i.e. transcriptome) can be sequenced (Zhong Wang, Gerstein, and Snyder 2009). RNA-Seq allows obtaining quantitative information on the expression of the entire transcriptome of a population of cells (ribosomal RNA, transfer RNA and messenger RNA) via retrotranscription to complementary DNA (i.e. cDNA; Mortazavi et al. 2008). RNA-Seq allows to identify cell-type specific patterns of gene expression and to look at differentially expressed genes between two types of cells to gain insight into their biological differences (Love, Huber, and Anders 2014; Turro, Astle, and Tavaré 2014).

Further technological advances enabled the sequence of the RNA present within a single cell, which gives a much finer capacity to discriminate subpopulations within the same cell type and it is being used to build a human single-cell atlas (Tang et al. 2009; Papatheodorou et al. 2020).

Sequencing technologies to study the chromatin state and structure

Chromatin immunoprecipitation followed by high-throughput DNA sequencing (i.e. ChIP-Seq) is used to study the distribution of proteins that interact with DNA in the genome. See Table 1.6 for a list of the most common histone modifications (Barski et al. 2007). ChIP-Seq returns information on the role of DNA regions, while indirectly reporting the

position of the proteins and their post-translational modifications because of the antibody precipitation step (Barski et al. 2007). For this reason, ChIP-Seq is also used to study transcription factor binding site specificity (Landt et al. 2012).

ChIP-Seq signal is measured as a signal DNA over the noise DNA (i.e. control). The signal is given by the DNA that was linked to a protein and precipitated because of the antibody-specific immunoprecipitation. ChIP-Seq controls are: (i) “input” DNA sonicated but not immunoprecipitated, or (ii) “IgG”, a mock antibody that does not have any specificity towards nuclear proteins (Landt et al. 2012).

The assay for transposase-accessible chromatin using high-throughput sequencing (i.e. ATAC-Seq; Buenrostro et al. 2013) is a parallel approach that is used to study open chromatin regions. The hyperactive Tn5 transposase inserts sequencing adapters in the genome, and it has a higher chance to insert them in regions where the DNA is not wrapped around histones (Buenrostro et al. 2013). A genomic region, to be functionally active, needs to be relatively free of histones, in order to give access to transcription factor motifs and allow protein-specific binding. For this reason, ATAC-Seq is thought to be more sensitive to cell dynamics, because chromatin accessibility is preceding transcription (Nimmo, May, and Enver 2015).

The 3D structure of the genome has a central role in cell function, influencing cell development, cell differentiation and gene expression (Schoenfelder and Fraser 2019; Zheng and Xie 2019; van Steensel and Furlong 2019). Chromosome conformation capture (3C) techniques have shown great potential to investigate the role of non-coding DNA. Amongst all the 3C derived assays, Hi-C, which looks at genome-wide interactions, has been widely adopted (Sati and Cavalli 2017; McCord, Kaplan, and Giorgetti 2020). Briefly, Hi-C consists of an enzymatic digestion of the genome, followed by a ligation to create chimeric DNA sequences of regions that are in close proximity when chromatin is folded into its structure (see chapter 2.3 for a more detailed description of the technique; Dekker, Marti-Renom, and Mirny 2013). Hi-C technology has contributed to the identification of TADs (see chapter 1.3.5), genomic structures which have also been confirmed more recently with microscopy experiments (Bintu et al. 2018).

The complexity of the Hi-C libraries (i.e. the number of different DNA fragments that are in a library) limits the experiment analysis resolution and the capacity to identify sub-TAD regulatory loops (Mifsud et al. 2015; Schoenfelder et al. 2015). To bypass this problem, Promoter capture Hi-C (or pcHi-C) uses probes to capture the regions of interest (i.e.

promoters), reducing the complexity of the libraries and increasing the information relative to regulatory loops (Mifsud et al. 2015; Schoenfelder et al. 2015; Javierre et al. 2016).

1.4.4 Clustered regularly interspaced short palindromic repeats (CRISPR)-Cas9 tools

The idea of surgical precision in gene editing has been around at least since the late 1980s (Thomas and Capecchi 1986; Capecchi 1989). Research into genome editing drove the investigations and discoveries of basic biological processes, such as homologous recombination and non-homologous end-joining mechanisms to repair DNA breaks (Jeggo 1998). It also pushed the boundaries for biotechnology invention, with the finding of meganucleases (i.e. restriction enzymes with recognition sites between 14 and 40 bp; Adli 2018), zinc fingers (Klug and Rhodes 1987), transcription activator-like effector nucleases (Joung and Sander 2013) and finally the CRISPR-Cas9 (Jinek et al. 2012).

CRISPR-Cas9 is an RNA-guided endonuclease, discovered because of its role in the defence from pathogens invasion in *Streptococcus pyogenes* (Marraffini 2015). The two molecular components required for the correct functioning of the CRISPR-Cas9 system are an RNA molecule, often called single guide RNA (sgRNA), and the endonuclease Cas9. The capacity to specifically direct the endonuclease in virtually any locus, just designing a 20-nucleotide long oligonucleotide that is complementary to the target, made the CRISPR-Cas9 system the widespread genome-editing tool that it is today.

CRISPR derived technologies also allow multiplexing (McCarty et al. 2020). It is therefore possible to assess, in the same experiment, the effect of perturbation of several regions at once, allowing combinatorial and additive testing of multiple loci. However, further studies are needed to understand and limit the deleterious effects of multiple testing, such as exogenous component toxicities and off-target effects (Morgens et al. 2017; Tycko et al. 2019). Indeed, off-target effects are still one of the main concerns for the designing of the CRISPR-Cas9 experiments and that are holding back most of its applicability in healthcare (Cho et al. 2014; Naeem et al. 2020). However, approaches such as base editors, which do not require DNA strand breaks to edit the genome, can accelerate the adoption of gene therapies in clinical practice (Gaudelli et al. 2017; Grünewald et al. 2019; Kurt et al. 2021).

Since its first application in 2012, CRISPR-Cas9 has been re-engineered in different flavours to edit DNA and epigenome (Jinek et al. 2012), but also to act as molecular probes for microscopy (Chen et al. 2013), biotin tag (Liu et al. 2017) and sequencing enrichment (Gilpatrick et al. 2020).

CRISPR interference

CRISPR interference (CRISPRi) is an adaptation of the CRISPR technology. CRPISPRi takes advantage of an enzymatically inactive form of Cas9 conjugated to protein domains to epigenetically repress chromosomal regions, usually with the aim of understanding the role of a region on gene expression (Gilbert et al. 2013). Part of the repression effect comes from Cas9 steric interference. When the Cas9 binds to a specific locus on the DNA, it competes with transcription factors and reduces their DNA binding, decreasing gene transcription (Gilbert et al. 2013). Furthermore, the Cas9-conjugated repressor domains, often the Krüppel associated box (KRAB; Larson et al. 2013; Gilbert et al. 2013), recruit chromatin remodelers to edit the epigenetics of the locus and silence it (Margolin et al. 1994; Urrutia 2003; Huntley et al. 2006).

Several protein domains have had their repressor activity tested in the CRISPRi system, namely chromo shadow, DNMT3A, LSD1, WRPW domains, however, KRAB showed the strongest repression activity (Gilbert et al. 2013; Kearns et al. 2015; Amabile et al. 2016). The CRISPRi based on the KRAB domain is able to repress up to 99% of the expression (Larson et al. 2013). There is an evident variability that depends on the regions and genes that are targeted (Yeo et al. 2018; McCarty et al. 2020). CRISPRi has been successfully adopted to repress gene expression via the epigenetic editing of proximal regulatory regions, such as promoters (Gilbert et al. 2013) and, with equal success, long-distance regulatory elements (i.e. enhancers, Thakore et al. 2015; Klann et al. 2017). Recently, independent studies optimised the CRISPRi system to make it more effective in the silencing of regulatory regions, via the addition of multiple repressing domains or proteins, in order to activate multiple distinct repressing pathways (Yeo et al. 2018; Li et al. 2020).

1.5 Databases and resources

The technological advances described in chapters 1.4.2 and 1.4.3 allowed the conceptualisation and completion of several large studies that collected phenotype, genotype, epigenome and transcriptome from hundreds of thousands of people and counting. This has prompted the development of large cohort studies to define the relationship between sequence variation and phenotype. This chapter describes some of these resources.

1.5.1 Biobanks

Biobanks are repositories of biological specimens (e.g. HipSci) or biological data (e.g. UK Biobank; Hewitt and Watson 2013). Originally, they have been thought to preserve biodiversity (e.g. Svalbard Global Seed Vault); but they quickly became an instrument to improve healthcare and personalised medicine (Hopkin 2008; Labant 2012; Hewitt and Watson 2013). The storage, use and distribution of biological samples and data also contributed to the development of stringent regulations on data privacy, data use and ethics (Cambon-Thomsen, Rial-Sebbag, and Knoppers 2007; Rothmayr 2009; Spagnolo, Daloiso, and Parente 2011).

The largest and best-characterised population cohort for biomedical studies is the UK Biobank (UKB; https://www.ukbiobank.ac.uk/). UKB is a prospective study that collected genomic data of half a million people and linked them to electronic health records (EHR) in order to understand the causes of common disorders with the aim of preventing disease onset and finding new treatments. Between 2006 and 2010 ~5 million people were approached with an invite to participate and 500,000 were enrolled across an age range from 40 to 69 years (at the time of enrolment; Munafò et al. 2018; Bycroft et al. 2018). Although UKB participant sampling is not completely unbiased, it is still the best representation available of an archetypal cohort of the UK population (Munafò et al. 2018). UKB has been instrumental in defining the genetic architecture of most common diseases and of biomedical relevant traits (Bycroft et al. 2018; Cortes et al. 2020; Vuckovic et al. 2020). In 2020 the WES data for 50,000 participants have been released (Van Hout et al. 2020) and during 2021 the WES data for all participants has been made available to a limited number of pharmaceutical companies and 125,000 participants for academic institutions (Wang et al. 2021).

A resource similar to UKB for size (830,000 participants) and scope is the Million Veteran Project (MVP; Gaziano et al. 2016). Recently, the largest meta-analysis study (746,667 Individuals) was published to gain insight into the different human haematological traits and, more importantly, in different ethnic groups using data mainly coming from UKB and MVP (Chen et al. 2020). There is an increasing awareness of the importance to further increase the ethnic diversity of biobanks and more initiatives have been taken recently to achieve this building more non-European cohorts, for instance, Three Million African Genomes, and GenomeAsia100K (GenomeAsia100K Consortium 2019; Wonkam 2021).

A limitation of UKB and MVP is that participants cannot be recalled based on genotype to investigate the association between genotype and phenotype in further detail. For this purpose, the NIHR BioResoure (NBR-BR) was founded in 2007 by professors Todd and Ouwehand (Cambridge Bioresource at the time). The NBR-BR has now 13 centres that create a national network with a repository for biosamples, genomic and clinical data ("NIHR BioResource Home Page"). Its main purpose is to provide researchers with access to a cohort to perform recall by genotype studies. At the moment, NIHR-BR has recruited more than 200,000 participants that have been genotyped with the UKB genome-wide array or WGS. One of the first publications originating from the NIHR-BR piloted the functional validation of a GWAS hit in humans (Dendrou et al. 2009). Indeed, Dendrou and colleagues validated the association between IL2RA and type 1 diabetes via the regulation of the expression of CDC25 (Dendrou et al. 2009). Since then, NIHR-BR has given an outstanding contribution to research with over 200 publications reporting on the association between genotype and phenotype (Turro et al. 2020b; Vuckovic et al. 2020; Thaventhiran et al. 2020; Collier et al. 2021; Stephenson et al. 2021; Gräf et al. 2018; Wei et al. 2019).

The NIHR-BR also piloted the rare diseases arm of the 100,000 Genomes Project (100KGP, 2014). The primary purpose of the 100KGP was to introduce WGS into the frontline of NHS care, initially for rare diseases and several cancers. The 100KGP has many unique features with one being that the genotype and phenotype data are available for research and the number of patient samples which will have been analysed by 2024 will have increased to 0.5 million.

Other resources of genomic data linked to the NHS EHR data are the blood donor health studies INTERVAL (Di Angelantonio et al. 2017), COMPARE (Bell et al. 2021) and STRIDES. These trials, which are delivered by a partnership between the University of Cambridge, NIHR-BR and NHS Blood and Transplant (NHSBT) have as their primary aim to inform improvements in the care of blood donors but, as is the case for the 100KGP, the genotype and phenotype data are available for research.

One more important resource is the genome aggregation database (gnomAD; Karczewski et al. 2020). This resource is the aggregation of genomic information on 201,904 individuals (125,748 WES + 76,156 WGS) from different ethnic groups and coming from more than 50 different projects (e.g. BioMe and TOPMed). Despite being a first-class resource for allele frequencies, gnomAD lacks phenotype information of the participants, limiting the use of the resource in biomedical research. However, some of the cohort studies

that are aggregated in gnomAD are on the cusp of releasing their genotype and phenotype data at the level of the single participants (e.g. TOPMed; Taliun et al. 2021).

There are also ambitious studies that are about to start, namely Our Future Health (https://ourfuturehealth.org.uk/) and All of Us (https://allofus.nih.gov/). These two projects aim to genotype up to 6 million people (5 million Our Future Health and 1 million All of Us) of different ethnic groups and backgrounds. Both projects have the goal to assess the value of the PRS in clinical settings and the benefit of early intervention on people having a higher genetic risk for a particular disease.

1.5.2 Phenotype ontologies

Biobanks also require hierarchical controlled dictionaries (i.e. ontologies) to communicate phenotypes. Indeed, while variants are relatively easy to describe because they are discrete and unambiguous, human phenotypes, especially diseases, have a series of possible interpretations and variability that comes from the nomenclature, anatomy or the severity of the phenotype.

The MYH9-related disease (MYH9-RD) is a good example. For historical reasons, MYH9-RD is also called Epstein syndrome, Fechtner syndrome, May-Hegglin anomaly and Sebastian syndrome. Its phenotypes are characterised by, but not necessarily all, thrombocytopenia, hearing loss, presenile cataracts, the elevation of liver enzymes, and nephropathy with a range of gravity from mild to severe. The age of onset for MYH9-RD spans from neonatal to late adulthood. All this variability limits the statistical power in association analysis, because it may lead to the creation of different groups describing the same observation or, even worse, aggregating different traits in the same one. For this reason, biomedical scientists created a series of codes that describe diseases, phenotypes and their relations unequivocally.

OMIM (https://www.omim.org; Hamosh 2002) was the first attempt to create a unique repository to define univocally human diseases. This effort started in the 1960s, by the work of Victor McKusick (i.e. Mendelian Inheritance in Man, MIM), moved online in the 1980s and was renamed OMIM (online MIM). Nowadays it contains 4,690 human genes associated with almost 6,000 traits including Mendelian, complex and somatic disorders (OMIM website accessed in September 2021).

Another human ontology, more comprehensive than OMIM, is the Human Phenotype Ontology (HPO; Robinson et al. 2008). The HPO annotation was started in 2008 and today

contains more than 15,000 terms that describe relations between phenotypes and genes (Köhler et al. 2021). This resource provides a tool to analyse the phenotype of the patients in a more statistically and computationally efficient way than free text (Turro et al. 2020). Similarly, Experimental factor ontology (EFO; Malone et al. 2010) is the European Bioinformatics Institute (EMBL-EBI) holistic attempt to describe all the experimental variables, chemical compounds, phenotypes and diseases that are related to molecular biology.

International Statistical Classification of Diseases and Related Health Problems (ICD) is not an ontology, and it was developed at the beginning of the 20th century to be used in medical practice and to describe anatomical sites. Since then, this code has been expanded and updated, in January 2022 the eleventh version of the ICD codes will be officially adopted in clinical practice ("Classification of Diseases; ICD"). ICD coding entered the world of biomedical research because large cohorts started to integrate the phenotype of their participants with hospital episode statistics (HES) and EHRs, some of which are based on ICD codes.

1.5.3 Pathogenic Variants

Most of the large-scale rare diseases genomic studies aim to identify new genes implicated in rare diseases and to catalogue the pathogenicity level of rare variants in these genes (Turro et al. 2020; Thaventhiran et al. 2020). Having this information organised in a single repository is convenient for genomic medicine and variant clinical interpretation.

ClinVar is one such public repository curated by the National Centre for Biotechnology Information (NCBI) that reports the relationship between human variants and phenotypes (NCBI website accessed in September 2021). Each variant has information on the evidence and interpretation of the variant and also the degree of confidence of the interpretation. Similar structure and Similar structure and scope are characterising the Human Gene Mutation Database (HGMD®), which is a proprietary resource of human pathogenic variants curated from the literature (HGMD® website).

There are also haemostasis specific databases, which have a superior level of curation and annotation of variants. An example is the one curated by the European Association for Haemophilia and Allied Disorders (EAHAD) which only reports variants occurring in *F5*, *F7*, *F8*, *F9* and *VWF* genes.

1.5.4 Transcription and Epigenetics

High-throughput sequencing studies have also produced a large amount of information on transcriptome and epigenome (ENCODE Project Consortium 2012; GTEx Consortium 2013; The GTEx Consortium 2015; Schultz et al. 2015; Skipper et al. 2015; Stunnenberg et al. 2016; Davis et al. 2018). Collecting information from transcriptome and epigenome is more challenging than genome because these biological features are cell-type specific. Moreover, in epigenomes many different modifications co-occur (e.g. acetylation or methylation) and evolve over the lifespan of the same person (Hernando-Herraez et al. 2019; Peleg et al. 2016).

Expression Atlas (https://www.ebi.ac.uk/gxa/home) is a resource curated by the EMBL-EBI that reports gene expression levels organised by tissues. Expression Atlas integrates, collates and reanalyses the information from several resources, using more than 3,000 independent experiments (EBI website accessed in September 2021). A resource, present within Expression Atlas, is GTEx (GTEx website accessed in September 2021). GTEx not only reports gene expression in different tissues but also correlates them with nucleotide polymorphisms, creating the most comprehensive repository of expression quantitative trait loci (eQTL) data.

The International Human Epigenome Consortium (IHEC; Stunnenberg et al. 2016, Hirst 2016) is an international research effort that aims to get information on 1,000 epigenomes (Eurice GmbH; http://ihec-epigenomes.org/). Also, this consortium is the result of the aggregation of several independent studies. Amongst many that are part of the IHEC, four have had great resonance in the scientific community: RoadMap (started in 2007; Bernstein et al. 2010), BLUEPRINT (started in 2011; Adams et al. 2012; Stunnenberg et al. 2016), 4D Nucleome (Dekker et al. 2017) and Encyclopedia of DNA element (ENCODE) which aims to collect information on all the different level of epigenetic information (started in 2003 and today at its 5th version; ENCODE Project Consortium et al. 2007, 2020). The goal of all these studies is to define the relationship between epigenetic modifications, human biology and ultimately human diseases.

1.6 Project aims

Despite the improvements brought about by high throughput sequencing, no conclusive diagnosis can be defined for ~50% of the referrals for a possible diagnosis of

BPTDs analysed by the TG HTS test (Downes et al. 2019). A part of the referrals may have concerned non-inherited conditions but in other cases, the underlying variants remain to be identified. The aim of this PhD project was to investigate further the genetic aetiology of inherited BTPDs, with a particular focus on the non-coding space as a possible harbour of causative variants.

First, I estimated the contribution of rare coding variants to BTPDs and their interplay with common variants. The rationale behind this study is: (i) pathogenic and likely pathogenic variants (P/LP) effect sizes may have been overestimated for a portion of these variants; (ii) some rare phenotypes may be caused by the interplay of rare and common variants. I investigated the contribution of common variants to the BTPD aetiology by using the relevant PRS. This workflow also allowed me to (i) look into the biology of PRS and its mapping onto pathways relevant for BTPDs and (ii) inspect the phenotype of autosomal recessive pathogenic-variant carriers and challenge the acceptance that these individuals have no discernible clinical phenotypes.

Second, I explored the role of non-coding DNA in the onset of BTPDs. My working hypothesis is that part of the undiagnosed BTPDs can be explained by rare non-coding variants that alter the transcription of key haemostasis genes. I experimentally derived high-resolution DNA interaction maps for the diagnostic-grade BTPD genes in HEP, EC and MK. I applied these interactions to define cell type-specific regulatory regions. I used these newly defined regulatory regions to interrogate the genome of individuals with BTPDs lacking clear genetic diagnosis, searching for possible explanatory variants. Ultimately, I performed *in vitro* validation for some of these regions that showed a high probability of having a regulatory role and that had rare variants identified in people with inherited BTPDs.

Chapter 2

Materials and methods

Chapter 2.1 Computational methods for the VarioPath project

Ethics

All the work that is related to chapter 3 has been performed under the UK Biobank application ID 13745. While the NIHR BioResource – Rare Diseases work (chapter 4 and chapter 5), such as accessing the biological samples to PCR the DNA and analysing the genotype and phenotype data, was performed under the REC Number 13/EE/0325.

Definition of the list of pathogenic variants

A list of variants was obtained from the major resources described below and filtered to retain only the variants with high likelihood of being pathogenic or likely pathogenic (P/LP). From HGMD Pro (v2019.4; http://www.hgmd.cf.ac.uk/), I extracted disease-causing (DM) or questionable disease-causing (DM?) variants. From ClinVar “pathogenic” or “likely pathogenic” variants (https://www-ncbi-nlm-nih-gov.ezp.lib.cam.ac.uk/clinvar/). I also complemented these large resources with in-house curated variants (Sivapalaratnam et al. 2017; Turro et al. 2020b; Thaventhiran et al. 2020). These manually curated pathogenic variants have been, mainly, obtained via the NIHR BioResource rare diseases projects. The variant coordinates used in the VarioPath project were in GRCh38, and if not, they were converted to this latest assembly using the online tool “AssemblyConverter” (https://www.ensembl.org/Homo_sapiens/Tools/AssemblyConverter, Ensembl v.100; Zhao et al. 2014). The final list of P/LP variants had 299,632 unique entries.

Genes and transcript adopted in the VarioPath studies

A list of 4,881 unique genes was compiled in order to collect the Mendelian disease-causing genes. The list was based on the union of (i) “green-flag” genes from Genomic England Panel App (Martin et al. 2019; https://panelapp.genomicsengland.co.uk/); (ii) OMIM (Amberger et al. 2019); (iii) OrphaNet (Rath et al. 2012); (iv) NBR diagnostic gene list (Simeoni et al. 2019; Megy et al. 2019; Turro et al. 2020b; Gleadall et al. 2020). All the mentioned resources have been accessed or downloaded on the 28th April 2020. Throughout this thesis I used the latest HGNC symbols and stable identifiers (Tweedie et al. 2021). The transcripts adopted for each coding gene were selected in order of priority as follows:

Matched Annotation from NCBI and EMBL-EBI (MANE; https://www-ncbi-nlm-nih-gov.ezp.lib.cam.ac.uk/refseq/MANE/), Locus Reference Genomic (LRG; https://www.lrg-sequence.org/), Canonical transcript from UniProt (https://www.uniprot.org/) and manually selected from Ensembl (i.e. ensembl canonical; https://www.ensembl.org/info/genome/genebuild/transcript_quality_tags.html).

Extraction of pathogenic variants from UKB

I searched for the pathogenic and likely pathogenic variants, collected from the various resources presented earlier, in the UKB WES data (data-field 23141, downloaded on 10th November 2020). These variants were extracted from UKB aggregated Variant Call Format Specification (VCF v4.4; https://github.com/samtools/hts-specs). The software used to manipulate the aggregated VCF files was `bcftools` version 1.10.2-17 (using htslib 1.10.2-22; Danecek et al. 2021). The pathogenic and likely pathogenic variants that were identified in the UKB participants were used for the following experiments only if they met the variant-call quality requirements. The minimum quality criteria for variants to be included were: (i) AF < 0.001; (ii) QUAL > 30; (iii) STATUS = PASS; (iv) DP > 20 (per each sample). Moreover, all the variants that had the same genotype as reference or missing genotypes were excluded. The command used to extract the variants from the UKB WES aggregated VCF was:

```
bcftools view --threads 20 -i 'INFO/AF < 0.001 && QUAL > 30' \
        # include variants if above the noted AF and the QUAL
    pathogenic_variant_from_200KWES.bcf | \
        # bcf files of the pathogenic variants present in UKB
    bcftools query -i'FORMAT/DP[0-]>20 & FORMAT/GT[0-]!="ref" & FORMAT/GT[0-]!="mis"'\
        # include if read depth is higher than 20
        # exclude reference and missing genotype
    -f '%CHROM\t%POS\t%REF\t%ALT\t%ID\t%INFO[\t%SAMPLE=%GT;%DP]\n' > \
    output_file.tab
```

Code 2.1 | Script and parameter used to select the variants to use in the VarioPath project.

VCF variant structure was normalised using the 'norm' function of the bcftools software. The VarioPath analyses were performed only on unrelated Europeans to limit the genetic confounding effect of the other ethnic groups. Relatedness and ethnicity were calculated using "Somalier" software (https://github.com/brentp/somalier) according to the developer guidelines. Somalier calculates the ancestries based on a predefined set of variants that are known to be polymorphic in the different ethnic groups according to the 1,000 Genomes Project (Fairley et al. 2020). It uses these variants to calculate principal components and subsequently label the different groups according to the known ethnicities

of the 1,000 Genomes Project. Somalier uses the coefficient of relationship to calculate the identity-by-state (IBS) factor 0, IBS1, IBS2 (Stevens et al. 2011), and kinship coefficient (Lange 2012) of the cohort. The parameters are then used to identify related samples. The ethnic-specific AF was calculated using the plink2 software (2.00a3LM; https://www.cog-genomics.org/plink/2.0/) with the command:

```
plink2 --bcf unrelated_genotype_file.bcf \ # bcf file with the variants and the genotype of the samples
    --freq --pheno population_file.txt \ # file with two columns: smapleID and predicted ethnicities
    --loop-cats population \ # header of the population file which has the predicted population
    --vcf-half-call 'haploid'
```

Code 2.2 | Script to calculate the ethnic group-specific AF with plink2.

Annotation of the pathogenic variant list

The pathogenic and likely pathogenic variants were annotated using the Ensembl Variant Effect Predictor (VEP; software version 100.2, cache version 99; McLaren et al. 2016). This annotation step allowed me to (i) estimate the impact of the variants; (ii) assign variants to genes and transcripts; (iii) convert genomic variants to their respective amino acid substitutions; (iv) annotate the variants with gnomAD AFs and (v) annotate the variants with the CADD scores. The command used to annotate pathogenic variants with VEP was:

```
vep -dir ${ensembl_dir} --cache --species homo_sapiens --offline \
        -i pathogenic_variant_from_200KWES.txt --no_check_variants_order --fork 6 \
        --force_overwrite --tab \
        --hgvs --symbol --buffer_size 50 --merged --transcript_version \
    --protein --symbol --uniprot \
        --canonical --mane --af_gnomad \
    --fasta ../reference_UKB_GRCh38/genome.fa \
        --plugin CADD,${ensembl_dir}/whole_genome_SNVs.tsv.gz --cache_version 99 \
        --no_check_variants_order --output_file ./annotation_withVEP.tab
```

Code 2.3 | Script used to run VEP and annotate the VarioPath variants.

VEP flags the impact of a variant on the transcripts or the proteins as “HIGH”, “MODERATE”, “MODIFIER”, or “LOW”. The description of these flags and their biological meaning is reported in Table 2.1.

SO term	SO description	IMPACT
transcript_ablation	A feature ablation whereby the deleted region includes a transcript feature	HIGH
splice_acceptor_variant	A splice variant that changes the two base region at the 3' end of an	HIGH

	intron	
splice_donor_variant	A splice variant that changes the two base region at the 5' end of an intron	HIGH
stop_gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript	HIGH
frameshift_variant	A sequence variant which disrupts the translational reading frame because the number of nucleotides inserted or deleted is not a multiple of three	HIGH
stop_lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript	HIGH
start_lost	A codon variant that changes at least one base of the canonical start codon	HIGH
transcript_amplification	A feature amplification of a region containing a transcript	HIGH
inframe_insertion	An inframe non-synonymous variant that inserts bases into in the coding sequence	MODERATE
inframe_deletion	An inframe non-synonymous variant that deletes bases from the coding sequence	MODERATE
missense_variant	A sequence variant that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved	MODERATE
protein_altering_variant	A sequence_variant which is predicted to change the protein encoded in the coding sequence	MODERATE
splice_region_variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron	LOW
incomplete_terminal_codon_variant	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed	LOW
start_retained_variant	A sequence variant where at least one base in the start codon is changed, but the start remains	LOW
stop_retained_variant	A sequence variant where at least one base in the terminator codon is changed, but the terminator remains	LOW
synonymous_variant	A sequence variant where there is no resulting change to the encoded amino acid	LOW
coding_sequence_variant	A sequence variant that changes the coding sequence	MODIFIER
mature_miRNA_variant	A transcript variant located with the sequence of the mature miRNA	MODIFIER
5_prime_UTR_variant	A UTR variant of the 5' UTR	MODIFIER

3_prime_UTR_variant	A UTR variant of the 3' UTR	MODIFIER
non_coding_transcript_exon_variant	A sequence variant that changes non-coding exon sequence in a non-coding transcript	MODIFIER
intron_variant	A transcript variant occurring within an intron	MODIFIER
NMD_transcript_variant	A variant in a transcript that is the target of NMD	MODIFIER
non_coding_transcript_variant	A transcript variant of a non-coding RNA gene	MODIFIER
upstream_gene_variant	A sequence variant located 5' of a gene	MODIFIER
downstream_gene_variant	A sequence variant located 3' of a gene	MODIFIER
TFBS_ablation	A feature ablation whereby the deleted region includes a transcription factor binding site	MODIFIER
TFBS_amplification	A feature amplification of a region containing a transcription factor binding site	MODIFIER
TF_binding_site_variant	A sequence variant located within a transcription factor binding site	MODIFIER
regulatory_region_ablation	A feature ablation whereby the deleted region includes a regulatory region	MODERATE
regulatory_region_amplification	A feature amplification of a region containing a regulatory region	MODIFIER
feature_elongation	A sequence variant that causes the extension of a genomic feature, concerning the reference sequence	MODIFIER
regulatory_region_variant	A sequence variant located within a regulatory region	MODIFIER
feature_truncation	A sequence variant that causes the reduction of a genomic feature, concerning the reference sequence	MODIFIER
intergenic_variant	A sequence variant located in the intergenic region, between genes	MODIFIER

Table 2.1 | Description of variant biological impacts that are used by the VEP software and adopted in the VarioPath project. Adapted from (Ensembl Variant Effect Predictor (VEP) website, accessed in September 2021)

Phenotypes in UKB, definition and extraction

Phenotypes and covariates of interest were extracted from UKB using both *ad hoc* and UKB-provided scripts. The BTPD phenotype definition and symptomatology, based on the ICD-10 codes, has been manually curated via experts opinions, literature review and the CALIBER phenotype resource (https://www.caliberresearch.org/portal/codelists). The disease phenotype definition used in this thesis is listed in Table 2.2. Every UKB participant,

who had at least one of the ICD-10 codes relative to the list of diseases, was defined as “case” in the case-control comparisons; those participants without any ICD-10 code in Table 2.2 as controls. Moreover, when a disease/phenotype is described only by the three-digit code, all the sub-codes present for that code were also used. For instance, aplastic anaemia is defined by the D61 ICD-10 code (Table 2.2). Therefore also the D61.9 and the other subcategories were adopted in the disease definition. The UKB phenotypes were extracted from the category “2000” (downloaded in November 2020). Specifically, the ICD-10 codes describing participant phenotypes were extracted from the “HESIN_DIAG" file (https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=2000). The covariates used in the case-control studies were age, smoking status, bmi, sex and the first ten genetic principal components. The covariates were extracted using the “ukbconv” software provided by UKB. The quantitative data used in this study were relative to the full blood count data (Category 100081; https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=100081). This data were standardised for their skewness and corrected for the technical variables (e.g. date of recruitment and centre of recruitment) as previously described (Astle et al. 2016; Vuckovic et al. 2020).

Disease	**ICD-10 codes**
Abdominal pain	R10.4
Acute lymphoblastic leukaemia	C91.0
Acute myeloid leukaemia	C92.0; C92.4; C92.5; C92.6; C92.8; C92.9; C93.0; C94.0; C94.2; C95.0; D46.3
Aplastic anaemia	D61
Arthralgia	M25.5
Arthritis	M05; M06; M07; M08; M09; M10; M11; M12; M13; M14; M19.90
Ataxia	R26.0; R27.0; R27.8
Atrial fibrillation	I48
Bleeding - epistaxis	R04.0
Bleeding - General	I61.0; I61.1; I61.2; I61.3; I61.4; I61.5; I61.6; I61.8; I61.9; I62.0; I62.1; I62.9; R23.3; D69.2; R04.1; R31; T79.2; O72.0; O72.1; O72.2; T81; M25.00; M25.01; M25.02; M25.03; M25.04; M25.05; M25.06; M25.07; M25.08; M25.09; I61.7; N92.3; R58; D69.9
Bleeding after tooth extraction other procedure	T81.0

Cardiac bleeding	I23.0; I31.2; S26.0
Cataract	H25.0; H25.1; H25.2; H25.8; H25.9; H26.0; H26.1; H26.2; H26.3; H26.4; H26.8; H26.9; H28.0; H28.1; H28.2; Q12.0; H59.0
Central nervous system bleeding	I61.0; I61.1; I61.2; I61.3; I61.4; I61.5; I61.6; I61.8; I61.9; S06.4; S06.5; S06.6; I60.0; I60.1; I60.2; I60.3; I60.4; I60.5; I60.6; I60.7; I60.8; I60.9; I62.0; I62.1; I62.9
Cholelithiasis	K80.0; K80.1; K80.2; K80.3; K80.4; K80.5; K80.8
Chronic kidney disease	N18.1; N18.2; N00; N10; N17; N01; N03; N05.2; N05.3; N05.4; N05.5; N05.6; N07.2; N07.3; N07.4; N18.3; N18.4; N18.5; N18.9; N19; N25; Y84.1; Z49; Z99.2; T86.1; Z94.0; N28.9
Coronary artery atherosclerosis	I25; I70; K76.1
Cutaneous bruising or bleeding	R23.3; D69.2
Diabetes	E11; E12; E13; E14; G590; G632; H280; H360; M142; N083
Pulmonary embolism	I26
Dyslipidemia	E78
Eczema	L20; L30.9
Eye bleeding	H35.6; H43.1
Gastrointestinal bleeding	K92.0; K92.1; K92.2; K62.5; K29.0; I85.0; I98.3; K22.6; K25.0; K25.2; K25.4; K25.6; K26.0; K26.2; K26.4; K26.6; K27.0; K27.2; K27.4; K27.6; K28.0; K28.2; K28.4; K28.6; K29.0; K62.5
Gout	M10.0; M10.4; M10.9
Growth delay	R62.8; R62.9
Gynaecological bleeding	N83.7; N93.8; N93.9; O03.1; O03.6; O04.1; O04.6; O05.1; O05.6; O06.1; O06.6; O07.1; O07.6; O08.1
Haemangioma	D18.0
Haemarthrosis	M25.0
Haematuria	R31
Haemolytic anaemia	D58.9
Hearing loss	H90.3; H90.4; H90.5; H90.6; H90.7; H90.8; H91.1; H91.2; H91.3; H91.8; H91.9; Z45.3; Z46.1; Z97.4
Hepatomegaly	R16.0; R16.2
Hereditary deficiency of oth	D68.2

clotting factors	
Hereditary elliptocytosis	D58.1
Hereditary factor IX deficiency	D67
Hereditary factor VIII deficiency	D66
Hereditary factor XI deficiency	D68.1
Hereditary haemolytic anaemia	D58.8; D58.9
Hereditary spherocytosis	D58.0
Hyperthyroidism	E05.0; E05.1; E05.2; E05.5; E05.8; E05.9; E06.9; H06.2
Hyposplenism	D73.0
Hypothyroidism	E05.5; E06.9; E03.5; E03.8; E03.9; E06.2; E06.3; E06.5
Idiopathic thrombocytopenic purpu	D69.3
Iron disorder	E83.1
Ischaemic stroke	I630; I631; I632; I633; I634; I635; I638; I639; I693
Jaundice	R17
Liver disease	K70; K71; K72; K73; K74; K75; K76; K77
Lower respiratory tract infection	J20; J21; J22
Lymphoma	C81; C82; C83; C84; C85; C88.3; C88.7; C88.9; C91.4; C91.5; C96
Maculopapular exanthema	R21
Meningitis	G00; G01; G03
Menorrhagia	N92.0; N92.1; N92.2; N92.4
Metastases	C77; C78; C79
Muscle weakness	M62.8; M62.9
Myalgia	M79.1
Myelodysplastic syndrome	D46.0; D46.1; D46.2; D46.4; D46.5; D46.6; D46.7; D46.9; C94.6
Myelofibrosis	D47.4
Myocardial infarction	I25; I21; I22; I23; I24
Non-melanoma skin cancer	C44
Obstetric bleeding	O20.8; O20.9; O46.8; O46.9; O67.8; O67.9; O71.7; O90.2
Oral cavity bleeding	R04.1

Other myeloproliferative diseases	C86; C88.0; C88.2; C88.4; C90.2; C90.3; C94.4; D45; D47
Other primary thrombocytopenia	D69.4
Otitis media	H66; H65
Pallor	R23.1
Peripheral arterial disease	I731; I738; I739; I743; I744; I745
Pneumonia	J09; J10; J11; J12; J13; J14; J15; J16; J17; J18
Postpartum haemorrhage	O72.0; O72.1; O72.2
Proteinuria	R80
Qualitative platelet defects	D69.1
Renal tubular acidosis	N25.8
Respiratory system bleeding	P26.1; R04.2; R04.8; R04.9
Restrictive cardiomyopathy	I42.5
Secondary pulmonary hypertensio	I27.2
secondary thrombocytopenia	D69.5
Sinusitis	J01; J32
Skin ulcer	L98.4
Splenomegaly	R16.1
Stomatocytosis	D58.8
Stroke NOS	G46.3; G46.4; G46.5; G46.6; G46.7; G46.8; I64; I69.4
Thrombocytopenia unspecified	D69.6
Transient ischaemic attack	G45.0; G45.1; G45.2; G45.3; G45.4; G45.8; G45.9; G46.0; G46.1; G46.2; I65; I66
Traumatic bleeding	T79.2
Upper respiratory tract infection	J39; J31
von Willebrand disease	D68.0

Table 2.2 | List of the phenotypes considered in the VarioPath project and their definition based on the ICD-10 codes.

Calculation of the OR via the burden aggregation test

The pathogenic and likely pathogenic variants present in UKB participants were aggregated by genes and used to calculate a cumulative effect size (i.e. Burden test; Guo et

al. 2018). Therefore, this score associates genetic regions with diseases. To perform these calculations, I used the raremetal software (https://genome.sph.umich.edu/wiki/RAREMETAL; Sanna et al. 2008; Willer et al. 2008). Genes were the discrete units used to perform the aggregation in the VarioPath project. For instance, all the variants that occurred in *MPL* were aggregated and considered together in their effect on the phenotype. This approach loses statistical power if the variants in a gene have opposite effects. For example, some variants are protective against the phenotype and others increase the risk of presenting the same phenotype. However, this approach allows exploring the effect of the variants that have so few carriers that would be removed from the experiments. Additive and dominant modes of inheritance were tested for all the genes and phenotypes.

The commands used to run the burden test with raremetal were:

```
raremetalworker --ped ${ped_dir}/${line} \ # this file contains the phenotypes
        --dat ${out_dir}/variopath_burden.dat \ # this file contains the covariates
        --vcf ${out_dir}/variopath_variants.vcf.gz \ # this file has the genotypes
        --traitName TRAIT \ # header that contains the phenotype
        --inverseNormal \ # normalisation method
        --makeResiduals --prefix \
        ${out_dir}/phenotypes/${name}/raremetalworker.${name} \ # output file
        --dominant[--recessive] \ # mode of inheritance to test
        --useCovariates \
        --xStart 2699520 --xEnd 155701383 # coordinated of the x chromosome variants
```

```
# Run raremetal burden test
raremetal --summaryFiles ${out_dir}/phenotypes/score_file_${name} \
    --covFiles ${out_dir}/phenotypes/cov_file_${name} \
    --groupFile ${group} \ # how to group the variants. By gene in this case
    --MAB --MB --BBeta --SKAT  --burden --useExact \ # statistical test to perform
    --longOutput --tabulateHits --altMAF --hitsCutoff 0.1 --prefix \ # output structure
    ${out_dir}/phenotypes/${name}/${name2} # output name
```

Code 2.3 | Commands used to perform the burden test analyses for the VarioPath project.

Estimation of the effect of single variants

The contribution of the single variants to the phenotype was estimated with the mathematical models reported below (Lynch, Walsh, and Others 1998). The role of a variant was considered only if the allele count for that variant was greater than five in the case of continuous phenotype (e.g. red cell count) or greater than 15 in the case of categorical phenotypes (e.g. thrombocytopenia). In the case of quantitative phenotypes, the mathematical model adopted was:

$$Y = \varepsilon + \beta_1(variant) + \sum_{i=1}^{n} \beta_{i,n}(covariates)$$

Where Y is the quantitative phenotype that has been normalised to a standard distribution. The linear model estimates the role of the variants and covariates on the phenotype. The β variant is the effect, in standard deviations, of the SNPs to the phenotype. In the case of qualitative phenotypes:

$$logit(\varepsilon[Y]) = \beta_0 + \beta_1(variant) + \sum_{i=1}^{n} \beta_{i,n}(covariates)$$

Where Y is the qualitative phenotype in consideration that has a Bernoulli distribution in response to the predictors. The logistic function calculates the odds associated with the predictors of being in the class disease. β is the effect of the SNPs on the phenotypes, and the summation refers to the impact of the covariates (i.e. smoking status, bmi, age, sex and genetic principal components).

The interplay between rare and common variants in VTE

The contribution that rare and common variants have to the different VTE characteristics has been estimated with the following mathematical model:

$$logit(\varepsilon[Y]) = \beta_0 + \beta_1(rare) + \beta_2(prs) + \sum_{i=1}^{n} \beta_{i,n}(covariates)$$

Where *rare* is a dichotomous variable that reports whether a person carries at least one rare variant, *prs* is the PRS score of each individual and the covariates are the same listed earlier. Different logit models have been tested based on additional covariates (Fig. 2.1). The one with the better performance was adopted to describe the outcome of these phenotypes. The PRS was used as a proxy of the effect of common variants, and it was calculated in a previously published study (Klarin et al. 2019).

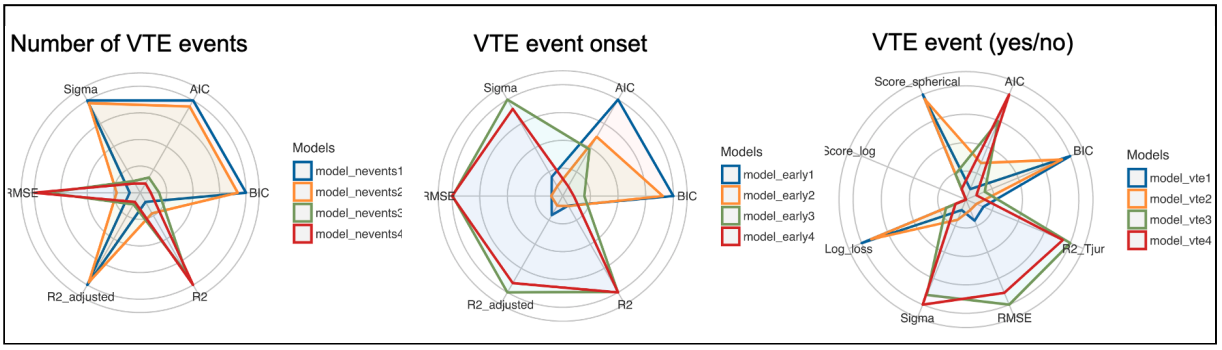


Fig 2.1 | Prediction capacities of the VTE phenotype models. *Four models that differ for the covariates used have been tested. The goal was to predict the number of VTE events, VTE manifestation, and onset age. The greater the area covered by the model, the higher its prediction capacity is.*

PRS variants with large effect sizes localise in the proximity of relevant biological pathways

The human interactome (i.e. PPI network) was built on the combination of the information coming from STRING (v11.0, score>0.75) and the Open Targets project (www.opentargets.org; accessed in November 2019). The Open Targets interactome is a compilation of IntAct, Signor and Reactome. Self-loops were removed leading to 18,410 nodes and 571,917 edges (Choobdar et al. 2019; Schwartzentruber et al. 2021). A PRS SNP was considered to overlap with a gene if the SNP laid within the gene or in the 10Kb surrounding the gene body. When several SNPs mapped to the same gene, only the one with the largest effect size was kept. SNPs were grouped into four quantile clusters depending on their effect sizes. Similarly, the interactome was divided into four groups based on their distance from the 93 BTPD genes. This method, applied to calculate the enrichment, was recently published (Barrio-Hernandez et al. 2021). The formula to calculate the odds ratio of a variants is:

$$\frac{g \in B \cap PRS_{\uparrow} \cdot g \notin B \cap PRS_{\downarrow}}{g \in B \cap PRS_{\downarrow} \cdot g \notin B \cap PRS_{\uparrow}}$$

Where g represents a gene, B is the list of 93 BTPD genes, PRS is the effect size of the variant associated with that gene, and the arrows indicate if that variant is part of the highest percentile (>50%) or lower percentile (<50%). These analyses and the PPI network were the result of a collaboration with Dr. Pedro Beltrao and Dr. Inigo Barrio-Hernandez.

Comparison between effect size and protein function

This analysis used only the missense variants that map onto protein crystal structures, based on Protein Data Bank (PDB; https://www.rcsb.org/). The deleterious effect of a variant was estimated with a support-vector machine (SVM) supervised model (Petrova and Wu 2006; Capra et al. 2009). The training set was composed of disease variants (i.e. pathogenic variants extracted from ClinVar and humsavar) and benign variants (i.e. variants recorded in gnomAD). The number of variants used in the training set was 84,105, and they were all selected to not overlap with the collection of variants used in VarioPath. Every variant was converted to its amino acid substitution and annotated with several descriptors used in the training of the model. The predictors were: (i) aa is used in disulphide bonds; (ii) aa binds proteins; (iii) aa binds DNA; (iv) aa binds ligand; (v) aa binds metal; (vi) conservation score of the residue; (vii) aa chemistry change; (viii) CADD score. The model generated was used to classify VarioPath variants as deleterious or not based only on their amino acid change characteristics. The effect sizes selected to compare with the SVM scores were the highest statistically significant scores identified across all the phenotypes tested. The statistical method adopted to compare the two scores was Pearson correlation. The training of the SVM model and calculation of the variant deleteriousness was performed by Professor Dame Janet Thornton and Dr Roman Laskowski.

2.2 Cell Biology methods for the identification of non-coding regions relevant to BTPDs

Culturing Human induced Pluripotent Stem cells

Human induced pluripotent stem cells (hiPSc) were obtained from the “Human Induced Pluripotent Stem Cell Initiative” (HipSci, http://www.hipsci.org/), namely Qolg (HPSI1113i-qolg_1) and Ffdk (HPSI0813pf-ffdk). In addition, megakaryocytes and hepatocytes were produced from a third hiPSC cell line, A1ATD-iPSCs (Kosuke Yusa et al, Nature 2011), kindly provided by Professor Ludovic Vallier. hiPScs were kept in culture using StemFlex medium (ThermoFisher Scientific), which was changed every day. Cells were incubated at 5% CO_2, 37°C. hiPSC pluripotency was routinely checked at the flow cytometer via the expression of TRA-1-60 and SSEA4 surface markers.

For the dCasKRAB experiments, stable hiPSCs that expressed the sgRNAs and the dCasKRAB components were obtained via Lentivirus infection (see below). For these cells, blasticidin (1.5µg/ml), puromycin (1µg/ml), and hygromycin (50µg/ml) supplemented the StemFlex medium to maintain an active expression of the region containing the dCasKRAB components. The dCasKRAB protein was under the control of the Tet-On system. Therefore, at the beginning of the differentiation medium, the hiPSC were supplemented with doxycycline (1µg/ml) to express the dCasKRAB protein.

Culturing the human embryonic kidney 293 T

Human Embryonic Kidney 293 T (HEK-293T) is a common immortalised cell type derived from the human embryonic kidney. I used it to screen the quality of all the plasmids used in this thesis. HEK-293T was cultured in high-glucose DMEM supplemented with 10% fetal bovine serum (FBS, SigmaAldrich), 5.5 ml of non-essential amino acids (SigmaAldrich), 5.5 ml penicillin/streptomycin (SigmaAldrich) and tylosin (ChemCruz). Incubator settings were: 5% CO_2, 37°C.

Culturing the imMKCL

imMKCL is an immortalised cell line that resembles MK biology (S. Nakamura et al. 2014), and it was used for the reporter assay experiments. These cells were cultured in IMDM (SigmaAldrich), supplemented with 15% FBS, 5ml of insulin-transferrin-selenium 100x, 0.5 ml ascorbic acids, 0.5 ml α-monothioglycerol, 1µl of doxycycline [20µg/µl], 100µl stem cell factor (SCF) [10µg/µl], 100µl thrombopoietin (TPO) [10µg/µl], 5ml of L-glutamine [200mM]. Incubator settings were: 5% CO_2, 37°C. The medium was rejuvenated every two days, and cells were split to keep a constant exponential growth phase.

Megakaryocyte differentiation from hiPSc

hiPS cells were reprogrammed into megakaryocytes using an adaptation of the forward programming (FOP) differentiation established in Dr Cedric Ghevaert laboratory (Moreau et al. 2016). Briefly, hiPSC were seeded at a density of 50,000 cells/well in a 6-well plate and cultured with StemFlex medium (FOP Day-1). The following day (Day0 of differentiation), lentiviruses (LV) were used to infect the cells. LVs transduced the three key transcription factors used in the MK differentiation, namely GATA1, TAL1 and FLI1 (FOP Day0). LV concentration was 40 multiplicity of infection per virus, and protamine sulphate was used to increase the transfection efficiency. On FOP Day0 and Day1, differentiating cells were cultured in AE6 medium supplemented with FGF2 [20 ng/ml] and BMP4 [10

ng/ml]. From FOP Day5 to FOP Day20, a chemically defined medium induced MK differentiation and maturation. Fresh medium was added every other day. The iMK chemically defined medium composition was: CellGro, TPO [160 ng/ml] and SCF [50ng/ml]. At FOP Day10, cells were detached from the wells using TrypLE and transferred to plates suitable for suspension cells. At the end of the differentiation protocol (i.e. FOP Day20), the quality of the differentiated iMK was tested by flow cytometry via the expression of MK specific surface markers, namely CD41a and CD42b (Fig 2.2).

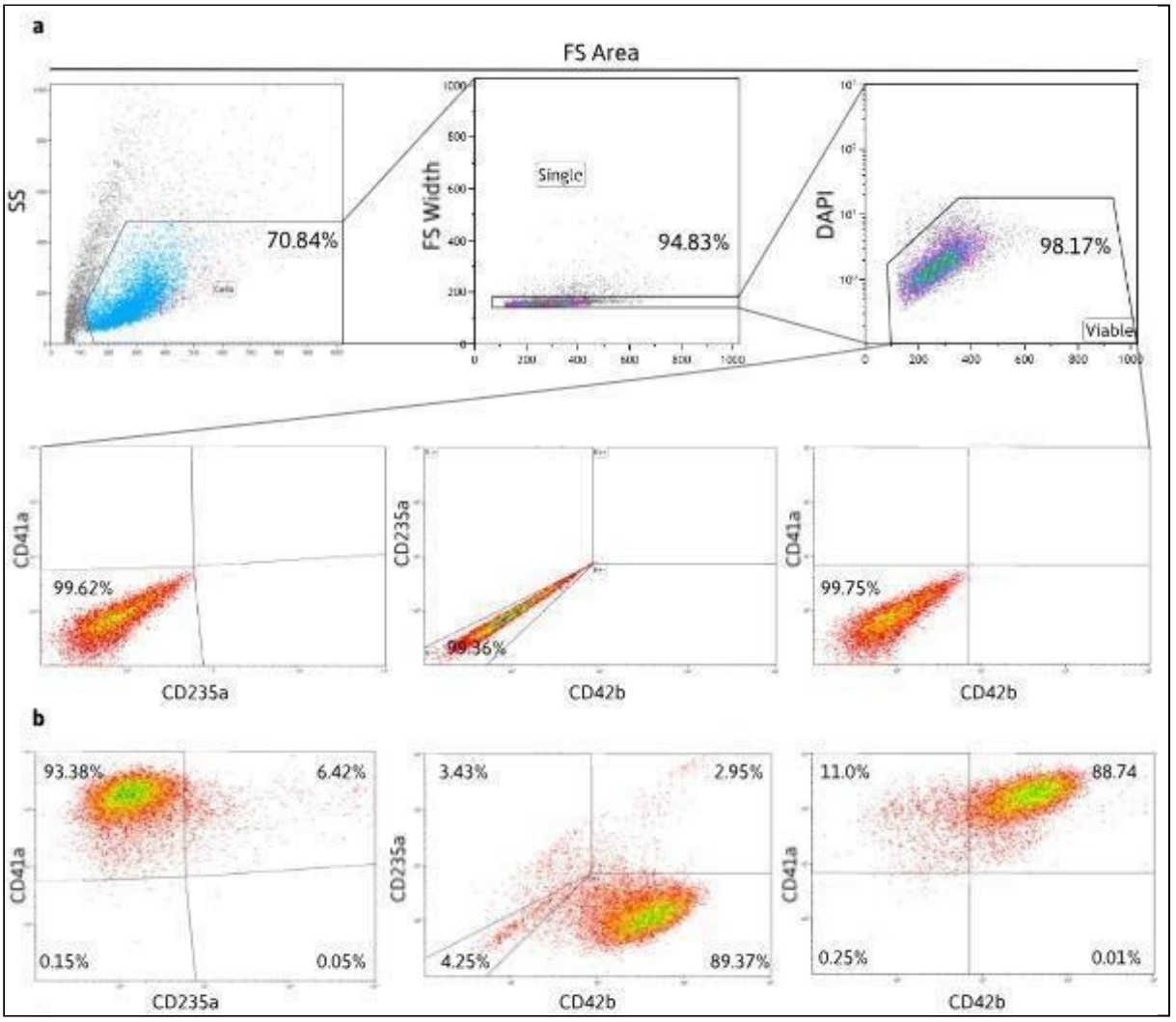


Fig 2.2 | Gating strategy adopted to test the differentiation of iMK. *Flow cytometry was performed at the end of the differentiation to test the amount of CD42 positive iMK, a proxy for the iMK maturation. a) gating strategy and flow cytometry for the negative control. b) example of a differentiated cell line.*

Endothelial cells differentiation from hiPSc

T225 flasks were coated with Matrigel. The coating step requires 60-minute incubation at room temperature (RT) with 0.33 mg/ml Matrigel solution. All the reagents used

for this protocol were filtered through a 0.22 µm PES filter. The Matrigel solution volume required for a T225 flask is 24ml. hiPSC were trypsinised to obtain a single-cell suspension used to seed cells at a density of $9.2x10^5$ per T225 flask. On the seeding day (Day-1), cells were cultured in the StemFlex medium. The differentiation protocol (Day 0) started with the replacement of the StemFlex medium with the custom made mesoderm induction medium (for a T225 flask: 18 ml F12, 18 ml IMDM, 363.3 µl lipid, 18 µl transferrin, 363.6 µl penicillin-streptomycin, 145.44 µl FGF2 [50 µg/ml], Chiron [10mM] 28.8 µl, ly294002 [20mM] 18µl). On Day2, the mesoderm induction medium was replaced with the iEC induction medium (for a T225: StemPro 60.646 ml, L-glutamine [200mM] 0.59 ml, VEGF-A [11.1 µg/ml] 1.134 ml, forskolin [10mM] 12.4 ul, L-ascorbic acid [250mM] 0.252 ml). On Day3, fresh iEC induction medium replaced the exhausted one. On Day5, the differentiating cells were resuspended and seeded in fibronectin-coated plates, $4.6x10^6$ cells in a T225 flask. From Day5 to Day9, there was a daily replacement of the iEC medium (for a T225 flask: 73.2 ml StemPro, L-glutamine 0.714 ml, 1.36 ml VEGF, 14.4 ml Forskolin, 0.29 ml Ascorbic acid). At the end of the differentiation protocol (Day9), cells were enriched for mature iEC using MACS beads (Miltenyi Biotec) for the VE-Cadherin (CD144) surface marker. After enrichment, mature iECs were tested for the expression of common EC surface markers, VECAD, CD34, PECAM-1, VEGFR2 and CD45 (Fig. 2.3).

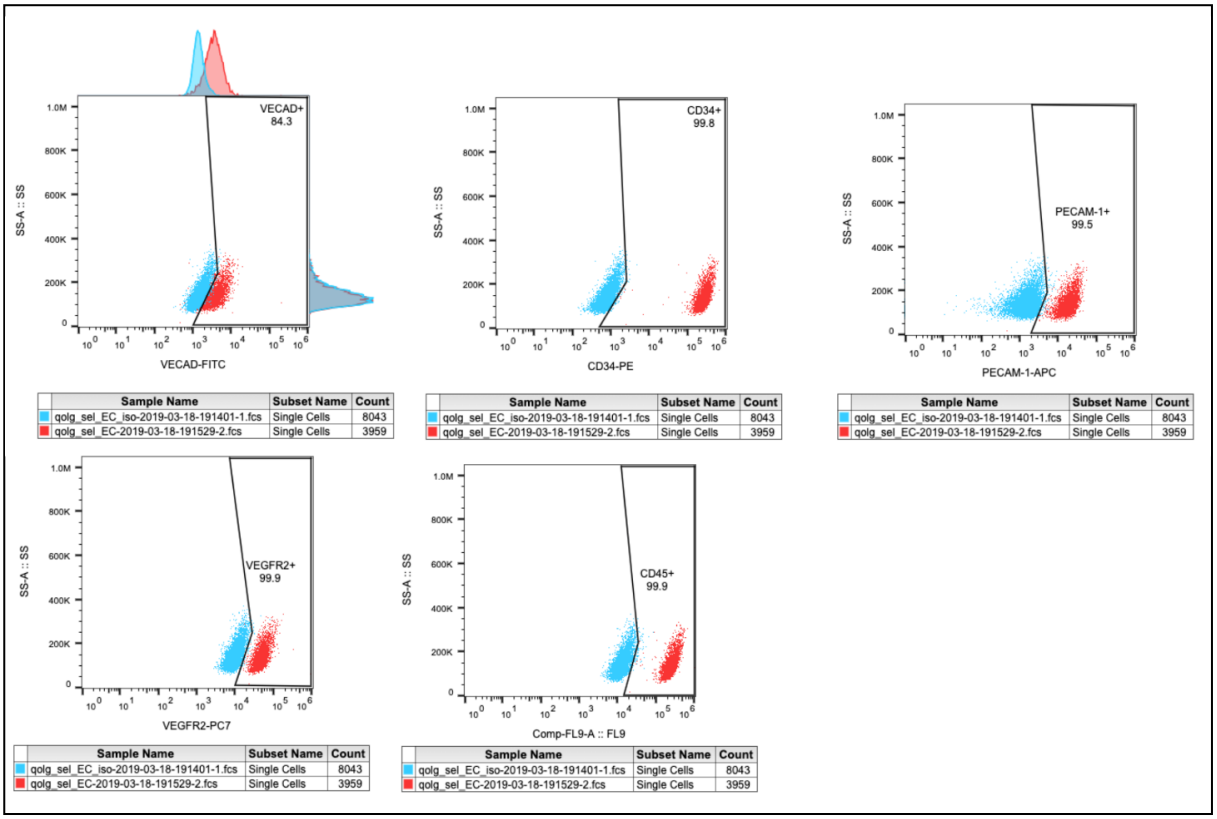


Fig. 2.3* | *Surface marker expression on the differentiated iECs. *At the end of the differentiation, enriched iEC showed a homogeneous population of mature iEC. These cells expressed all the markers that have been used to test the quality of the iEC.*

Hepatocyte differentiation from hiPSc

The HEP differentiation protocol was described elsewhere in detail (Touboul et al. 2010). Briefly, hiPS cells were grown in fibronectin-coated plates for the length of the whole differentiation protocol (i.e. 30 days). hiPSC were cultured in Essential8 (ThermoFisher) medium on the days preceding the differentiation protocol. On Day-1 of the differentiation, cells were trypsinised and seeded at $20x10^4$ cells/cm^2. The differentiation protocol requires hypoxia incubators and daily change of the medium from Day0 to Day11, then every other day up to Day30. Day1-3 were used to induce the endoderm differentiation. The medium for these three days was RPMI supplemented with activin [10µl/ml], FGF2 [20µl/ml], BMP4 [1µl/ml], ly294002 [0.2µl/ml] and chiron [1µl/ml]. From Day4 to Day8, the RPMI medium required only activin 5µl/ml supplement. In the last days of the differentiation, cells were cultured with Hepatozyme medium supplemented with human oncostatin [1ul/ml] and hepatocyte growth factor [1ul/ml]. At the end of the differentiation, hepatocytes identity was assessed via quantifying relevant transcripts using qPCR, namely *NANOG*, *AFP*, *A1AT*, *ALB*, *HNF4a* (Fig. 2.4). Also, the activity of CYP3A4 was controlled to verify the success of the differentiation.

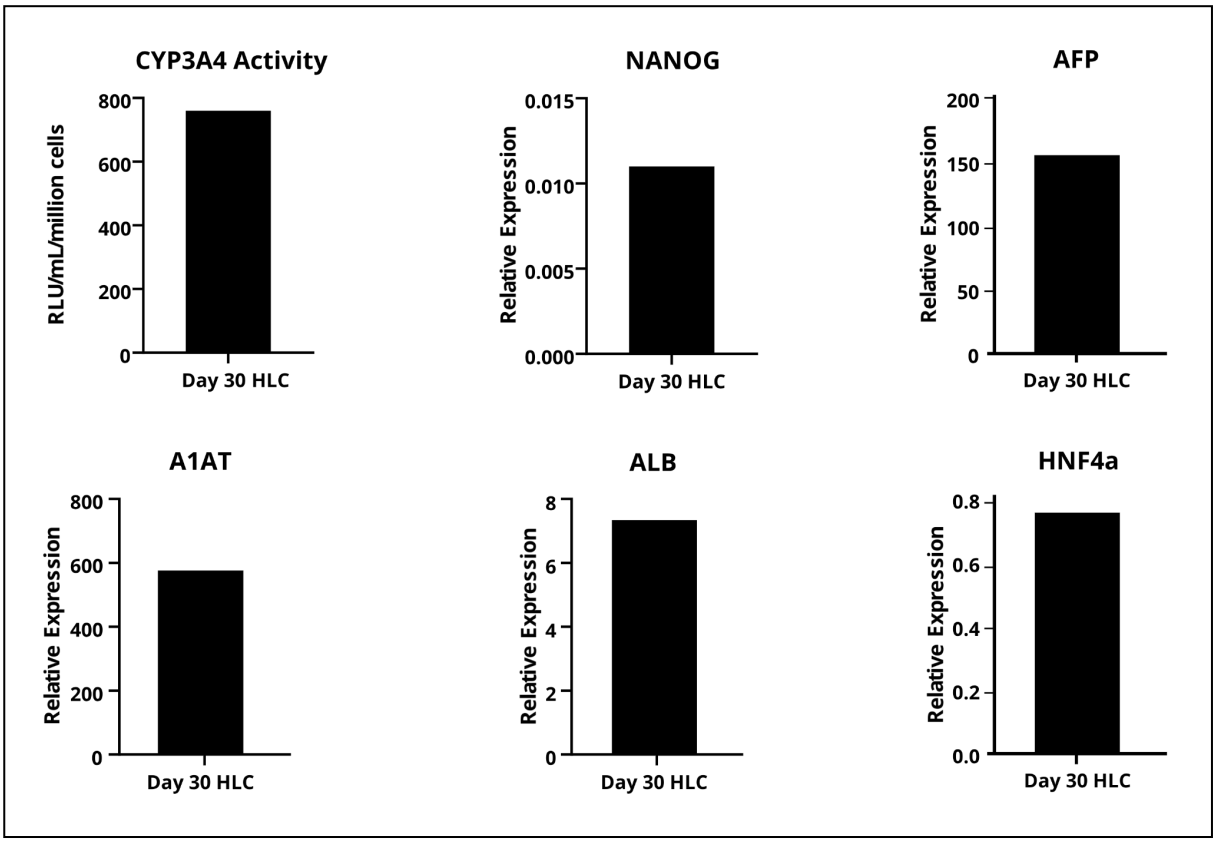


Fig 2.4 | QC plots of the HEP differentiation. *At the end of the differentiation, HEPs were a homogeneous population that expressed the correct molecular markers of mature HEP.*

2.3 TG Hi-C relevant protocols and analysis

DNA purification

One volume of phenol:chloroform:isoamyl alcohol (25:24:1) was added to the cell lysate containing the DNA. The phenol:chloroform mix allows the removal of the organic component from the solution. The organic component phase separation was obtained via a centrifugation step at 2,430xg for 10 minutes. Next, the upper/aqueous phase was transferred in a new tube and supplemented with 1 volume of chloroform. A second centrifugation step, again 2,430xg 10 minutes, created another phase separation. Finally, the upper/aqueous layer containing the DNA was transferred in a clean tube. This mix was diluted in tris-HCl and complemented with 100 µg/ml of glycogen and 0.1 volumes of sodium acetate 3M pH 5.2. DNA precipitation was triggered with 20.6 volumes of ice-cold 100% ethanol and incubated for 16 hours at -20°C. DNA was precipitated with a centrifuge of 2,430xg for 1 hour at 4°C and washed with 70% ethanol. The DNA pellet was resuspended in 150 µl of 10 mM Tris-HCl, pH 8.

Libraries production

The libraries have been created following and adapting three protocols: Dekker laboratory Hi-C protocol (Belaghzal, Dekker, and Gibcus 2017), KAPA Library Preparation Kit protocol (KK8230) and Thrombogenomic protocol (Ilenia Simeoni et al. 2016). The cells used for the libraries production were selected to be at least 80% positive for lineage markers, according to the characteristics described in chapter 2.2. 5 Mln cells per condition were fixed in 1% formaldehyde for 10 minutes. The formaldehyde effect was quenched with 125 mM of glycine, incubating the solution for 5 minutes at room temperature and then for further 15 minutes on ice. Next, cells were resuspended in Hi-C lysis buffer (10 mM Tris-HCl pH=8.0, 1 M pH=8.0, 10 mM NaCl, 0.2% NP40) for 15 minutes on ice and cell membranes were mechanically broken with insulin syringes and needles. To pellet the nuclei, and remove the debris, I used a series of 3 centrifugation (3000xg for 5 minutes) and washed the pellet (NEBuffer 3.1 1X) after each centrifugation. After the last centrifugation, nuclei were resuspended in 360 µl of NEB 3.1 Buffer 1X, supplemented with SDS 0.1% (final concentration), and incubated at 65°C for 10 minutes. Afterwards, SDS activity was quenched with triton-X at a final concentration of 1% v/v. iECs and iMKs were digested using DpnII as a restriction enzyme, while HindIII was used for iHEP. The digestion was performed overnight (i.e. 16 hours) at 37°C and restriction enzymes were inactivated at 65°C for 25 minutes. The digested fragments ends were filled with a four-hour reaction at 23°C with the following mix (per sample):

Milli-Q	2 µl
10x NEB 3.1 (NEB)	6 µl
dCTP (Invitrogen)	1.5µl
dGTP (Invitrogen)	1.5 µl
dTTP (Invitrogen)	1.5 µl
Biotin dATP (Invitrogen)	37.5 µl
Klenow polymerase (NEB)	10 µl

The use of biotinylated nucleotides is crucial, during the fill-in, for the following pulldown procedures. Biotinylated adenine was chosen because of its central position in the overhangs created by DpnII digestion, increasing the incorporation efficiency (Belaghzal, Dekker, and Gibcus 2017). The resulting blunt ends were ligated with a 20-hour incubation at +4°C with (per sample):

Milli-Q	282 µll
NEB T4 ligase buffer + 25% PEG (5X)	240 µl
Triton X 10%	120 µl
BSA 10 mg/ml	12 µl
T4 ligase enzyme (NEB)	3 µl

After the ligation, fragments were de-cross-linked from proteins and purified. The RNA was removed with 30 µl of RNase and a 30-minute incubation at 65°C. 50 µl of Proteinase K (≥30 units/mg) and overnight incubation at 65°C removed the proteins linked to the DNA. DNA purification was performed with phenol:chloroform protocol (see “DNA purification” paragraph above). I performed a “dangling ends” removal step to reduce the unspecific pull-down due to nucleotides linked to the edges of the DNA fragments. This reaction was performed with the following mix (every 5 µg of DNA) incubated for 4 hours at 20°C:

10X NEBuffer 2.1	5 µl
10 mM dATP	0.125 µl
10 mM dGTP	0.125 µl
3000 U/ml T4 DNA POL (NEB)	5 µl
Milli-Q	up to 5 µl
DNA	up to 5 µl

Ultimately, the circular fragments resulting from ligation were linearised using a sonication step. I used the Sonicator Diagenode Minichiller 300, with ten intermittent cycles of 30 seconds. According to the manufacturer instructions, fragment ends were repaired with the KAPA library preparation kit for Illumina platforms (KAPA Biosystems) using the following reaction mix:

2X KAPA HiFi HotStart ReadyMix	25 µl
Adapter-ligated HiC library	20 µl
PCR Oligos (5uM)	5 µl

At this point, samples were ready for the biotin pool-down. I used two µl of Dynabeads MyOne Streptavidin C1 (Thermo Fisher) every µg of DNA. I washed the dynabeads with a tween buffer (5 mM Tris-HCl pH=8.0, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween, Milli-Q dH2O) 2 times. After the washing steps, samples and beads were resuspended together in a beads buffer (10 mM Tris-HCl pH=8.0, 1 mM EDTA, 2M NaCl,

Milli-Q dH2O) and incubate for 15 minutes at room temperature. Then, beads-DNA complex was purified, and fragments were resuspended in TLE pH 8.0 buffer (10 mM Tris-HCl, 0.1 mM EDTA, Milli-Q dH2O). Next, this mix was used for the A-Tailing reaction. A-Tailing was performed with a KAPA library preparation kit for Illumina platforms following manufacturer protocol. The first amplification PCR was performed with the following parameters and cleaned up with Ampure XP beads (Agencourt) following the producer protocol.

98 C	45 seconds	
98 C	15 seconds	12 cycles
60 C	30 seconds	
72 C	30 seconds	
72 C	1 minute	

After the 1st PCR amplification, the hybridisation of the probes was performed following Thrombogenomics guideline: 1.25 µg of DNA was supplemented with five µl of COT human DNA and two µl of IDT Blockers. The mix was lyophilised with a DNA vacuum concentrator (Eppendorf) for 1 hour at 65°C. Then, the pellet was resuspended in 7.5 µl of the hybridisation buffer and 3µl of the Hybridization Component. Ultimately, before adding the capture probes, DNA was denatured at 95°C for 10 minutes to facilitate probes annealing. Next, 4.5 µl of biotinylated probes SeqCAP EZ probes (TG v3.0) were added to each sample and incubated for 60 hours at 47°C. Afterwards, hybridised probes were rescued using avidin coated beads (SeqCap Pure Capture beads Kit, Roche), following the manufacturer instructions. Finally, the capture mix was cleaned using Ampure XP beads (Agencourt) and a second PCR was performed with the following parameters:

Kapa HiFi HotStart Ready Mix	25 µl
Retrieved DNA	25 µl
Primer Forward (5 uM)	2.5 µl
Primer Reverse (5 uM)	2.5 µl

98 C	45 seconds	
98 C	15 seconds	
60 C	30 seconds	
72 C	30 seconds	
72 C	1 minute	

Lastly, a final Ampure purification was performed to remove any primer-dimer complex from the library. Libraries were quantified with the KAPA library quantification kit and evaluated for their size distribution on the 2100 Bioanalyzer, with High sensitivity kit (Agilent). Libraries were pooled to have 14nM final concentration for each cell type and sequenced using HiSeq 4000 (paired-end, 75 bp) or MiSeq (paired-end, 150bp).

Processing Hi-C raw data, HiCUP Pipeline

To analyse Illumina TG Hi-C raw sequences, I used the HiCUP pipeline, developed by Dr Steven Wingett at the Babraham Institute (Wingett et al. 2015). The workflow of the pipeline is hicup_truncater > hicup_mapper > hicup_filter > hicup_deduplicator. This pipeline takes forward and reverses Illumina reads and truncates them on the restriction enzyme recognition sites, eliminating any possible hybrid reads coming from the ligation of distant sequences. Subsequently, HiCUP aligns the reads to the NCBI GRCh38 human genome assembly, using the Bowtie2 algorithm (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml). After mapping, the pipeline filters out common Hi-C artefacts, such as re-ligation of consecutive fragments or di-tags mapping in the same restriction site. Lastly, it filters PCR duplication and outputs a single BAM file which can be used for the following statistical analysis. Reference digested genome (GRCh38) used in the alignment process was obtained from the hicup_digester tool.

Identify statistically significant interactions, CHiCAGO pipeline

DNA interactions, functional to gene expression, decrease in number with the increasing genomic distance from the viewpoint (Cairns et al. 2016b). CHiCAGO adopts this characteristic and translates it to the sequencing reads. Indeed, the software models these random interactions (i.e. null-hypothesis) as a function of two variables: Brownian movement and technical noises. The interactions due to Brownian movements are modelled using a negative binomial random variable, while technical noises are modelled with a Poisson random variable (Cairns et al. 2016b). The regions where the number of mapped reads exceeds the null distribution are considered statistically significant (Cairns et al. 2016b; Freire-Pritchett et al. 2021). The CHiCAGO parameters to use in the analysis change according to the restriction enzymes used in the library production (i.e. DpnII or HindIII; Code 2.4 and Code 2.5 respectively; Freire-Pritchett et al. 2021).

```
minFragLen=75L,
maxFragLen=1200L,
```

```
binsize=1500L,
MaxLBrowndist=75000L,
weightAlpha=24.5,
weightBeta=-2.16,
weightGamma=-21.2,
weightDelta=-9.2,
brownianNoise.seed=1989L
```

Code 2.4 | Parameters used in the CHiCAGO analysis for the DpnII digested libraries.

```
minFragLen=150L,
maxFragLen=40000L,
binsize=20000L,
MaxLBrowndist=1500000L,
weightAlpha=34.11573,
weightBeta=-2.586881,
weightGamma=-17.13478,
weightDelta=-7.076092,
brownianNoise.seed=1989L
```

Code 2.5 | Parameters used in the CHiCAGO analysis for the HindIII digested libraries.

The interactions I decided to keep as statistically significant after the CHiCAGO analysis have a less stringent CHiCAGO score > 4. The threshold was empirically determined via the observation of the number of interactions distribution (Fig. 2.4)

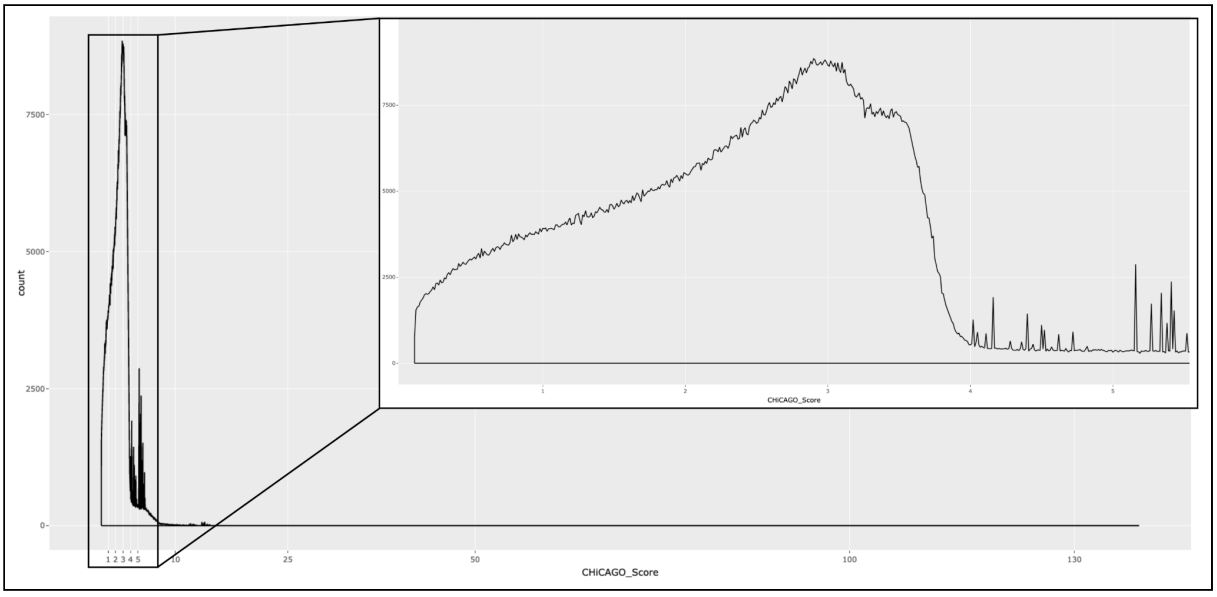


Fig. 2.4 | Number of interactions per CHiCAGO score. *On the x-axis, there is the CHiCAGO score; on the y-axis, there is the number of reads that support that interaction. There is a clear peak that characterises the interactions with a CHiCAGO score lower than four.*

TG Hi-C genomic features annotation

The definition of the genomic features, either for baits or preys, was achieved via the annotation with data from previously published studies (see below). If a region identified via the TG Hi-C experiments overlapped a region defined in the studies used as a reference, the label given in the reference studies was transferred to the TG Hi-C regions. The regions were considered overlapping even if just 1nt was shared between the two regions. When the regions were not directly overlapping, but in the 1 Kb genomic space surrounding the genomic feature, they were flagged as proximal to the genomic feature. The reason for this liberal annotation comes from a limitation of the Hi-C technology and chromatin interactions in general: interactions can be in a 2Kb window from the functionally active regions (Peter H. L. Krijger et al. 2020).

The reference studies used to define iMK genomic features were the outcome of the BLUEPRINT consortium (Adams et al. 2012; Stunnenberg et al. 2016, and Hirst 2016b). The features assigned to iMK interactions were transcription factor binding site (i.e. GATA1, FLI1, TAL1, CTCF) and regulatory regions (i.e. H3K27Ac and ATAC-Seq). These experiments have been downloaded from the European Genome-Phenome Archive (EGA; https://ega-archive.org/) under the data set ID EGAD00001001871. iHEP regulatory regions were defined according to the H3K27Ac ChIP-Seq unpublished data and experiments from Professor Ludovic Vallier laboratory. iEC regulatory features were provided by Doctor Matthew Sims.

The transcript levels specific to the relevant cell types for iHEP and iEC were provided by Professor Ludovic Vallier laboratory and Doctor Matthew Sims. The iMK transcript levels were obtained from RNA-Seq from the BLUEPRINT consortium (L. Chen et al. 2014, 2016).

Comparison with previously published MK interaction data

The statistics about the iMK interactions identified in previous studies are coming from (Javierre et al. 2016). The authors of the manuscript made available raw and analysed data on the osf platform, at the following link https://osf.io/u8tzp/. Javierre and colleagues aligned their data to the GRCh37 version of the human genome, while my data were aligned

to the version GRCh38. To properly compare the data, I lifted the genomic regions identified in my analysis to the previous assembly of the genome. I used liftOverBedpe (https://github.com/dphansti/liftOverBedpe) to convert the genomic coordinate of the identified regions from one assembly to the other. Finally, to limit the pcHi-C interactions to the same number of regions and genes, I subsetted the interaction file from Javierre et al. according to the chromosomal regions overlapping the promoters of the TG Hi-C regions.

Prioritisation score calculated for the variants in the TG Hi-C regions

This score has been used to select the variants to test in the functional validation experiments. The variants were ranked according to their characteristics and overlapping features. The formula to calculate the score was:

$$S_i = \sum_{i=1}^{n} gf_i \, log\frac{1}{af_i} \sqrt[3]{l_i}$$

Where gf is the result of the sum of all the scores associated with each genomic feature, the weight of the genomic elements considered the possibility that a genomic feature is pathogenic (Table 2.3). The af parameter stands for the allele frequencies observed in the NIHR cohort. The idea is that rarer variants have a higher probability of having large effect sizes. Ultimately, l is the length of the variants; larger variants have a higher chance to compromise the function of a region.

Evidence	Score	Evidence	Score
ATAC-Seq	2	CTCF Chip-Seq	2
H3K27Ac ChIP-Seq	3	ClinVar	5
GWAS	5	Splice site	8
Exon	10		

Table 2.3 | Weights of the functional elements. The weights have been adopted in the calculation of the prioritisation score.

To further increase the probability of selecting functionally significant regions, I also filtered the variants to be in areas that overlapped with RedPop regions (Turro et al. 2020b). RedPop is a software that looks for anticorrelation between the histone acetylation and open chromatin sequencing signals. The areas with the highest anticorrelation have a strong

likelihood of being the transcription factor binding site and, therefore, relevant to the regulatory effect of that region.

Statistical association of the non-coding regions to BTPD phenotypes

BeviMed is an R package that builds Bayesian models to associate rare variants to the phenotype of interest (i.e. Mendelian disease; Greene et al. 2017). I used it to calculate the associations of variants in the TG Hi-C regions to the BTPD phenotypes. BeviMed splits the cohort into case-control groups according to the phenotypes, and it can also model the expected mode of inheritance of the diseases (Greene et al. 2017). The cohort of cases and control has been limited to the unrelated European cases. The prior utilised for the Bayesian association have been set to take into account the non-coding space. The associations presented in chapter 5 focused on variants that have AF < 0.002 and occurred in the TG Hi-C prey regions. To be more inclusive, 500bp surrounding the TG Hi-C regions have been used in the statistical associations (Peter H. L. Krijger et al. 2020). In addition, autosomal dominant and autosomal recessive modes of inheritance have been tested for all the phenotypes. BeviMed outcome is a posterior probability of association that ranges from 0 to 1. The closer the association is to one, the higher is the probability that the variants and region in consideration are associated with the phenotype. All the BTPD phenotypes that have been tested in the BeviMed associations are listed in Table 2.4. The statistical association has been performed in collaboration with Professor Ernest Turro and Dr Daniel Greene.

Diseases and phenotypes tested
Roifman syndrome
Hermansy-Pudlak Syndrome
Gorham-Stout disease
Pseudohypoparathyroidism type 1B
Autistic behaviour with impaired platelet aggregation or Abnormal dense granules without Thrombocytopenia
Thrombocytosis (i.e. plt > 450)
Impaired collagen-induced platelet aggregation with abnormality of the skeletal physiology or reduced mineral density
Impaired collagen-induced platelet aggregation and abnormal platelet membrane protein expression with abnormality of the nervous system
Thrombocytopenia (plt < 150) without abnormality of the integument unless Subcutaneous haemorrhage

Abnormal bleeding with normal plt and one of joint hypermobility or increased susceptibility to fractures or osteoporosis or fragile skin
vWF antigen < 40 U/dL or reduced vWF activity or quantity
Normal platelet count but impaired arachidonic acid-induced platelet or thromboxane A2 agonist-induced platelet aggregation
Impaired thrombin-induced platelet aggregation with normal platelet count and other normal responses to agonists
Abnormal platelet count and abnormal response to at least four agonists
Abnormal thrombosis
Abnormality of erythrocytes and Increased mpv
Impaired ADP-induced or epinephrine-induced platelet aggregation with normal platelet granules
Hearing impairment and thrombocytopenia
Abnormality of limb bone morphology and thrombocytopenia
Fibrinogen < 1 g/L
Wiskott–Aldrich like phenotype
Nervous system abnormality and thrombocytopenia
Impaired ristocetin-induced platelet aggregation but other agonists responses are normal
Increased mpv without thrombocytopenia
Abnormal alpha granules and Abnormal dense granules
Asthma or eczema and at least one of abnormal platelet count, abnormal bleeding or impaired platelet aggregation
Neutropenia
Monocytopenia
Factor V < 0.7 IU and Factor VIII < 0.5 IU
Fibrinogen > 6 g/L
Factor V < 0.4 IU
Factor VII < 0.4 IU
Factor VIII < 0.4 IU
Factor X < 0.4 IU
Factor XI < 0.4 IU

Factor IX < 0.4 IU
Factor XIII < 0.4 IU

Table 2.4 | The phenotypes, derived from HPO terms, that have been used to group people in the case-control during the statistical comparison in BeviMed analyses.

Analysis of the variant effects on the CTCF binding sites

This experiment used all the variants that passed the filtering criteria described in the previous paragraphs and overlapped an MK CTCF ChIP-Seq region (L. Chen et al. 2016). The DNA sequence that flanks the variants was extracted, as a string of nucleotides, using the reference sequence (GRCh38) and the NIHR-BR BTPD cases. These regions were screened for the CTCF Jaspar motif (http://jaspar.genereg.net/; Jaspar matrix profile MA0139.1; Fig 2.5), and the maximal motif score was calculated. The motif score is a logarithmic likelihood ratio; higher scores indicate that the sequence is more likely to be a CTCF binding site than a random sequence. Two maximal motif scores were calculated per region, one for the reference sequence and one for the alternative sequence. Comparing the two scores relative to the same regions gives the effect of the variants on the CTCF motif. When the reference and alternative score difference is significant (i.e.>10 units), the variant has a notable impact on the CTCF binding capacity. Dr Nick Owens performed the calculation of the CTCF maximal motif score in reference and alternative regions.

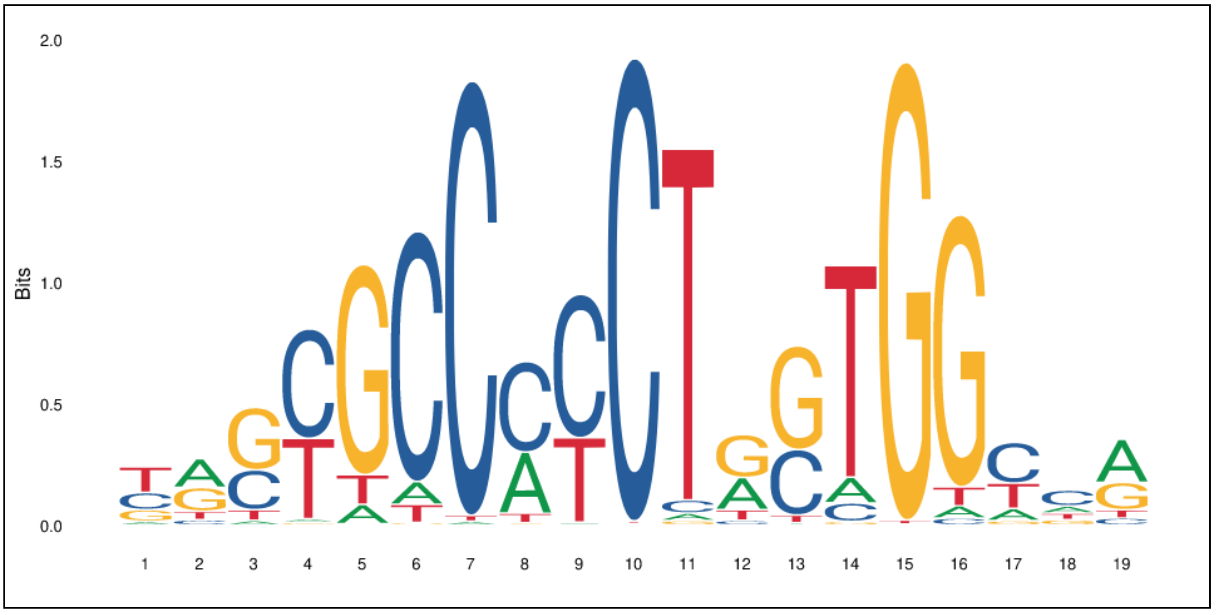


Fig 2.5 | The CTCF binding motif *("Matrix Profile: CTCF - MA0139.1 - from JASPAR 2018")*

Bootstrap experiment to estimate differences in the AF of the variants in TG Hi-C regions

To check whether the regions identified with the TG Hi-C approach are more conserved than the rest of the genome, I performed a bootstrap experiment to compare TG Hi-C regions on chromosome 11 to other randomly selected areas on the same chromosome. I focused on chromosome 11 because it has the greatest number of base pairs captured by the TG Hi-C system (4,000,592 bp) and the largest diversity of TG genes. Per each of the 25 AF sampling iterations, I created a random reference that had a similar size and distribution to the genomic space identified by the TG Hi-C experiments. Then, I extracted the AF of the variants present in these regions and compared their frequency distribution with the one of the TG Hi-C regions. The statistical test adopted for this comparison was the Kolmogorov–Smirnov test.

Promoters definition in the TG Hi-C experiments

The iMK promoter regions used in the TG Hi-C analysis have been empirically determined using ChIP-Seq and RNA-Seq experiments (L. Chen et al. 2016). Briefly, the method adopted determined the transcription starting site based on the transcripts identified in the MK RNA-Seq data. The transcripts were used only if the $\log_2$(FPKM) was larger than one $\log_2$FPKM. Once the transcription starting site was determined, the promoter regions

were defined using the H3K27Ac ChIP-Seq peaks overlapping that nucleotide. If no overlap was identified, the canonical promoter from ENSEMBL was adopted.

2.4 Molecular Biology methods

Cloning experiments

The cloning experiments were performed in NEB 5-alpha Competent E. coli (New England Biolabs) and transformed according to the manufacturer instruction. Briefly, bacteria were thawed on ice and then incubated for 30 minutes with 50ng of plasmid DNA. Thirty seconds of heat shock at 42°C allowed the plasmid DNA to enter bacterial membranes. To recover bacterial growth, they were incubated for one hour at 37°C with nutrient rich medium (SOC medium), then plate on agarose plated with the correct antibiotic selection. Restriction enzyme digestion and Sanger sequencing were used to assess the quality of the plasmid. In case of large quantity or better quality of the plasmid DNA was required, the plasmid DNA was extracted with QIAGEN MaxiPrep Kit according to the manufacturer protocol.

The reporter assay used two plasmids:

Plasmid	Experiment
pGL4.54[luc2/TK]	Reporter assay - Reporter gene - Contains the regions to test
pGL4.74[hRluc/TK]	Reporter assay - Normalisation gene

pGL4.54[luc2/TK] was digested using KpnI (New England Biolabs) according to the manufacturer instructions. After enzymatic digestion, the linearised plasmid was purified using DNA Clean & Concentrator™-5. The regions to test (Table 2.5) were cloned in the pGL4.54[luc2/TK] plasmid using the InFusion® protocol (Takara) kit, using as plasmid overhangs 5'-CTCATTTAGACTCAG-3' for the primer reverse and 5'-TGGCCTAACTGGCCG-3' for the primer forward. This approach allowed me to keep the same directionality of the reference human genome. The synthesis of the regions in Table 2.5 was impossible due to their complexity (i.e. GC content or repetitiveness of the regions). For this reason, I decided to clone them using PCR amplification. The genomic DNA template was extracted from controls samples and cases that carry the variant of interest (Fig.5.2.A and Fig.5.2.C)

Region	Assigned gene	PF sequence	PR sequence
A	VWF	GCAGAACATGGGTGCCGGTGA	CGGGAGGCGGAGGTTACAGTGA

B	CD9	AACAGGGTGGCTGCGGGGAG	AGGGGAGGTTTGGAGGCTGCC
C	THBD	AGCGGTGTTATCAGGGGCCCA	TCCATCGTGCGGCCCTGTCC
D	MPIG6B	GCCCAACCCACCAAGCAGCT	AGCAGACCCCTCACAGACCCCT
E	ABCC4	GCGGTCTTCTGGCAGCACTGA	GGTCCTCTGTTCTTTGGGCCCATCC
F	SERPINE1	GCGTGCCCAGCTCTTCACCC	TGCGGCTGTGAGTCACCCTGT
G	PTGS1	TGGGAGCTGGGCAGTGGGTG	TGACCTGGGCAGCAGAGTCCT

Table 2.5 | Genes that have been tested with the reported assay and primers used to amplify such regions.

The oligonucleotides used to silence the regions in the dCasKRAB experiments are listed in Table 2.6. These regions were cloned in the LentiguidePuro plasmid (Addgene 52963; Sanjana, Shalem, and Zhang 2014) after linearisation with the BsmBI restriction enzyme. The 5'-CACCG...C-3' on the sense strand and 5'-AAACCC...C-3' on the antisense strand were used to create overhangs with the restriction-enzyme-created ends. Ultimately, these plasmids were used to produce lentiviruses and infect hiPSc to obtain a stable expression of the sgRNAs.

Region	Oligo sense	Oligo antisense
MCFD2/BC200	CACCGATGCTCAAGGATTACTCCGTAGG	AAACCCTACGGAGTAATCCTTGAGCATC
MCFD2/BC200	CACCGCACTAGCAGGACTGCAACCGAGG	AAACCCTCGGTTGCAGTCCTGCTAGTGC
MCFD2/BC200	CACCGCCAGCGTACATTTGCCTCATGGG	AAACCCCATGAGGCAAATGTACGCTGGC
BLOC1S6_BV	CACCGGTCGCCTAGCAAAGTACAGGAGG	AAACCCTCCTGTACTTTGCTAGGCGACC
BLOC1S6_BV	CACCGCAGTTGCGCGAGACTCCAACGGG	AAACCCGTTGGAGTCTCGCGCAACTGC
BLOC1S6_BV	CACCGCGGTTGTTATTTGGAACACCTGG	AAACCCAGGTGTTCCAAATAACAACCGC
BLOC1S6_SV	CACCGAAGATCTCGCTACCCGACGGCGG	AAACCCGCCGTCGGGTAGCGAGATCTTC
BLOC1S6_SV	CACCGCTAGGTTGCTATTCCAACGGCGG	AAACCCGCCGTTGGAATAGCAACCTAGC
BLOC1S6_SV	CACCGCTGGCGGAGCGACCACCACGTGG	AAACCCACGTGGTGGTCGCTCCGCCAGC
THBD	CACCGAGCTCTAGACGACGTAGCGTGGG	AAACCCCACGCTACGTCGTCTAGAGCTC
THBD	CACCGGGCCCAACCATTACTTAAGTGGG	AAACCCCACTTAAGTAATGGTTGGGCCC
THBD	CACCGACATCTGTTAAACTCTCGATAGG	AAACCCTATCGAGAGTTTAACAGATGTC

Table 2.6 | Oligos for sgRNAs. These have been used to silence the regions in the dCas9KRAB experiments.

The dCasKRAB system also needed the expression of the dCasKRAB and rtTA components in the same cell system. This co-expression was obtained via the cotransfection

of lentiviruses containing the rtTA and the dCasKRAB viruses. The 2 plasmids required for the LV production were obtained from Addgene:

Plasmid	Experiment
pCSIIhyg-rtTA (Addgene 139480)	Expresses rtTA for TET-ON system
TRE-KRAB-dCas9-IRES-GFP (Addgene 85556; Fulco et al. 2016)	KRAB-dCas9 expression TET regulated

Sanger sequencing

Plasmids and PCRs were routinely sequenced to check the quality of the amplification or cloning. The sequencing was performed by Source Bioscience company. The plasmids were shipped at 100ng/ul concentration and PCR amplicons at 10ng/ul. Custom designed primers were provided at a concentration of 3.2pmol

Nucleofection experiments

Nucleofection™ was used to express the reporter plasmids in the imMKCLs. The nucleofection experiments were performed following the manufacturer guidelines. imMKCLs were nucleofected using the nucleofector machine program “Y1”. The Nucleofection reaction was set to have $8x10^5$ cells per cuvette, and in each cuvette, cells were resuspended in 100µl of nucleofector solution (82% P2 and 18% P1). The amount of DNA that was used per nucleofection was about three micrograms per condition. If multiple plasmids were used simultaneously, then the amount of each plasmid was scaled down to have three micrograms of DNA total.

Lentivirus production

To produce the lentiviruses (LV) used in the dCasKRAB experiments, I used HEK293T cells. One almost-confluent petri dish (~6 Mln cells) was used for the production of each virus. The vector plasmid was co-transfected with Pax2 and VSV-G packaging vectors. The amount of DNA used in the production of LV was as follows: 8µg of PAX2, 2.8µg of VSV-G and 20µg of insert vector. To transfect the DNA into the cells, I used 75µl of Polyethylenimine [1mg/ml] (PEI) and resuspended the DNA-PEI mix in 1 ml of DMEM basal medium. Then, the mix was vortexed for 15 seconds and incubated at room temperature for 20 minutes. After this incubation time, the DNA-PEI complex was added onto the cells and set for 6-16 hours. The medium was replaced with a fresh one at the end of the transfection,

and cells were incubated for two days (5% CO2, 37°C) to allow the production of LVs. LVs were concentrated using the Lenti-X reagent following the manufacturer instructions.

Reporter assay experiments

The dual-luciferase reporter assay allowed me to study the role of a regulatory region thanks to the expression of a reporter gene (i.e. firefly luciferase; McNabb, Reed, and Marciniak 2005). The expression of the reporter gene is under the control of the cloned DNA region. The dual-reporter genes used in this study also utilise a secondary control gene (i.e. renilla luciferase) used to set the baseline expression and controls for transfection efficiency (Matthews, Hori, and Cormier 1977; Wood et al. 1984). The plasmids pGL4.54[luc2/TK] and pGL4.74[hRluc/TK] were co-nucleofector into imMKCLs. The pGL4.54[luc2/TK] contained the different regions I wanted to test, while pGL4.74[hRluc/TK] stayed the same in all the experiments. After nucleofection, cells were incubated for two days (5% CO_2, 37°C) to allow enough time for the cells to express the plasmid genes. After this two-day period, cells were washed two times in a fresh medium and then lysed with 100µl of lysis buffer. The lysis step occurred for 20 minutes at room temperature on a rocking platform. To increase the lysis, cells underwent two freeze and thaw cycles. All the luciferase absorbance steps were performed in 96-well polystyrene plates with solid white bottoms. The recording machine was the SpectraMax M5. Each well had a 100µl LARII reagent, to which I added 20µl of lysate. After this step, the firefly emission (560nm) was recorded and quenched with the “Stop&Glo” reagent. After this last reagent was added, the renilla emission (480nm) was measured. The regulatory effect of the tested regions was reported as relative fluorescence units, and it was calculated as the average of reporter gene expression (i.e. Luciferase) normalised by the average expression of the renilla luciferase.

dCasKRAB experiments, RNA extraction and qPCRs

The CRISPR interference method is derived from the CRISPR/Cas9 technology (Fig.5.1; see also chapter1.4.4 for an accurate description of the CRISPR interference; Jinek et al. 2012; Fulco et al. 2016). The catalytic activity of the Cas9 enzyme was inactivated with two point mutations. Moreover, the Cas9 protein was fused to a Krüppel associated box (KRAB) domain. The KRAB effector domain binds co-repressor proteins and induces the heterochromatin state and transcriptional repression in the region of interest (Margolin et al. 1994; Urrutia 2003; Huntley et al. 2006). hiPSC, which stably expressed rtTA, dCasKRAB and sgRNAs, were differentiated in iMK using the protocol described before (Moreau et al. 2016). The differentiation medium has always been supplemented with antibiotics (see

hiPSc section chapter 2.2) and doxycycline to express all the required components. The mature iMK differentiated cells, which had regions that have been silenced with the dCas9KRAB system, had their RNA extracted and transcript checked with qPCR.

The RNA was extracted using the Monarch® Total RNA Miniprep Kit (New England Bioscience). One million differentiated cells (per condition) were pelleted, washed and resuspended in 300µl of lysis buffer. The lysed cells were aliquoted in a “light-blue” column and spinned for 30 seconds to remove the gDNA. Then, I added one volume of ethanol and transferred the mix to the “dark-blue” columns. I spun for 30 seconds and washed with the washing buffer three times. Ultimately, RNA was diluted in 50µl of nuclease-free water and snap-frozen. Retro-transcribed to cDNA was performed using 40µl of RNA and the “High capacity cDNA reverse transcription” kit following the manufacturer instructions. The qPCR was performed using a Stratagene Mx3000P qPCR machine and “Luna Universal qPCR Master Mix” (New England Biolabs). The primer used to quantify the transcripts are in Table 2.7.

Region	Primer forward (PF) sequence	Primer reverse (PR) sequence
MCFD2	CCTGTTACCCAGTTCCAAAGT	GATAAAGACATACCCGAGACTGG
THBD	GGGTGATTAGAGGGAGGAGAA	TGTAACGAAGACACAGACTGC
BLOC1S6	GGTGAGCAAGATGCCAGATAG	CTAAAGCTTGGTGGAGGTAGTG
GUSB	ACGTGGTTGGAGAGCTCATT	CTCTGCCGAGTGAAGATCCC

Table 2.7 | Primers for the qPCRs amplify the cDNA of the genes silenced with the dCasKRAB system. *GUSB* is the reference gene that was used to normalise the transcript expression.

The differences in the gene expression were calculated with the formula $2^{-\Delta\Delta Ct}$, where Ct is the qPCR cycle threshold. The $\Delta\Delta$ symbols refer to the Ct differences between the silenced sample (i.e. the one expressing the gRNA and dCasKRAB) and the Ct of the not-treated sample (i.e. the wild-type cell line). The ΔCt is given by the Ct differences between the average Ct of the gene of interest (e.g. *MCFD2*) and the reference gene (i.e. *GUSB*).

Softwares

The statistical tests performed in this thesis have all been executed using the R software environment (https://www.r-project.org/). The version used for this thesis was 3.6.3

and it was run under the Scientific Linux (v7.9) installed onto the University of Cambridge High-Performance Computing Service.

2.5 Materials

Reagent	Supplier	ID
T4 DNA ligase	NEB	M0202
T4 DNA ligase buffer	NEB	M0202
Phusion® HF DNA Polymerase	NEB	M0530L
Kapa library quantification kit Illumina	Roche	KK4824
Glucogen	Thermo fisher	R0561
Dual-Luciferase® reporter assay system	Promega	E1910
Luna® Universal qPCR master mix	NEB	M3003X
High capacity cDNA Reverse transcription kit	Applied biosystems	4368814
DpnII	NEB	R0543M
NlaIII	NEB	R0125L
Proteinase K from Tritirachium album	Sigma-Aldrich	P8044-250MG
NimbleGen SeqCap hybridization and Wash Kit	Roche	05634253001
Klenow Fragment (3’ -> 5’)	NEB	M0212L
DNA Polymerase I, Large (Klenow) Fragment	NEB	M0210L
Biotin-14-dATP	Invitrogen	19524-016
100 mM dTTP	Invitrogen	55085
100 mM dCTP	Invitrogen	55083
100 mM dGTP	Invitrogen	55084
xGen® Universal blockers-TS Mix	IDT	1075476
Protease inhibitor cocktail	Thermo	78429
T4 DNA polymerase	NEB	M0203L
Dynabeads® MyOne™ Streptavidin C1	Invitrogen	650.01
37% formaldehyde	Sigma-Aldrich	25259
LentiGuide-Puro	Addgene	52963
pCSIIhyg-rtTA	Addgene	139480
TRE-KRAB-dCas9-IRES-GFP	Addgene	85556

pGL4.54[hLuc]	Promega	E506A
pGL4.74[hRluc/TK]	Promega	E692A
rhVTN-N	ThermoFisher	A14700
SOC Outgrowth Medium	NEB	B9020S
Blasticidin S hydrochloride	AlfaAesar	J67216
Amaxa™ Human Stem Cell Nucleofector™ Kit2	Lonza	VPH-5022
DMEM/F12	ThermoFisher	11330-032
L-Ascorbic Acid 2-phosphate	Sigma-Aldrich	D1159-100MG
Insulin-Transferrin-Selenium (ITS -G)	ThermoFisher	41400045
FGF2	Cambridge Stem Cell Institute	N.A.
BMP4	R&D	314-BP-010
Rock inhibitor Y-27632	Sigma-Aldrich	Y0503
rhTPO	bio-techne	Bulk 288-TP
Protamine Sulphate	Sigma-Aldrich	P4505
TrypLE	Life Technologies	12563029
CellGro	CellGenix	0020902-0500
rLV-WPT-GATA1	Vectalys	N.A.
rLV-WPT-FLI1	Vectalys	N.A.
rLV-TRIPU3-TAL1	Vectalys	N.A.
EasySep™ Release Human PE Positive Selection Kit	StemCell Technologies	17654
rhSCF	Gibco	PHC2116
CD235a FITC Mouse Anti-Human	BD Pharmingen	559943
CD42b PE Mouse Anti-Human	BD Pharmingen	555473
CD41a APC Mouse Anti-Human	BD Pharmingen	559777
CD324 PerCP-Cy™5.5 Mouse Anti-Human	BD Pharmingen	563573
CD326 PE-CF594,Mouse,Anti-Human	BD Pharmingen	565399
Tylosin	ChemCruz	sc-253815
Penicillin/Streptomycin	Gibco	15140-122
MEM NEAA	Gibco	11140-035
Puromycin	Gibco	A11138-031
Hygromycin B	Invitrogen	10687010

StemFlex™ Medium	Gibco	2279513
DMEM - High Glucose	Sigma	RNBJ6691
DMEM/F-12(Ham)	Gibco	11330-032
Sodium Bicarbonate (7.5%)	Gibco	25080-094
Agarose	Sigma	A9539-500G
Dimethyl sulfoxide	Sigma	RNBH9956
Sodium Acetate solution	Sigma	71196-100
Nuclease Free Water	Qiagen	1039498
GeneRuler 1Kb, DNA Ladder	Thermo Scientific	SM0313
DNA Clean & Concentrator™-5	ZYMO RESEARCH	D4014
QIAquick® Gel extraction kit	QIAGEN	278706
QIAquick® PCR Purification KIT	QIAGEN	28104
DNeasy® Blood & Tissue Kit	QIAGEN	69506
QIAprep® Spin Miniprep Kit	QIAGEN	27106
HiSpeed® Plasmid Maxi Kit	QIAGEN	12663
Matrigel®	Corning	354230
Fibronectin	Corning	356008
Y-27632 Rock inhibitor	Millipore	SCM075
T225 Flasks	Corning	3293
F-12	Thermofisher	31765027
IMDM	Thermofisher	21980032
Lipid concentrate	Thermofisher	11905031
Transferrin	Sigma	T1147
Chiron (8μM)	Cambridge Stem Cell Institute	CHIR-99021
Ly294002 (20mM)	Adooq Bioscience	A01547
StemPro-34	Thermofisher	10639011
Forskolin	Sigma	F6886
VEGF-A	Gibco	9393
L-Ascorbic acid	Cambridge Stem Cell Institute	N.A.
CD144	BD Pharmigen	560411
TRA-1-60	Merck Millipore	FCMAB115F

SSEA-4	BD Pharmigen	560128
Human oncostatin	BioTechne	295-OM
Hepatocyte growth factor	Peprotech	100-39H
Lenti-X concentrator	Takara	PT4421-2
IGV	N.A.	N.A.
Benchling	N.A.	N.A
FlowJo	BD Bioscience	N.A.
raremetal	N.A.	N.A.
bwa	N.A.	N.A.
plink2	N.A.	N.A.
somalier	N.A.	N.A.

Chapter 3

VarioPath: investigating the role of pathogenic variants in disease aetiology

3.1 Introduction and aims of the chapter

Genotyping of the DNA samples from large cohorts, such as UKB, has begun to better define the relation between common disease pathophysiology (i.e. phenotype) and genotype. In parallel, studies that aim to better understand the aetiology of rare diseases, for instance the 100,000 Genomes Project, have incrementally increased the number of identified P/LP variants (Downes et al. 2019; Karczewski et al. 2020; Turro et al. 2020b; Thaventhiran et al. 2020; Taliun et al. 2021). However, the role of these variants in a healthy population cohort, like UKB, has yet to be determined (MacArthur and Tyler-Smith 2010; Karczewski et al. 2020). Moreover, these large cohort studies allow for the first time to explore the interplay between common and rare variants of the P/LP categories (Vuckovic et al. 2020; Thaventhiran et al. 2020; Goodrich et al. 2021).

The American College of Medical Genetics and Genomics (ACMG) recommends reporting secondary genetic findings (i.e. spurious discoveries of pathogenic variants), in high-throughput sequencing studies, for only 59 genes (ACMG Board of Directors 2015; Kalia et al. 2017), and the 100,000 Genomes Project has been even more restrictive on reporting the variants. The main reason to curtail the reporting of secondary findings is that the effect sizes and penetrance of rare variants in genes implicated in rare diseases have not yet been defined.

Some studies shifted the genetic aetiology of rare conditions from variants with large effect size, affecting the function of a single gene, to the aggregated effect of many variants with minor effects in a large number of genes (Khera et al. 2018; Oetjens et al. 2019). These studies built their hypothesis on the idea of Fisher's infinitesimal model, which states that every variant and gene contributes minimally to a phenotype (Fisher 1919; Barton, Etheridge, and Véber 2016). Boyle and colleagues recently revised Fisher's infinitesimal model based on the observations made by GWAS (Boyle, Li, and Pritchard 2017). In their 2017 manuscript, the authors presented the omnigenic model. They propose that a small number of common variants in a cluster of disease-relevant highly connected genes have relatively large effect sizes on the phenotype, whilst other phenotype-associated variants that map on genes in the periphery of this gene network have a far smaller effect (Boyle, Li, and Pritchard 2017).

In 2016, UKB released the genotype data generated with a genome-wide array for half a million participants (Cortes et al. 2017; Bycroft et al. 2018). These results have been used to define PRS for many common diseases for medically relevant traits, like the FBC parameters (Vuckovic et al. 2020). However, the effect of rare variants (AF < 1/1,000) could not be studied because the accuracy of rare variant imputation from the array genotyping technology is limited (Van Hout et al. 2020). Recently, UKB released the WES data for 200,624 participants, offering the most extensive resource that links genotype to phenotype at the level of the individual participants. Despite a few known existing biases in the sampling strategy (Fry et al. 2017; Munafò et al. 2018; Haworth et al. 2019), this new resource offers for the first time an opportunity to define the association between rare variants (MAF < 0.001) and phenotypes in a prospectively defined population cohort.

In this chapter, I describe how I made use of the UKB cohort to study the phenotype of individuals who are carriers of P/LP variants in one of the 93 BPTD genes (Fig. 3.1). Amongst the different physiological processes and phenotypes that can be utilised as a model, haemostasis has several advantages: (i) protein functions for many of the BPTD genes are exceptionally well characterised, (ii) the phenotypes of DVT and PE are relatively well captured in the UKB HES records, (iii) the FBC measurements of UKB participants have been extensively quality controlled and corrected for batch effects (William J. Astle et al. 2016), (iv) the genetic architecture of haematopoiesis has been thoroughly explored by GWAS (Soranzo, Rendon, et al. 2009; Soranzo, Spector, et al. 2009; Gieger et al. 2011; van der Harst et al. 2012; William J. Astle et al. 2016; Vuckovic et al. 2020), and (v) BPTDs was one of the rare disease domains included in the rare diseases pilot study for the 100,000 Genomes Project (Turro et al. 2020b). Human genetics widely assume that individuals who are carriers of autosomal recessive (AR) P/LP variants have no discernible phenotypic consequences. I collected P/LP variants from available databases (see below), determined their prevalence in UKB and, if present in an adequate number of participants, calculated their effect sizes for the four FBC platelet traits (plt, mpv, pct, pdw) and DVT/PE (VTE together). This allowed me to determine, for the first time, the effect sizes of a subset of the P/LP variants in BPTD genes and to identify a novel phenotype for carriers of LoF variants in the *MPL* gene, which encodes the receptor for thrombopoietin. Subsequently, I studied the contribution of PRS to VTE manifestation and platelet traits, using previously published PRS (Klarin et al. 2019; Vuckovic et al. 2020). Then, building on these analyses, I explored the omnigenic model proposed by Boyle and colleagues in a newly developed protein-protein interaction (PPI) network (Barrio-Hernandez et al. 2021; Schwartzentruber et al. 2021) and concluded in favour of this theory. I also computationally determined the deleteriousness of

missense P/LP variants, in a subset of the BPTD genes, based on the aminoacid change and protein structure information. Finally, I correlated the deleteriousness inferred from protein information to the effect sizes calculated in the UKB cohort.

The VarioPath work described in this chapter is a collaborative effort. Dr Elspeth Bruford (EMBL-European Bioinformatics Institute, EBI) and Dr Karyn Megy (NIHR BioResource at the University of Cambridge) led the gene selection and curation process. The extraction of phenotype data from UKB resulted from a collaboration with the Department of Public Health and Primary Care at the University of Cambridge, with help and suggestions from Dr Luanluan Sun and Prof. Emanuele di Angelantonio. Dr Dragana Vuckovic from Imperial College London has supervised the statistical analysis. The mapping of the variant on their protein crystal structure and the estimation of their deleteriousness have been performed by the collaboration with Professor Dame Janet Thornton and Dr Roman Laskowski.

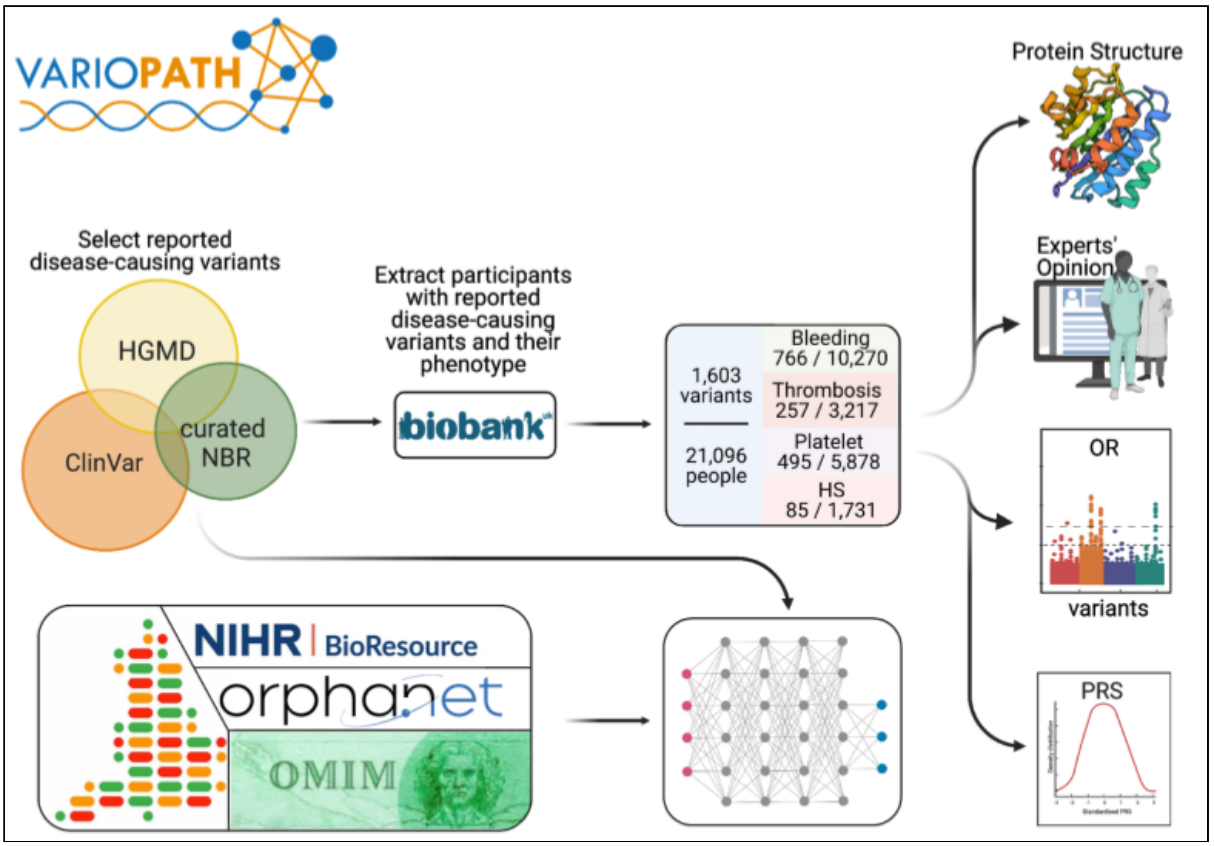


Fig. 3.1 | Workflow adopted in the VarioPath project. *The pathogenic and likely pathogenic variants have been collected from HGMD, ClinVar, and NBR. UK Biobank WES data has been used to calculate AF in the UK population and define the phenotype of the disease domain selected. Gene and variants information has been used to create PPI networks. HS: hereditary spherocytosis; OR: Odds Ratio; PRS: polygenic risk score.*

3.2 Genes and inherited diseases

Literature review and query of the major gene-relevant disease databases (see chapter 2.1) returned 4,851 genes associated with human phenotypes and rare diseases (see chapter 2.1). The recorded mode of inheritance for these genes is mainly autosomal recessive (AR; 2,123 genes), followed by autosomal dominant (AD; 1,118 genes; Fig. 3.2). Strikingly, the third-largest group by number of genes is "not known" (951 genes), meaning that for almost 20% of the inherited disorders, the mode of inheritance is unknown or not yet recorded in the databases used for the project. In total, 198 genes on chromosome X are implicated with rare diseases, with 125 genes being associated with disease phenotypes mainly in males (e.g. F8, F9) and 47 genes being causal of dominantly inherited disorders with disease phenotypes in both males and females.

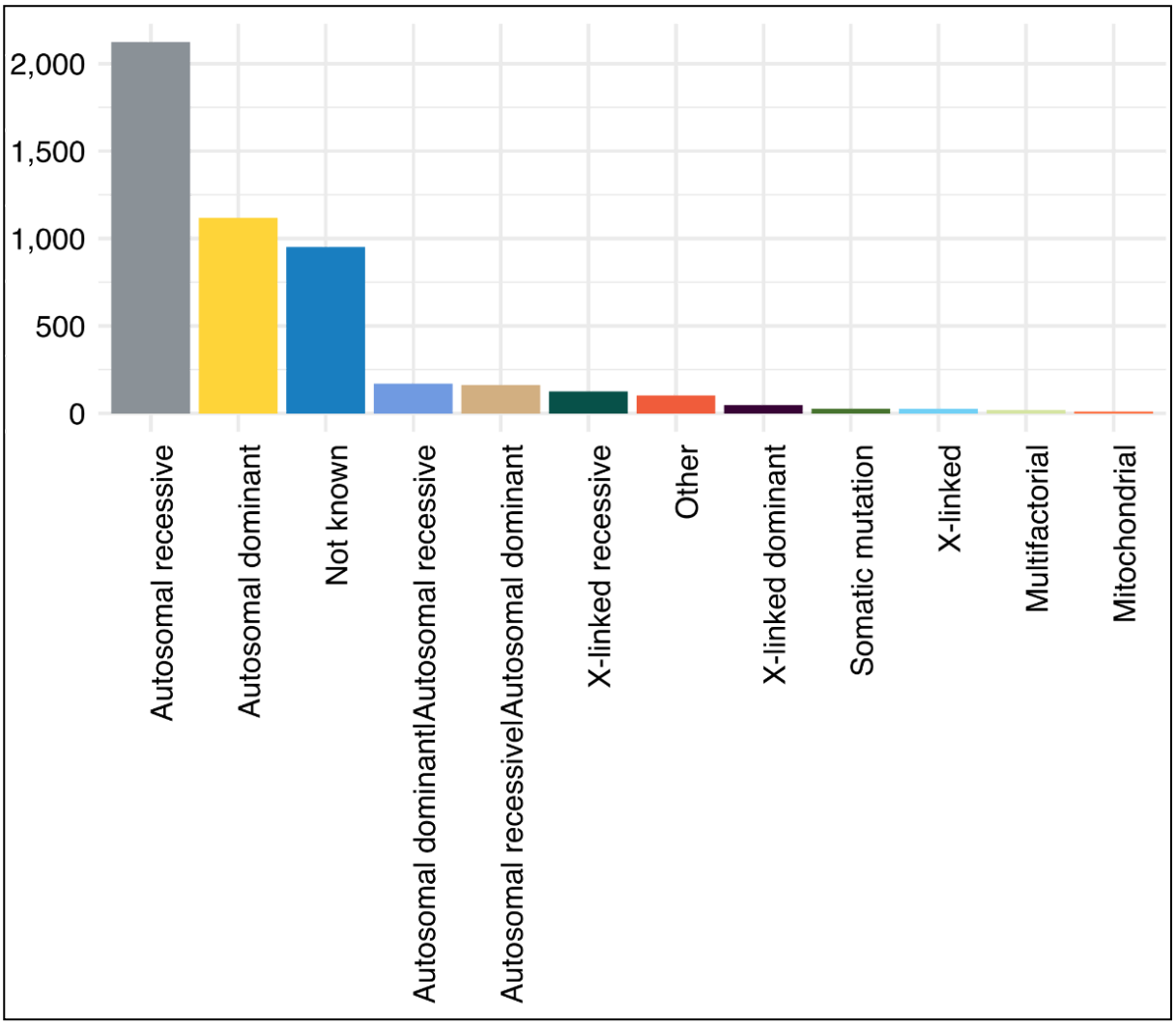


Fig. 3.2 | Distribution of the mode of inheritance for genes linked to human disease. *The difference between "Autosomal dominant|Autosomal recessive "and "Autosomal recessive|Autosomal dominant" is that the mode of inheritance listed first is the most common associated with that gene.*

The 4,851 genes can be grouped in 34 rare disease domains, involving several tissues and organs (Fig. 3.3). The number of genes per domain varies, with significant differences, as a function of (i) the dimension and complexities of biological pathways involved in a particular disease, (ii) how well characterised the domain is and (iii) how general the domain description is. The two extremes are intrahepatic cholestasis of pregnancy, with only two genes, and neurology and neurodevelopmental disorders, a broad category including many pathologies and 2,021 genes.

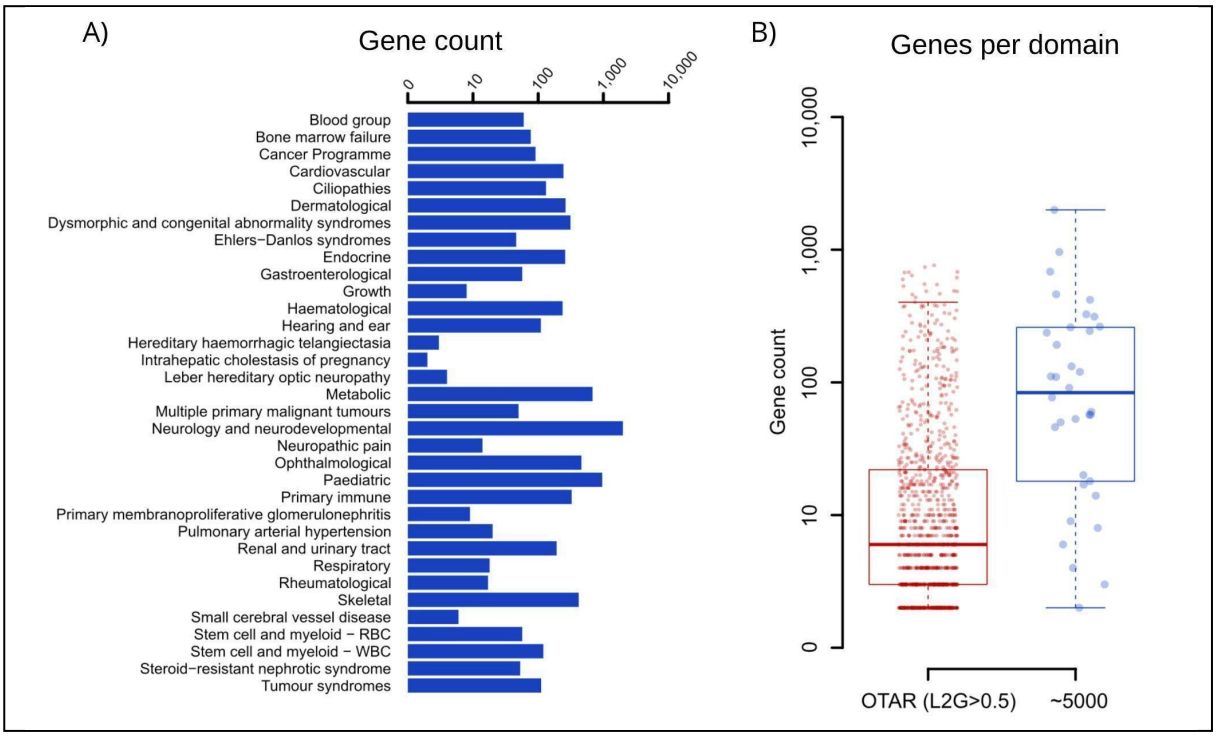


Fig. 3.3 | Number of genes per rare disease domain. *Classification of the domain according to the sources. B. Comparison of the genes per domain defined in VarioPath (blue) to the Open Targets (OTAR) definition used in the L2G algorithm (red).*

Interestingly, several genes belong to multiple rare disease domains, suggesting that a gene and its associated molecular pathway are involved in the aetiology of several pathologies (Fig 3.4). The largest intersection of genes is between "paediatric" and "neurology and neurodevelopmental" disorders, possibly because of the large number of genes that are within these two domains (973 and 2,021 genes respectively), and because most neurodevelopmental disorders tend to manifest at a young age (Fig 3.3; Scandurra et al. 2019; Morris-Rosendahl and Crocq 2020). Another interesting group is composed of "cancer programme"," haematological", "bone marrow failure", "stem cell myeloid", "dermatological" and "primary immune" (Fig. 3.4). Except for dermatological conditions, these domains have their common root in haematopoiesis. The dermatological disease domain may be linked to these haematological conditions because of the overlap between the skin and infections or immune system disorders, i.e. primary immune disorders and autoimmune diseases (Richmond and Harris 2014).

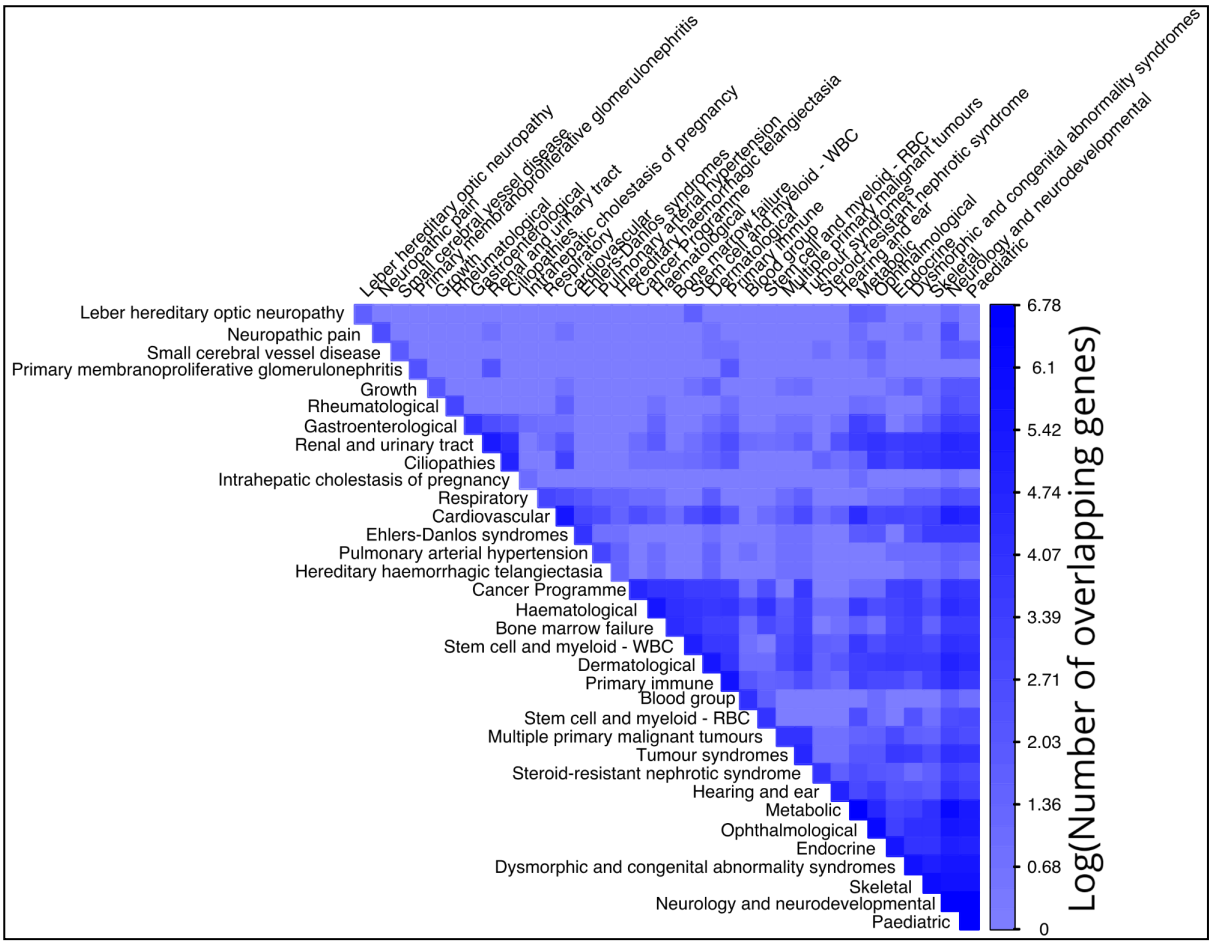


Fig. 3.4 | Heatmap showing the log number of genes overlapping between different disease domains.

An attempt to cluster the rare disease domain genes based on the number of overlapping genes is less informative because it is impossible to estimate the relationship or distance between two domains that lack any overlapping gene (Fig. 3.5). For example, it is impossible to evaluate the relation between "rheumatological" and "neuropathic pain" because there are no shared genes (Fig. 3.5,B). To overcome this analytical limitation, it is possible to apply a diffusion approach to a PPI network, which measures the distance between two networks even if they do not share any genetic component (Choobdar et al. 2019; Schwartzentruber et al. 2021). This method calculates the distance between two pathways based on the number of edges and nodes that link their proteins. Indeed, using the diffusion approach, I estimated disease domain distances (inferred from the protein network distances; Fig. 3.5). For example, the domains "rheumatological", which remained unclustered using the gene overlapping hierarchical approach (Fig. 3.5,B), can be positioned in proximity to the closest disease domains (Fig. 3.5,A).

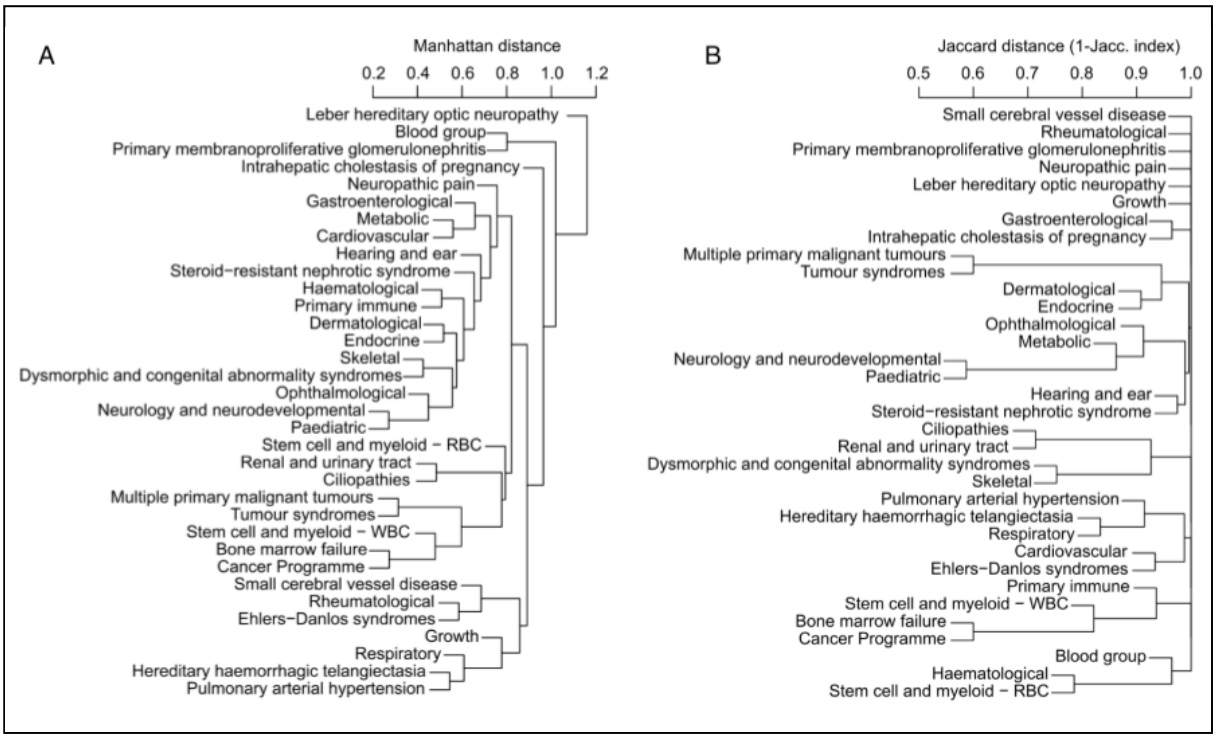


Fig. 3.5 | Hierarchical clustering of the rare disease domains using Manhattan distance (A) and Jaccard distance (B; Barrio-Hernandez et al. 2021)

3.3 Pathogenic variants

All the P/LP variants have been collected with an approach similar to the one employed for genes (see chapter 2.1). The resources used in this thesis are mainly HGMD, ClinVar and NIHR BioResource (NBR; see chapter 2.1 for more details). These three resources combined produced a list of 299,632 unique P/LP variants. The overlap of variants between the different resources is presented in Fig. 3.6. The most considerable overlap (65,503 variants) is between HGMD and ClinVar, the two most comprehensive resources.

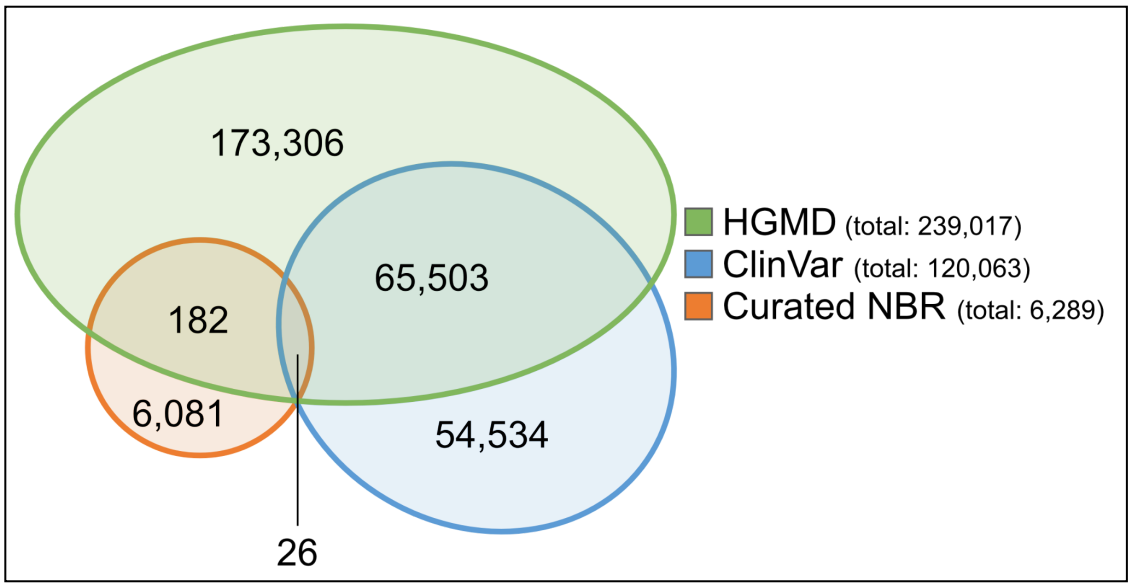


Fig. 3.6 | Number of pathogenic (P) and likely pathogenic (LP) variants and their intersection across the resource used in the VarioPath project.

The catalogued P/LP variants mainly consist of SNVs and small INDELs, 210,466 (70.2%) and 82,204 (27.4%) respectively, which cumulatively are ~98% of the entire collection of variants (Table 3.1). The retrieval of P/LP variants from the databases confirmed the relatively under-representation of disease-causing structural variants (SVs). The main explanation is that SVs are not reported systematically in HGMD and ClinVar. The major contributor to the rare-disease-causing SVs is the Deciphering of Developmental Disorders project (DDD; Firth, Wright, and DDD Study 2011; Wright et al. 2015), but their efforts are mainly limited to the neurology and neurodevelopmental disorders domain (Firth, Wright, and DDD Study 2011). For this reason, it was decided to exclude SVs from further analysis in the VarioPath project. A better assessment of their clinical role can be performed once the whole genome sequencing data of UKB will be released.

Total number of variants	Number of MNV	Number of large DEL	Number of INDEL/frameshifts	Number of SNV
299,632	5,113	1,849	82,204	210,466

Table 3.1 | The number of variants in the VarioPath project split by variant type. MNV: multi-nucleotide variants; DEL: deletion, INDEL: insertion and deletion; SNV: single nucleotide variants.

After variant effect predictor (VEP) annotation of the P/LP variants, the vast majority of the variants showed "HIGH" (133,731; ~44%) or "MODERATE" (129,942; ~43%) deleterious consequences on their transcripts (e.g. premature stop codon or missense variants; see chapter 2 Table 2.1 for the complete list; Fig. 3.7). Considering the source and the selection process of the P/LP variants, the observed distribution of the variants in the deleteriousness categories is expected. However, there are ~10,000 P/LP variants where VEP predicts a minimal detrimental effect on the transcript (i.e. "MODIFIER" or "LOW"; Fig. 3.7). This observation is either because the presence of the variant in the databases is not correct or the VEP prediction is not accurate.

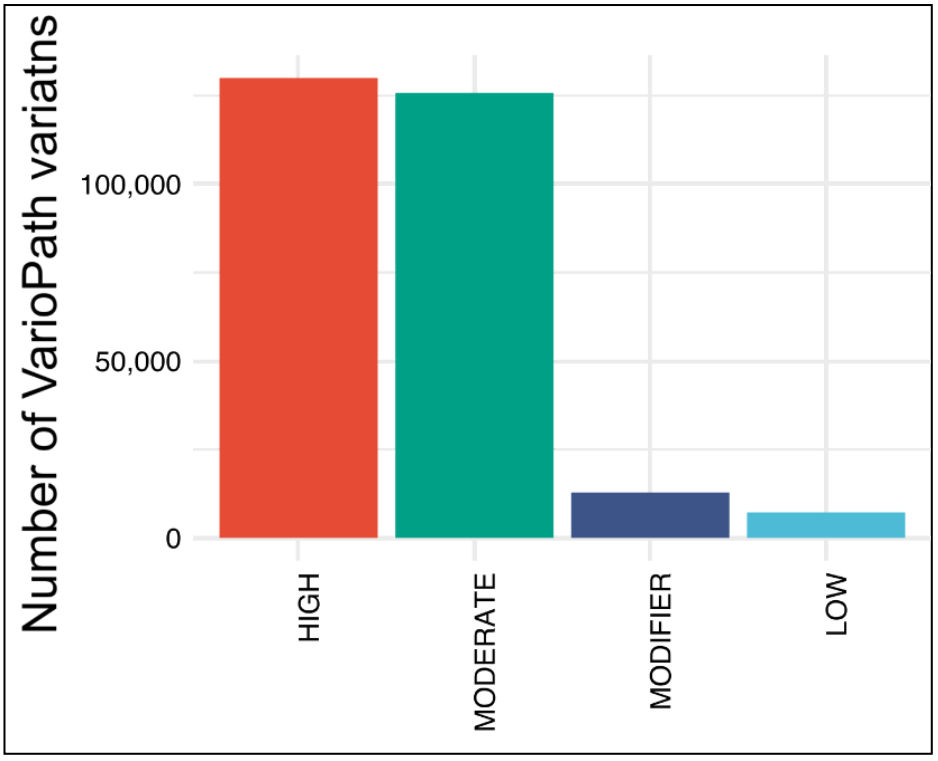


Fig. 3.7 | Predicted pathogenicity of the variants calculated using the VEP algorithm. *The definition of variants within each effect group are listed in chapter 2.*

3.4 Pathogenic variants in UKB

Out of the 299,632 P/LP variants (hereafter 300K) considered, 52,735 were observed in at least one participant of the unrelated European UKB cohort. When considering the entire list of variants, without any AF filter, the number of P/LP variants per individual is, on average, 147 (range: 107-191; sd: 9.8). When constraining the query on lower AF, the average number of pathogenic variants per individual is 26.8 (range: 5-59; sd: 5.36) for P/LP with an AF ≦ 0.01 and 5.9 (range: 1-26; sd: 2.68) variants per individual if AF ≦ 0.001 (Fig. 3.8). A small number of P/LP variants (19) have an AF > 0.9. This is a typical database error, where the reported effect allele for the P/LP variants is the reference (major) allele instead of the minor allele. This aberration in variant annotation is well known to the clinical genomics community and corrected when such variants are clinically reported.

The order of magnitude of the P/LP rare variants per individual is similar to the one previously predicted in another UKB study using the WES results on the first 50,000 individuals (Van Hout et al. 2020). Indeed, they estimated that in a group of 200,000 individuals, the number of observable P/LP variants genome-wide should be around five; therefore, the prediction made by Van Hout and colleagues is corroborated by my analysis. Because of the 300K AFs observed in the UKB WES cohort (Fig. 3.8) and the current AF filter used in clinical genomics projects (Turro et al. 2020b), I decided to focus the analysis on the P/LP variants with AF ≦ 0.001.

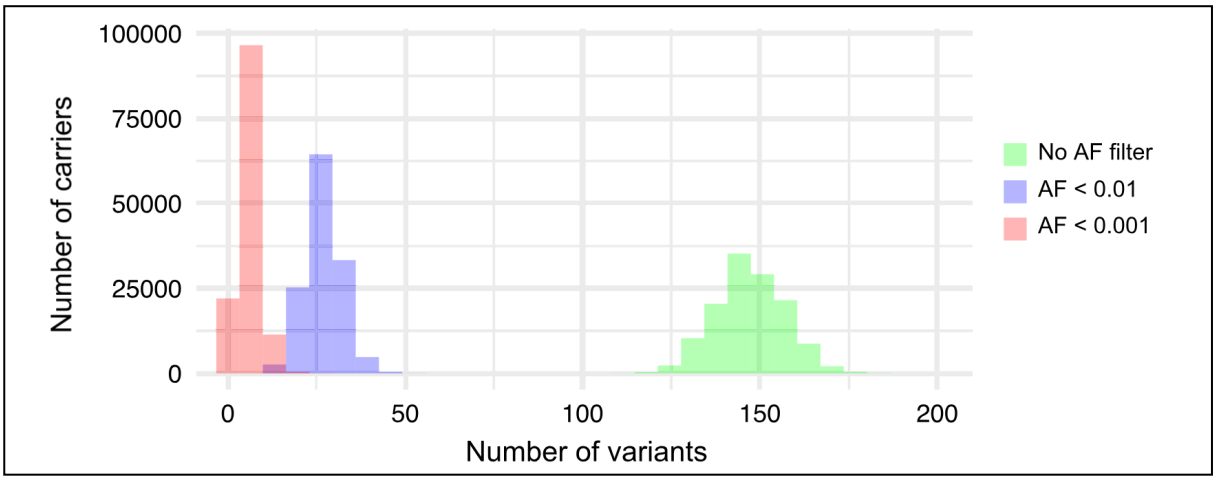


Fig. 3.8 | Number of P/LP variants co-occurring in UKB carriers grouped according to their AF. *No AF threshold applied (green), ~50,000 individuals carry P/LP variants. AF thresholds at ≦ 0.01 (blue) and ≦ 0.001 (red). Variants are considered P/LP if one of these labels is present in any of the variant resources.*

To investigate the number of P/LP variants and their distribution in the UKB WES cohort in greater detail, I focused on the 93 BTPD genes (Table 1.4). The number of P/LP variants per gene follows a negative binomial distribution, with a few genes carrying most of the variants (Fig. 3.9). Reassuringly, the phenomenon that most rare disease cases are explained by P/LP in a few genes is well documented in the clinical genomics community. Moreover, it has been recently confirmed for BTPD and other blood-related conditions (Downes et al. 2019; Thaventhiran et al. 2020; Turro et al. 2020b). It is interesting to note that the genes that carry the largest number of variants tend to have a dual mode of inheritance in which both, dominant and recessive, have been observed (Fig,3.9).

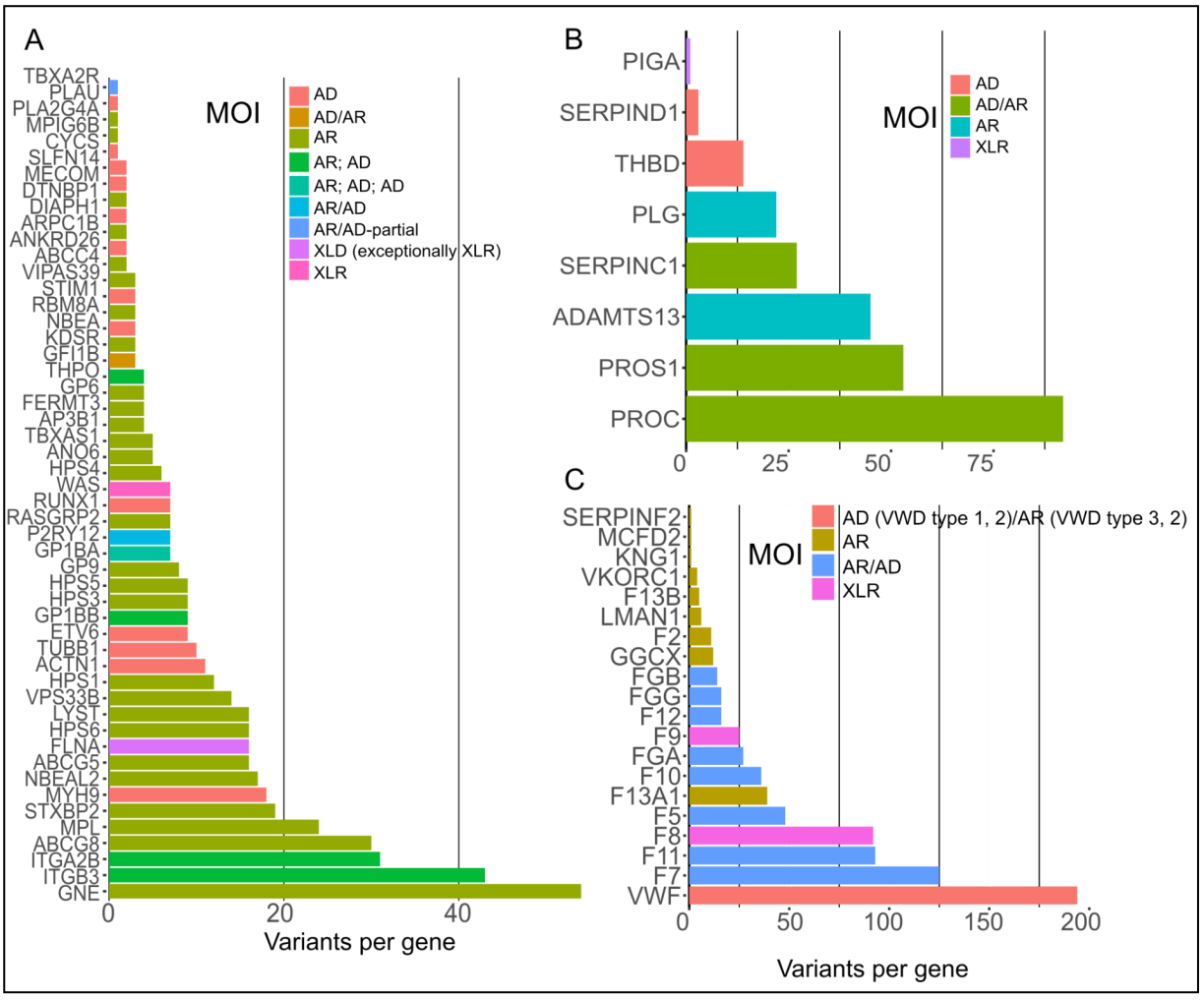


Fig. 3.9 | Number of P/LP variants per gene split by BTPD sub-domain. *A) platelet disorder genes B) thrombosis disorder genes C) bleeding disorder genes. Genes are colour coded according to their mode of inheritance. Only variants with AF < 0.001 are represented in this figure. MOI: mode of inheritance; AD: autosomal dominant; AR: autosomal recessive; XLR: x-linked recessive; XLD: x-linked dominant; VWD: von Willebrand disease.*

Ultimately, I counted the number of individuals carrying a P/LP variant in the UKB cohort at the level of the single variants (Fig. 3.10). The variant burden per gene follows the same distribution of Fig. 3.9, where a small group of variants account for the majority of variants observed in that gene. The consensus of the P/LP label across the different resources shows that the variants with the highest number of cases tend to have conflicting information about their pathogenicity levels between different databases (Fig. 3.10); for instance, the same variant is reported as pathogenic in one database and either variant of uncertain significance (VUS) or benign in another. The results depicted in Fig. 3.10 shows that for a considerable portion of the P/LP variants there were an adequate number of individuals with a particular single variant to explore effect sizes on a variant-by-variant basis.

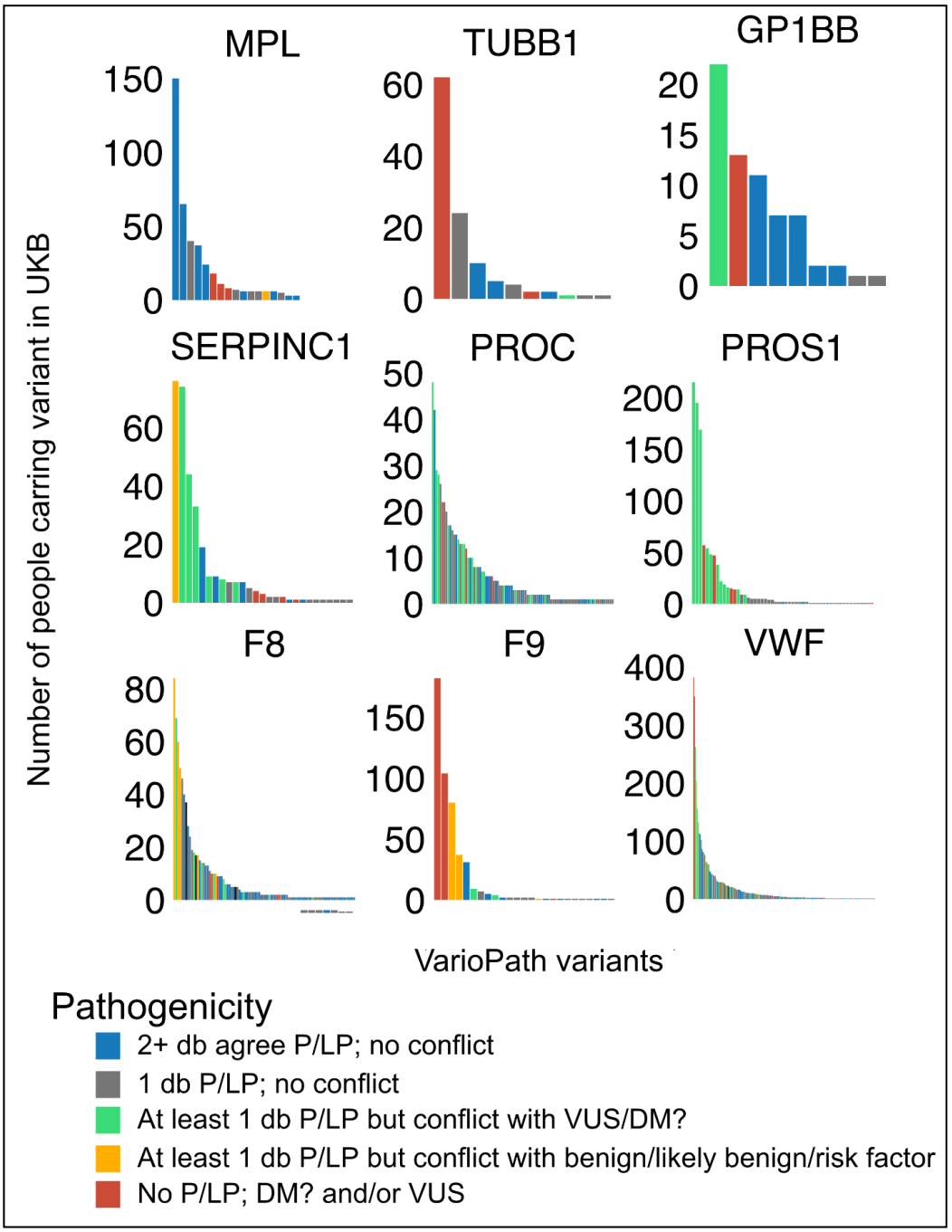


Fig. 3.10 | Number of individuals in UKB that carry pathogenic variants (AF < 0.001). *X-axis has all the pathogenic variants present for that gene in the databases interrogated. db: database (one amongst HGMD, ClinVar or Curated variants); DM: disease-modifying; DM?: questionable disease modifying; VUS: variant of uncertain significance.*

3.5 Effect sizes of rare variants in platelet disorder genes

BTPD relevant phenotypes were used to calculate the effect sizes of P/LP variants. Reassuringly, these variants, aggregated by genes in a burden association analysis or in single variant associations, confirmed the expected link between BTPD genes, the platelet phenotypes (Table 3.2) and thrombosis conditions (Fig. 3.14).

GP1BA encodes the α chain of the platelet Glycoprotein (GP) Ib/IX/V complex. The α and β chains (encoded in the *GP1BB* gene) in complex with GPIX and GPV form the receptor for the plasma protein VWF (see chapter 1.1.2.4). P/LP variants in GP1BA, GP1BB, GPIX (but not GP5) are causal of BSS (mode of inheritance recessive) when present on both alleles. However, recently, it has become apparent that rare variants in GP1BB are causal of an autosomal dominant but milder form of thrombocytopenia, which is only occasionally accompanied by bleeding (Sivapalaratnam et al. 2017). Evidence for the presence of a similar class of P/LP variants in GP1BA and the cause of mild thrombocytopenia has been reported (Downes et al. 2019). These two recent studies confirm earlier observations from single case/pedigree studies suggesting that variants in GP1BA and GP1BB might be associated with mild forms of thrombocytopenia (Andrews and Berndt 2013; Smolkin and Perrotta 2017). Common- and low-frequency variants in the GP1BA and GP1BB genes have also been associated with the platelet count and volume (Gieger et al. 2011; William J. Astle et al. 2016; Vuckovic et al. 2020). However, the difference in the effect sizes between the GWAS-identified variants and the P/LP variants has not been fully investigated yet. The effect size for all P/LP variants in GP1BA is +0.77 sd for mpv and -0.35 sd for plt (Table 3.2). In the most recent and comprehensive platelet traits GWAS (Vuckovic et al. 2020), variants in *GP1BA* have been reported to have, on average, an effect size of < 0.05 sd (highest observed 0.21 sd) and ≦ -0.07 sd for plt (highest observed -0.15 sd). The association analysis for P/LP variants at the single variant level for GP1BA produced similar results (Fig. 3.11). For example, the chr17:4933190:C:T variant (rs371226354; p.Gln196Ter), which is localised in one of the leucine-rich repeat domains of GP1BA, has an effect of -1.39 sd. This effect translates to a platelet count of, on average, 70×10^9 plt/L (p-value=8.62e-05) lower than individuals lacking this variant.

Similarly, P/LP variants in *GP9* are causal of BSS if inherited in homozygosity. The current assumption is that there are no consequences for the number and volume of

platelets in carriers of BSS-causing variants in GP9. However, (Vuckovic et al. 2020) showed, in GWAS analysis, that UKB participants who carry the GP9 P/LP variant chr3:129061921:A:G (rs5030764; p.Asn61Ser) have statistically significant differences on all four platelet traits. It is reassuring that the association analysis using the WES-genotyping results for this variant results in an effect of -0.62 sd (p-value=2.76e-16) for platelet count, similar to what Vuckovic and colleagues reported in their manuscript. Same effect and directionality is also confirmed by the analysis at the gene level (Table 3.2), with a general effect size of -0.69 sd for platelet count.

Since the initial report from a single pedigree that a variant in TUBB1 was causally associated with autosomal dominant form of macrothrombocytopaenia (Kathleen Freson et al. 2005), a large number of additional P/LP variants in *TUBB1* have been included in the P/LP databases used in this study (Fig. 3.10; Downes et al, Blood 2019; Turro et al, Nature 2020). The burden association test in the UKB-WES data confirmed that P/LP variants in *TUBB1* are associated with all four platelet traits (Table 3.2). The aggregated effect size for plt is -0.69 sd. This has large differences, but same directionality, with the more modest effect sizes (i.e. ≤ -0.1 sd) observed for the common and low frequency variants in TUBB1 associated with GWAS. The reported effect sizes of +0.66 sd for mpv; +1.29 sd for pdw; -0.33 sd for pct; -0.59 sd for plt are very similar to the effect sizes observed if the WES-determined results for this variant were associated with the four platelet traits (Fig. 3.11).

Another novel finding of my analyses is about P/LP variants in *MPL*. This gene encodes for the TPO receptor, the key growth factor for megakaryopoiesis and HSC differentiation in general (Hitchcock and Kaushansky 2014; Hitchcock et al. 2021). Homozygous and compound heterozygous LoF variants in *MPL* cause congenital amegakaryocytic thrombocytopenia (or CAMT). This extremely rare condition is diagnosed shortly after birth because of the very low platelet counts that characterise CAMT-affected children (van den Oudenrijn et al. 2000). Moreover, CAMT is a pre-leukemic condition warranting curative treatment by allogeneic stem cell transplantation (van den Oudenrijn et al. 2000; Germeshausen and Ballmaier 2021). A small number of CAMT-causing LoF variant carriers (n=27) were identified in the UKB. Under the current assumptions, these participants should have a phenotype consistent with the UKB average values. However, the burden test results show that people carrying P/LP variants in MPL have an increased platelet count if compared with the UKB participants who lack this type of variants in the gene (Fig. 3.12).

Furthermore, the observed directionality of these LoF variants is opposite to the expected, and not only results in an increased platelet count, but also a higher plateletcrit (Table 3.2).

Burden test Genes in plt phenotypes								
Gene	MOI	Trait	#Variants	mean AF	min AF	max AF	effect size (β)	p-value
mean platelt volume								
GP1BA	AR; AD	mpv	6	2.61692e-05	4.61889e-06	8.31255e-05	0.7705030	7.25069e-06
GP9	AR	mpv	10	2.85861e-04	4.61889e-06	1.70871e-03	0.1320240	1.08150e-03
TUBB1	AD	mpv	10	1.46393e-04	4.61889e-06	1.12219e-03	0.6982320	2.70443e-35
platelt distribution width								
GP9	AR	pdw	10	2.85450e-04	4.61889e-06	1.70439e-03	0.1073630	8.14667e-03
TUBB1	AD	pdw	9	1.58582e-04	4.61889e-06	1.09930e-03	1.7243100	2.99409e-199
plateltcrit								
GP9	AR	pct	10	2.86380e-04	4.61889e-06	1.71181e-03	-0.1701350	2.60840e-05
TUBB1	AD	pct	10	1.60898e-04	4.61889e-06	1.12423e-03	-0.4347950	1.87393e-14
MPL	AD	pct	27	1.31614e-03	4.61889e-06	3.31763e-02	0.0730950	5.72533e-10
platelt count								
GP1BA	AR; AD	plt	6	2.61880e-05	4.62141e-06	8.31854e-05	-0.3506780	4.15552e-02
GP9	AR	plt	10	2.86067e-04	4.62141e-06	1.70994e-03	-0.1970480	1.12185e-06
TUBB1	AD	plt	10	1.45112e-04	4.62141e-06	1.12300e-03	-0.6918350	3.01363e-34
MPL	AR	plt	27	1.31659e-03	4.62141e-06	3.31864e-02	0.0746287	2.45257e-10

Table 3.2 | Burden association test between platelet disorder relevant genes and platelet traits. mpv = mean platelet volume; pdw = platelet distribution width; pct = plateletcrit; plt = platelet count. Effect sizes (β) are reported as differences in standard deviation from the mean.

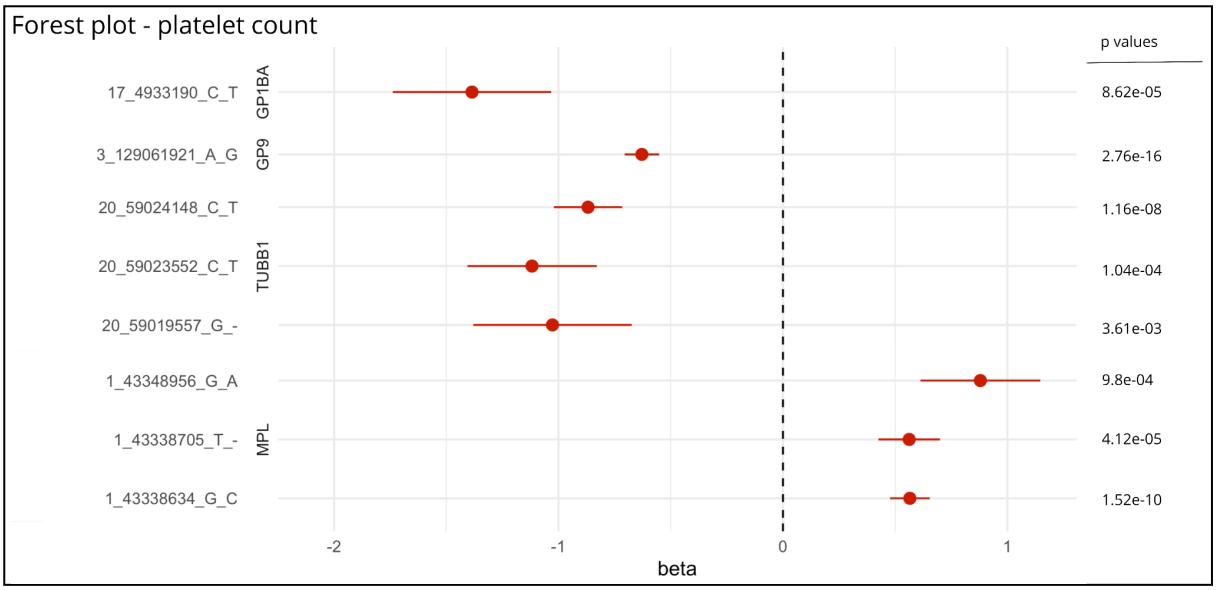


Fig. 3.11 | Forest plot for variant associations to platelet count. *Horizontal red lines indicate the 95% confidence interval.*

Single variant analysis for *MPL* tested for AD mode of inheritance, not only confirmed the effect directionality of the burden test (i.e. increased number of platelets), but also showed large effect sizes for these CAMT-causing LoF variants (Fig 3.11). Indeed, carriers of *MPL* LoF variants chr1:43338705:T:-, chr1:43348956:G:A and chr1:43338634:G:C (rs587778515, rs754859909 and rs28928907 respectively) have a platelet count between 0.5 and 1 sd higher (i.e. 25-50$\times 10^9$ plt/L) than the rest of the UKB population. The first two, with respective beta coefficients of 0.56 sd (p-value=4.1e-05) and 0.88 sd (p-value=9.8e-04), encode for premature stop codons. These premature stops reside at amino acids 126 and 474, which are both localised in the extracellular domain of the receptor (Plo et al. 2017). Fig. 3.12 shows the platelet counts for the carriers of the premature stop at residue 474 compared to the remaining UKB cohort. From this analysis, it could be concluded that, on average, the platelet count is 1 sd higher in individuals with this premature stop codon. (Fig. 3.11).

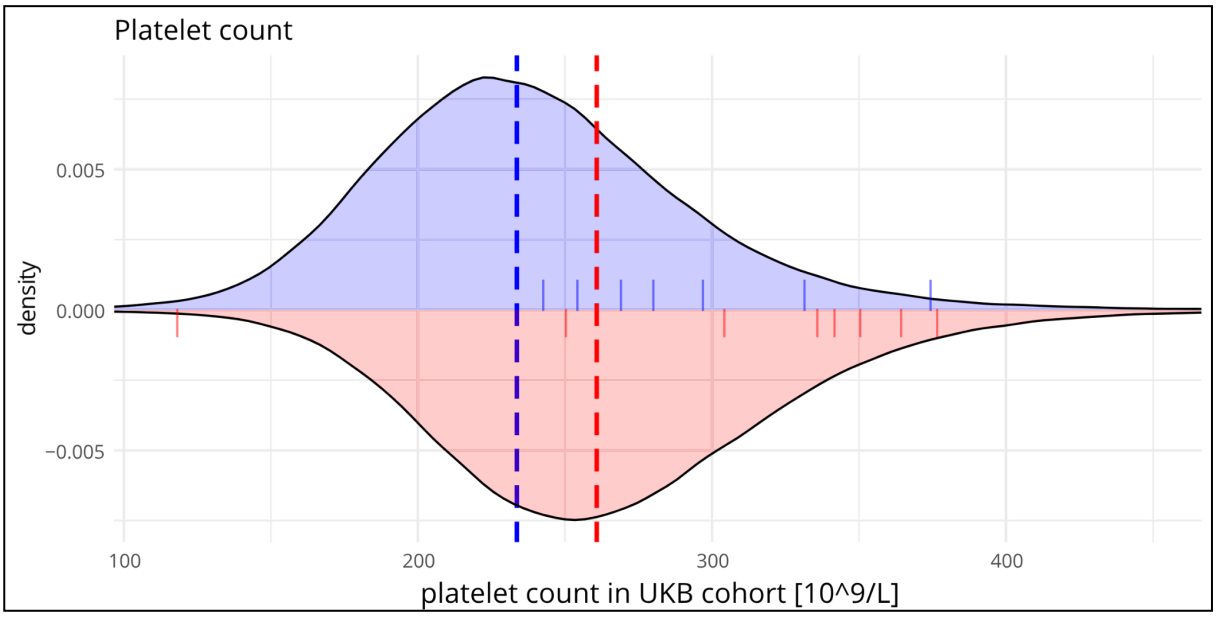


***Fig. 3.12* | Platelet count for variant chr1:43348956:G:A in UKB.** The vertical solid lines point to the individual platelet count for the carriers of the chr1:43348956:G:A variant. The density distribution shows the reference platelet count in the UKB WES population. The vertical dotted lines track the median value of the platelet count. The data are sex-stratified, with males and females, reported in blue and red, respectively.

Chr1:43338634:G:C (rs28928907) is localised in the TPO cytokine receptor domain of the MPL protein (Plo et al. 2017). This nucleotide transversion converts an arginine at residue 102 to a proline (arginine conservation score 0.94). The amino acid change removes a positive charge in the TPO binding site and inserts some physical constraints in the protein structure because of the proline side chain (Fig. 3.13; Butterworth 2005). It has previously been demonstrated that chr1:43338634:G:C causes the MPL protein to be retained in the ER and therefore leads to a reduction of the MPL protein on the membrane (van den Oudenrijn et al. 2000; Tijssen et al. 2008; N. E. Fox et al. 2009; Stockklausner et al. 2015).

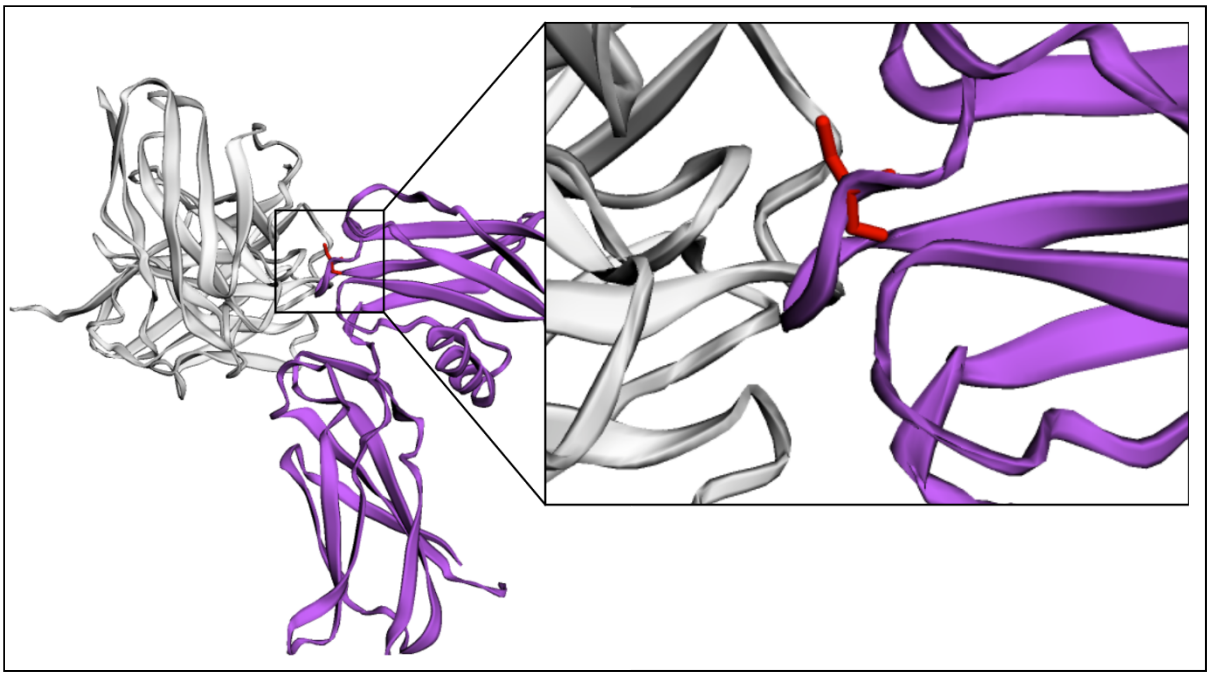


Fig. 3.13 | MPL receptor 3D structure and chr1:43338634:G:C variants (MPL:R102P). *In purple, there is the structure used to model the MPL receptor (4y5v). The red residue shows the position of the residue 102. The white molecule component is a diabody used in the crystallography process. Source EBI website: https://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/DisaStr/GetPage.pl?uniprot_acc=n/a&template=VPlist.html*

3.6 Effect sizes of VTE genes rare variants

Similarly to the platelet traits, gene aggregated statistical tests (i.e. burden aggregated test) for rare P/LP variants in genes known to be implicated in VTE confirmed their role in disease onset (Fig. 3.14). Moreover, for a small number of rare P/LP variants (n=7; 3 in PROC; 3 in PROS1; 1 in SERPINC1), which had enough carriers, it was possible to estimate the OR (Table 3.3). It is worth noting that, in these experiments, the ORs observed for PROC, PROS and SERPINC1 are lower than previous estimates. *PROC* is a glycoprotein that cleaves the thrombin-thrombomodulin complex and P/LP LoF variants in this gene are associated with AD thrombophilia (Davie, Fujikawa, and Kisiel 1991; Gandrille et al. 1993; Sakata et al. 2000; Bulato et al. 2018). The burden test analysis suggests that P/LP variants in PROC increase the risk for a VTE event by 2.76 (p-value = 9.91e-08; Fig. 3.14). Interestingly, the OR for DVT and PE are not the same (OR DVT: 4.24, p-value = 9e-12: OR PE: 2.17, p-value = 4.47e-03; Fig. 3.14), suggesting that the contribution of the gene to the two different types of VTE is uneven. The single-variants association showed strikingly high OR of 15.91 (p-value = 4.16e-04), 14.44 (p-value = 2.59e-05) and 13.03 (p-value = 7.75e-04) for the variants chr2:127423378:G:T (rs989908811, p.Gly231Arg), chr2:127421339:G:A (rs767626189, p.Ala71Thr) and chr2:127426090:T:C (rs199469470, p.Phe243Leu; Table 3.3).

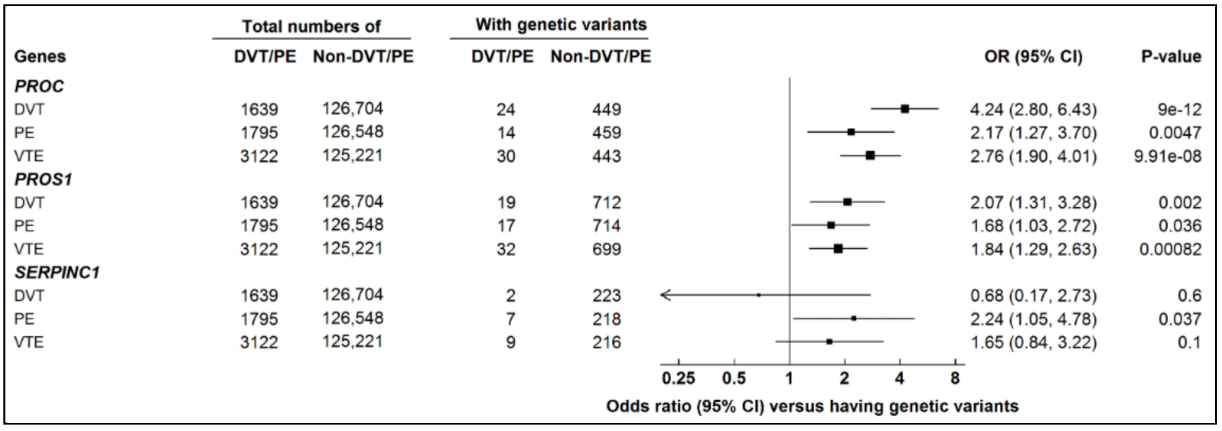


Fig. 3.14 | Forest plot reporting the OR and 95% confidence intervals (CI) for P/LP variants in genes implicated in VTE.

PROS1 has been associated with both AR and AD thrombophilia (Poort et al. 1999; Fischer et al. 2010) and its role in coagulation is to increase the inhibitory activity of *PROC* (Gierula et al. 2019). The burden test for *PROS1* shows that P/LP variants in this gene are associated with an increased number of VTE phenotypes (OR= 1.84; p-value = 8.2e-04), as well as for DVT (OR= 2.07; p-value = 2e-03) and PE (OR= 1.68; p-value = 3.6e-02; Fig.

3.14) independently. It was interesting to note that the single variant analysis, which could be applied for three PROS1 P/LP variants, confirmed the increased risk for VTE events, albeit the OR were significantly higher than the one observed for the burden test (Table 3.3).

Finally, variants in SERPINC1, which encodes for the primary thrombin inhibitor, are a well-characterised cause of thrombophilia (see chapter 1.1.2.3.4; Davie, Fujikawa, and Kisiel 1991). The primary mode of inheritance is AD but cases of juvenile VTE may also be caused by P/LP variants in SERPINC1 on both alleles. The number of VTE events in UKB participants carrying a single P or LP variant in SERPINC1 was low (VTE, n=9, DVT, n=2 and PE, n=7). Therefore, the statistical analysis was not robust and powered enough to calculate reliable OR, even for the burden association test. The OR for VTE, calculated with the burden test and P/LP variants in SERPINC1, seems to be in the range of 1.65, although the confidence interval straddles the value of 1 (which means the absence of effect). Comparably, the estimated OR for PE (i.e. 2.24) and DVT (i.e. 0.68; Fig. 3.14) are not a robust approximation of the actual effect sizes because of the small number of P/LP variant carriers. At the single variant associations, the PE event occurred in two UKB participants out of the 38 carriers of the chr1:173909819:C:G variant (rs372820797; p.Ala296Pro). Although the number of cases with this variant is only two, the estimated OR for this variant is 4.22, but a p-value of 4.83e-02 must be considered as nominally significant considering the extensive use of multiple testing in this study.

Single variant analysis							
Gene	**Trait**	**Genetic variant**	**Carriers**	**OR**	**LCI**	**UCI**	**p-value**
PROS1	DVT	3_93877159_T_C	14	17.42	3.88	78.26	1.94e-04
PROS1	DVT	3_93927336_T_C	36	4.43	1.06	18.6	4.1e-02
PROS1	VTE	3_93877159_T_C	14	9.09	2.02	40.88	4.03e-03
PROC	DVT	2_127423378_G_T	12	15.91	3.42	73.97	4.16e-04
PROC	DVT	2_127421339_G_A	20	14.44	4.16	50.1	2.59e-05
PROC	DVT	2_127426090_T_C	15	13.03	2.92	58.25	7.75e-04
SERPINC1	PE	1_173909819_C_G	38	4.22	1.01	17.64	4.83e-02

Table 3.3 | Association of genetic P/LP variants in genes known to be implicated in thrombotic events in the venous circulation. LCI: lower confidence interval; UCI: upper confidence interval.

3.7 Common and rare variants interplay

The majority of the UKB participants with a VTE event are not carriers of one of the known rare P/LP variants in the main VTE aetiological genes: PROC, PROS1 and SERPINC1. Recently Klarin et al. have reported a PRS for VTE (Klarin et al. 2019) leveraging the data of the entire UKB cohort of 0.5 million participants (genotyped with SNP array) and the MVP cohort of 0.8 M participants. The GWAS data based on array genotyping were used to calculate the PRS. The score was calculated using 297 variants in 100 genes, with the utilised variants having a AF > 0.001 (mean AF: 0.2; sd: 0.19).

In this study, I adopted the pre-calculated PRS from Klarin et al. in the WES genotyped data. As expected, this analysis showed that VTE events are enriched in UKB participants with a VTE-PRS score in the upper quantile (χ2, p-value=4.998e-04; Fig. 3.15, top-right panel). This observation is also confirmed by the visualisation in Fig. 3.16. The distribution of the PRS of the individuals who had a VTE event is higher than the UKB participants who didn't have a VTE event. The 163 UKB participants with a VTE event who are carriers of a single P/LP variant in any of the VTE relevant genes had a similar distribution across the PRS quantile, showing again an enrichment of cases in the upper PRS score quantile group (χ2, p-value=3.648e-02; Fig. 3.15, bottom-right panel). The UKB participants without a VTE event, regardless of their rare variant status, tend to be equally distributed across the three lower quantiles of the PRS score and there is a minor depletion of cases grouped in the upper-quantile of the VTE PRS score (χ2, p-value=1.499e-03; Fig 3.15). There are no statistically significant differences in the number of UKB participants who had VTE grouped according to their PRS quantiles, when comparing the rare variant carriers to non-carriers (χ2, p-value=0.327). Ultimately, it could be reasoned that the increased risk of VTE in UKB participants conferred by being in the upper-quantile PRS group is additive to the increased risk conferred by a single P/LP variant in PROC, PROS or SERPINC1. However, the number of these cases was too small to test this assumption.

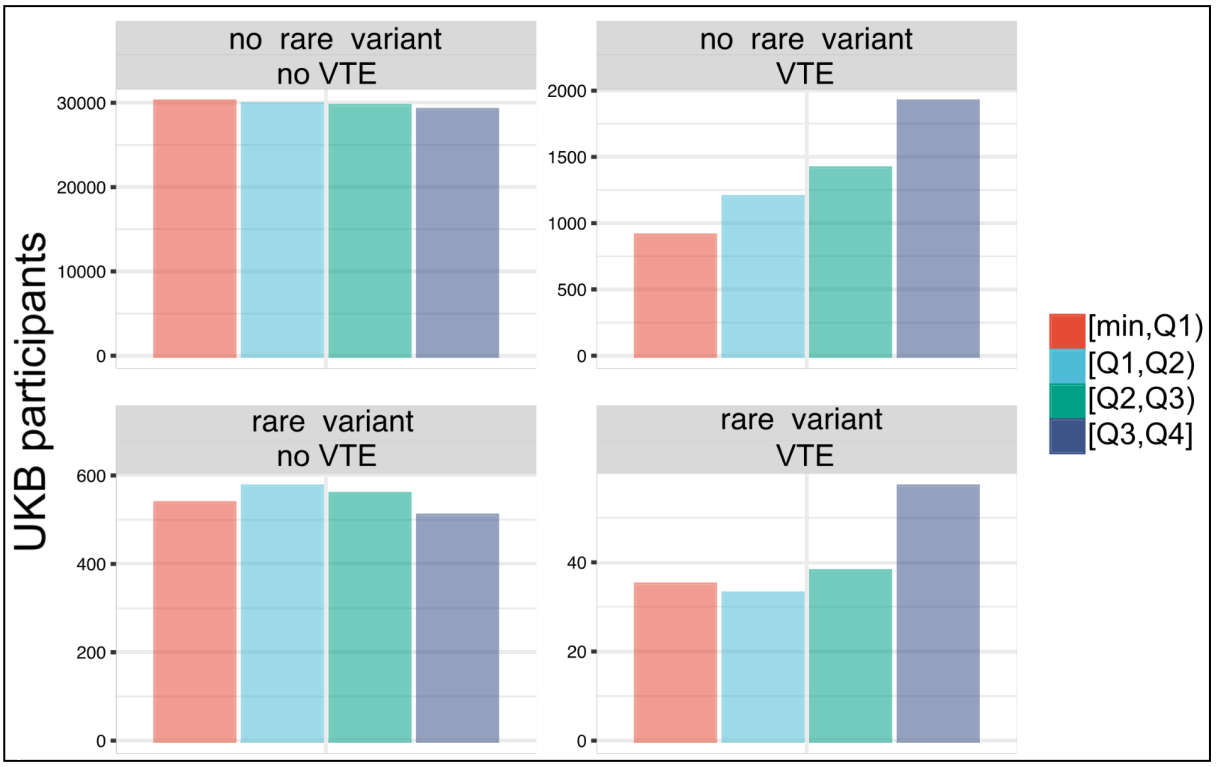


Fig. 3.15 | Number of UKB participants with or without recorded VTE events. *Participants are categorised in groups according to their genetic characteristics of rare P/LP variants in thrombosis-implicated genes and the PRS for VTE. Q is quantile (25% intervals). Q1-Q4 = First, second, third and fourth quantile, respectively*

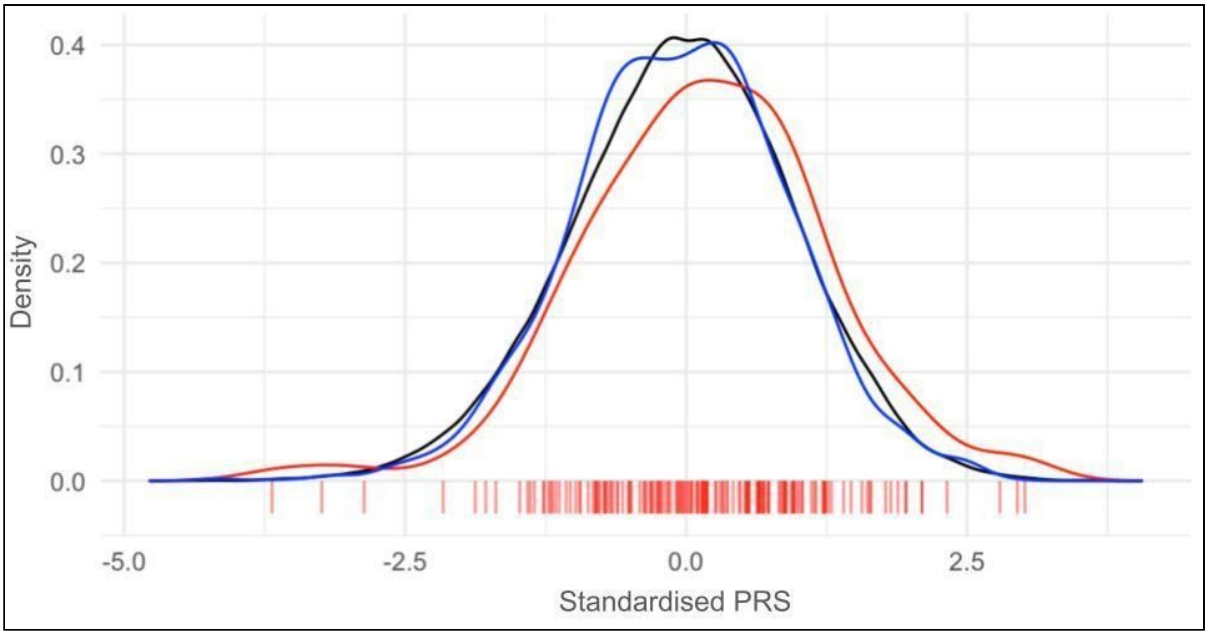


Fig. 3.16 | PRS score distribution according to P/LP variant carriers. *PRS score distribution for the 163 UKB participants who carry a single P/LP variant in PROC, PROS or SERPINC1 and had at least one VTE event (red line; the red vertical lines are the PRS scores of individuals who had VTE and carry rare variants). Participants without a VTE who carry a rare variant are depicted by the blue line. The black line represents the background PRS distribution in UKB.*

Using a generalised linear model, I was able to estimate the independent effect of rare variants and PRS on the VTE phenotypes (Table 3.4). The VTE phenotypes considered are: (i) a binary category that describes if a UKB participant had or not at least one VTE event (i.e event in Table 3.4); (ii) the number of VTE events that are recorded per UKB participant; and (iii) the age at the presentation of the first event (i.e. onset in Table 3.4). The outcome of this analysis shows that rare variants almost double the probability of having a VTE event (OR: 1.70; p-value: 6.18e-10), recapitulating the findings in Fig. 3.14. The contribution of the common variants, calculated as PRS, has instead a lower effect (OR: 1.30; p-value: 2e-16). The effect of rare variants is also evident in the number of events and age of manifestation (Table 3.4). PRS and rare variants both contribute to the number of VTE events, with OR of 1.05 (p-value: 6.00e-03) and 1.26 (p-value: 2.60e-02) respectively. Ultimately, there is a negative effect of PRS (albeit not statistically significant) and rare variants (OR: 0.63; p-value 3.50e-02) on age of the first VTE event manifestation, meaning that carriers of rare variants in VTE genes and higher PRS scores tend to have events earlier in life (Table 3.4).

Predictor term	**effect OR**	**p-value**
VTE event		
PRS	1.30	2.00e-16
Rare variant	1.70	6.18e-10
Number of events		
PRS	1.05	6.00e-03
Rare variant	1.26	2.60e-02
Onset		
PRS	0.99	7.94e-01
Rare variant	0.63	3.50e-02

Table 3.4 | OR of having a VTE event, number of events and onset based on PRS and rare diseases.

3.8 Placing the PRS into a biological context

GWAS studies have identified thousands of variants associated with quantitative traits (Claussnitzer et al. 2020), the risk of many common diseases (Lambert et al. 2021) and a few rare diseases (Lyons et al. 2012; Rhodes et al. 2019). Mapping associated variants to genes, proteins and molecular pathways is the key challenge for the post-GWAS era. A generic approach is to apply simple algorithms like VEP to gain a broad insight in the genes, transcripts and proteins implicated in a certain trait or disease. Transcripts and proteins can be organised in networks of nodes and edges, either based on analysis of transcript co-expression (Vuckovic et al. 2020) or experimentally inferred PPIs respectively. With an increase in the number of participants in GWAS the observed effect sizes of newly identified common variants becomes infinitely small. For example, the effect sizes of the first four variants found to be associated with the count and volume of platelets in GWAS with a few thousand participants ranged between 1.02 and 1.03 units (Soranzo, Rendon, et al. 2009; Meisinger et al. 2009), whilst in the most recent GWAS with 408,112 participants of European ancestry effect sizes as small as 0.0113 units were observed for these traits.

In 2017, Boyle et al proposed the omnigenic model as an alternative to the infinitesimal one, which was the widely accepted model for explaining the many common variant associations observed for common diseases. In their model most phenotypes, diseases or quantitative traits are directly explained by a limited number of genes, encoding proteins which play pivotal roles in the trait of interest. First, they named these genes "core" genes. Second, they reasoned that function-altering mutations in these genes, like the P/LP variants in genes implicated in rare diseases, would have large effects on the relevant trait. Third, other non-core genes harbouring GWAS-associated variants were grouped as "peripheral" genes. Fourth it was argued that the separation between core and peripheral genes was not a binary choice but on a more gradual scale. Finally it was postulated that the majority of peripheral genes were on average only a small number of edges and nodes removed from a core gene, providing a connectivity through which the small effects from the peripheral genes could flow to the core network of proteins (Boyle, Li, and Pritchard 2017; Fig 3.17).

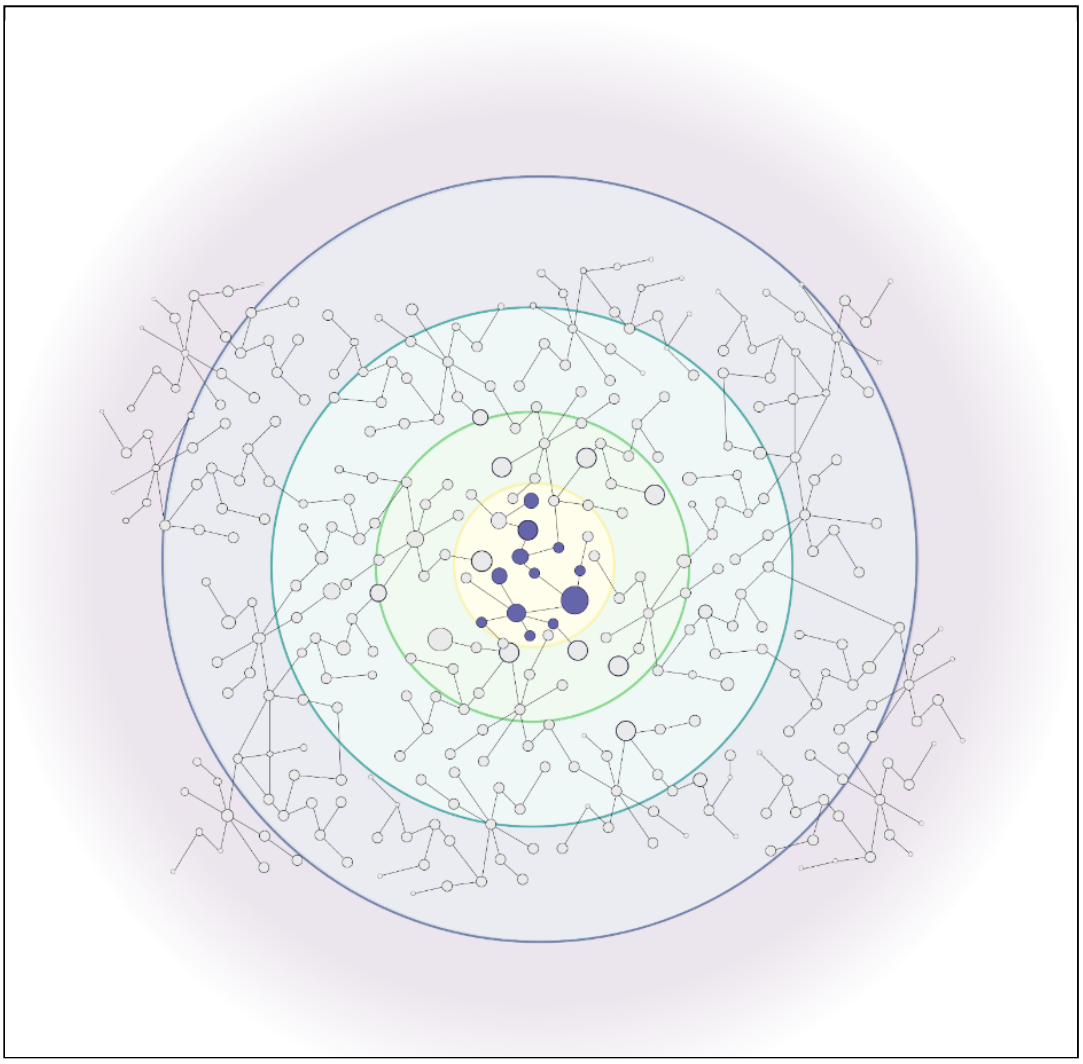


Fig. 3.17 | Visual representation of the omnigenic model. *Purple and grey nodes represent the core and peripheral genes of the network relevant for a biological trait. The surface area of the node is proportional to the effect sizes of GWAS-identified variants in these genes. The overall trend is that effect sizes are larger at the core of the network versus the periphery. The concentric disks are marking the relative distances of nodes from the core set of nodes in the network. In this chapter, the core of genes would be the 93 BTPD genes.*

As discussed, the PRS for a trait is calculated by summarising the effect sizes of genetically independent variants found to be associated with that trait by GWAS (Läll et al. 2017; Khera et al. 2018). However, so far, our understanding and biological interpretation of the PRS is limited (Hari Dass et al. 2019). To explore the omnigenic model, I used the PRS for VTE and platelet traits and interpreted these in the context of the PPI network of 18,410 nodes and 571,917 edges, which is the PPI network that most resemble the human cell "interactome" to date (see chapter 2.1; Barrio-Hernandez et al. 2021).

The likelihood of a PRS-variant for VTE being localised in one of the 93 BTPD genes is increased by 1.85 OR if compared with remaining genes of the protein network (Fig. 3.18;

Group [90,100] is the group of "core" genes; Groups [50,0) is the group containing the "periphery" genes). Therefore, these data show an enrichment of the PRS-variants in the "core" genes. However, according to the omnigenic model it could be reasoned that the PRS-variants localised in the BTPD genes should be characterised by higher effect sizes. Interestingly, this is not the case for PRS-variants in VTE. Indeed, the PRS-variants in BTPD genes (i.e. “core” genes; Fig 3.18; group [90,100]) have comparable effect sizes that span across the whole spectrum (grouped by quantile; Fig 3.18; “Q’s”). This observation, which seems in contrast to the omnigenic model, may be an artefact of the VTE PRS calculation (Klarin et al. 2019). The VTE PRS is based on a relatively small number of variants (n=297) that map to less than 100 genes (Fig; 3.18; left). This limited genetic resolution may be due to the number of cases (~14,000 in UKB) which were included in the GWAS (Klarin et al. 2017). As a consequence of this under-powered GWAS the genetic architecture for VTE is only partially resolved and a large number of associated variants remain to be identified. For these reasons, the PRS for VTE might not be the ideal archetype to explore the omnigenic model.

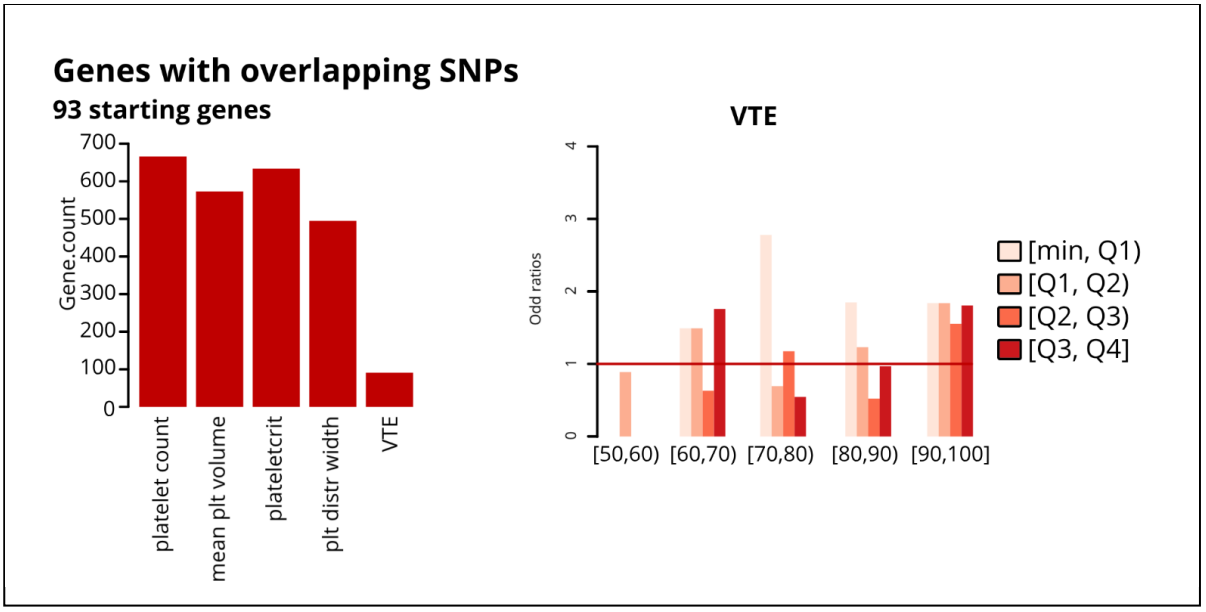


Fig. 3.18 | Number of genes that overlap variants used in PRS calculation*. Left) the number of variants used in the PRS calculation that map to a gene (or the 10 Kb surrounding the gene). Right) Distribution of the effect sizes of variants in the surrounding of the BTPD core gene pathway (group [90,100]) and the effect sizes grouped by quantiles. Q is quantile (25% intervals).*

In contrast to VTE, the GWAS for platelet traits is well powered and offers almost 10 times the variants (Fig 3.18; Klarin et al. 2019; Vuckovic et al. 2020). In fact, the most recent GWAS by Vuckovic and collaborators has identified more than 10,000 variants associated with blood cell traits (Vuckovic et al. 2020). Based on these results, PRS variants for the four

platelet traits (i.e. mpv, plt, pdw and pct) can be used to test their distribution in the interactome. The PRS variants used to predict the mpv, plt, pdw and pct traits map to 654, 739, 555, and 700 genes, respectively (Fig. 3.18).

The PRS for platelet count, indeed, shows an enrichment (Wilcoxon test; p-value=10e-05) for the largest effect sizes (i.e. quantile [Q3, max) in Fig 3.19) in the BTPD gene network when compared to the whole human interactome (Fig. 3.19).

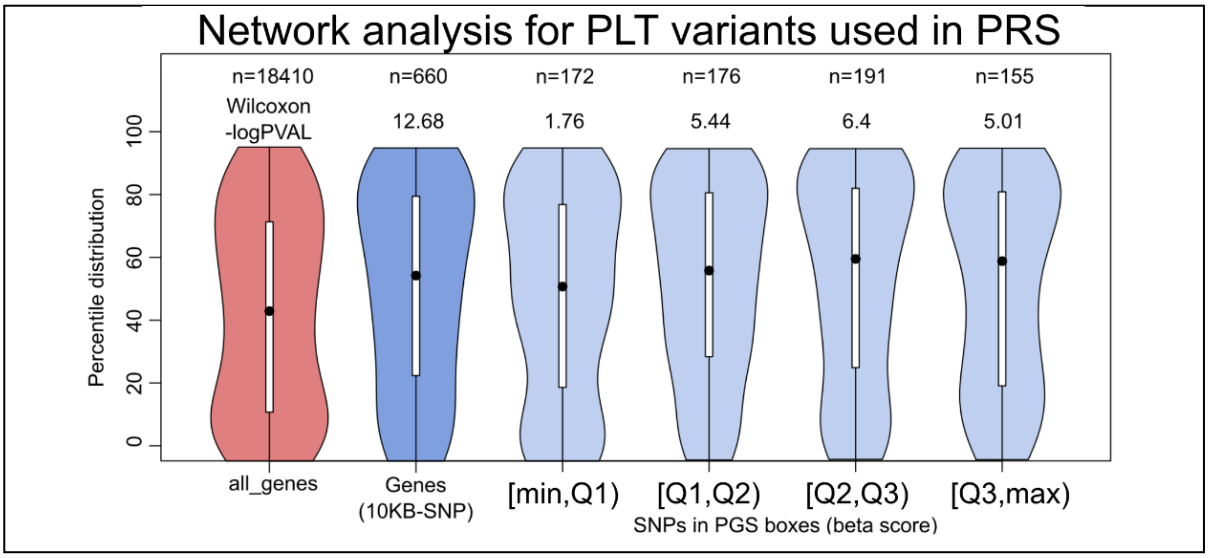


Fig. 3.19 | Distribution of the effect sizes in the core BTPD interactions. *The red violin plot shows the distribution of the plt PRS with respect to the entire protein network. Blue violin with the dark shade shows the distribution of PRS variants with respect to the PPI network of the 93 BTPD genes. Violin plots with the light blue shade show the distribution of the effect sizes for plt PRS variants categorised according to their effect sizes. Q is quantile (25% intervals).*

Similar levels of enrichment for the largest effect sizes of the PRS variants are identifiable also in all the other platelet traits (Fig. 3.20). As hypothesised at the beginning of this analysis, the PRS variants with the largest effect sizes (group "[Q3, max)" are enriched in the centre of the PPI network, which in these experiments correspond to the 93 BTPD genes (Fig. 3.20).

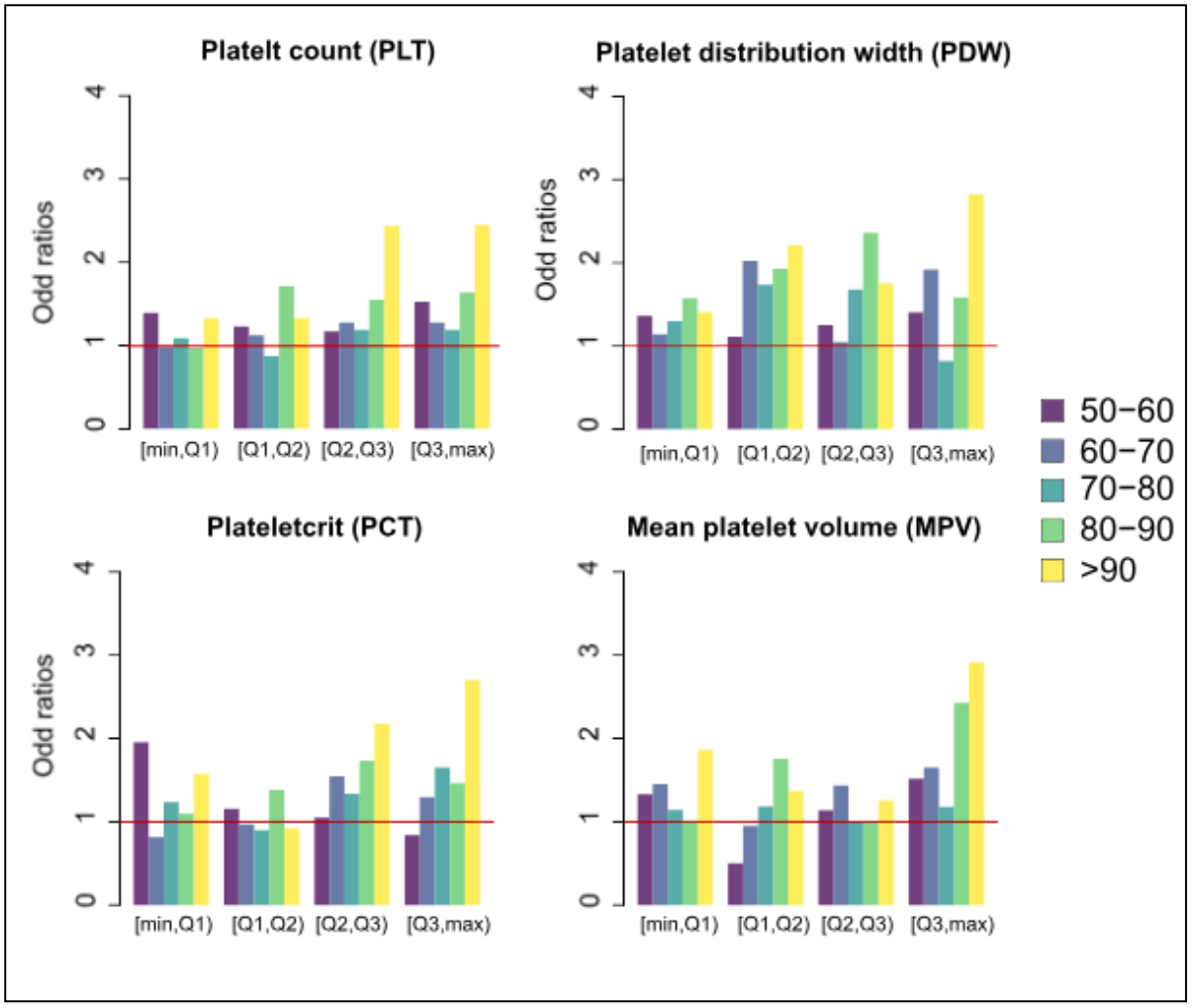


Fig. 3.20 | Distribution of the PRS variant effect sizes in the proximity of the BTPD genes PPI network for the four platelet traits*. Colours are referring to the percentile distribution from the pathway (i.e. distance from the pathway) and colour coded as in Fig. 3.17. The yellow bar is the closest to the BTPD core network. Q is quantile (25%) intervals.*

3.9 Comparison of the results obtained by statistical approaches and by protein structure-function analysis.

Inferring the functional consequences of missense variants by exploring information gleaned from protein structure has been one of the foundations of informing decisions about their level of pathogenicity. To compare the predicted pathogenicity of the two independent approaches, namely protein-structure and statistics based, I scored the functional consequences of P/LP missense variants modelling their effect on the protein structure consequences. Then, I classified the variants in two categories (i.e. “Disease” if pathogenic and “Natural” if benign) using a support vector machine (SVM) model. This method allowed me to quantify the deleteriousness of a single amino acid polymorphism in a protein domain for which structural information is available (Fig. 3.21). The SVM outcome is a deleteriousness score that can be compared to the OR estimated with the statistical approach. This model can discriminate, even if not perfectly, between pathogenic variants, functional-neutral and benign variants (see chapter 2.1 for more details on the method). A total of 23,321 P/LP variants and 2,296,890 benign variants have been used to train the SVM model, which is then be used to score the P/LP missense variants that have been identified in UKB. The SVM score ranges from -10 to +10, to reflect the spectrum 'deleterious-neutral-benign' respectively (Fig 3.21).

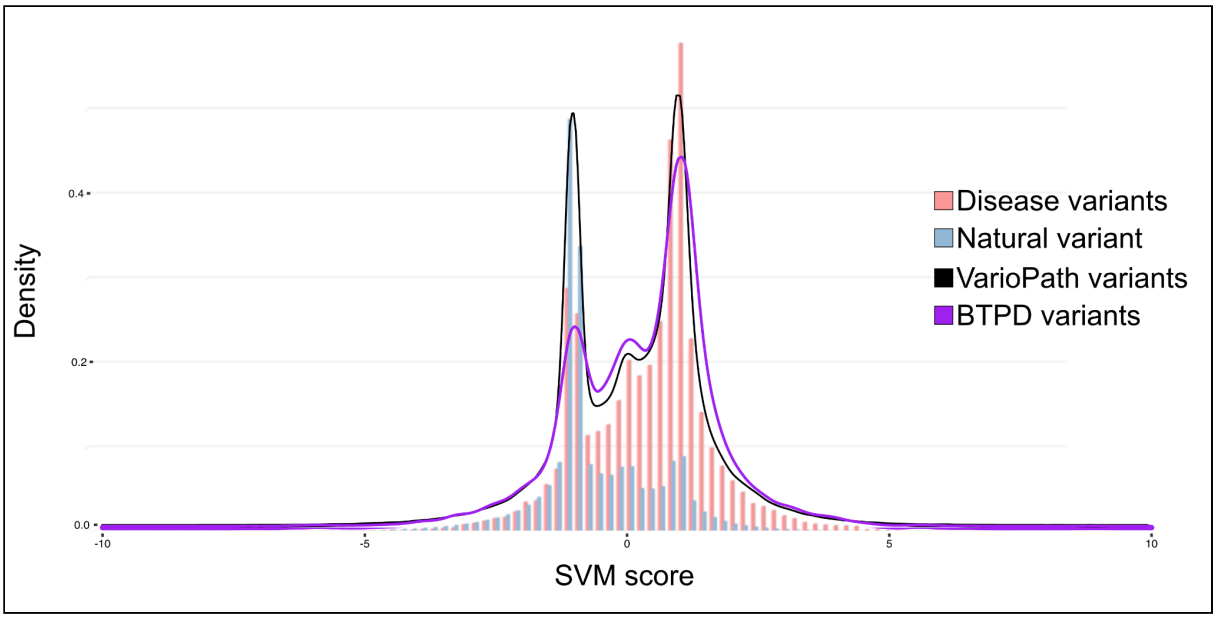


Fig. 3.21 | Support vector machine classification of the P/LP variants according to their deleteriousness effect on the protein structure and function. *Disease variants: the variants that have been reported in ClinVar as P/LP and have been used as disease reference to train the SVM model. Natural variants: the missense variants, identified in gnomAD, that have been used as benign reference in the training of the SVM model. VarioPath variants: the P/LP missense variants with carriers identified in the UKB cohort. BTPD variants: the P/LP missense variants that map to any of the BTPD genes and with carriers identified in the UKB cohort.*

In the following step, I determined the correlation between the variant effect sizes obtained from the single-variant association test (Fig. 3.11 and Table 3.2) with the score assigned by the SVM analysis. This comparison shows a small but significant correlation between the two approaches (r^2=0.095; p-value=1.20e-03; Fig. 3.22), which seems not to be influenced by gene-pLI score or mode of inheritance (Fig 3.22). This lack of a strong correlation suggests that the prediction of the pathogenicity of variants could benefit by integrating the results from both methods.

There are several possible explanations for these findings. First, a considerable portion of missense variants may impact the protein trafficking or the post-translational modifications, leading to absence or not-functional proteins. The SVM analysis method, and crystal structure approaches in general, disregard these two molecular mechanisms, therefore assigning a nominal and erroneous deleteriousness score. Second, it is well established that P/LP variants in ClinVar may have an incorrect pathogenicity label, hence the SVM classification model may be biased from these misclassifications. Ultimately, the parameters used to train the SVM model that relies on crystal information make extensive use of protein homology between different domains. There are ample examples where the use of homology modelling leads to incorrect assumptions of the functional consequences of

an amino acid polymorphism (L. R. Forrest, Tang, and Honig 2006; Haddad, Adam, and Heger 2020).

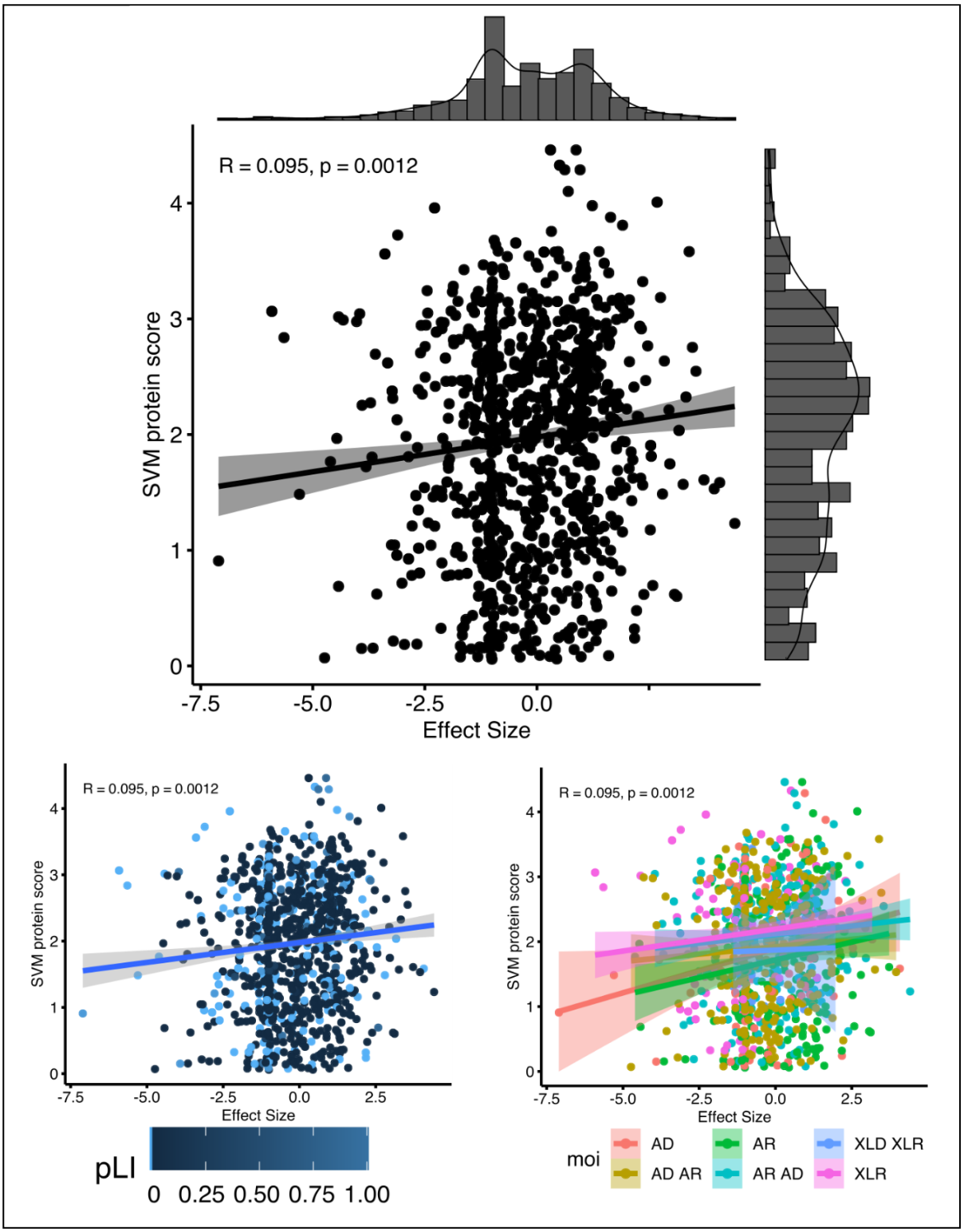


Fig. 3.22 | Correlation between effect size and the deleteriousness on the protein structure. *R: Pearson correlation coefficient; p: p-value; pLI: probability of LoF intolerant; moi: mode of inheritance; AD: autosomal dominant; AR: autosomal recessive; XLD: X-linked dominant; XLR: X-linked recessive.*

3.10 Discussion

The knowledge that healthy individuals carry pathogenic variants has been reported from the early days of WES and WGS analysis, which commenced a decade ago (MacArthur and Tyler-Smith 2010; Van Hout et al. 2020). However, to what extent carrying such variants has consequences for wellbeing and health has not been thoroughly investigated until the first release of UKB WES data became available (Goodrich et al. 2021; I. S. Forrest et al. 2021). To address this important question in a robust manner the pathogenic variants must be measured by direct genotyping, because inferring the frequency of this class of rare variants (AF < 0.001) by imputation of array genotyping data has been proven not reliable (Bycroft et al. 2018).

The intersection of the P/LP list of variants with UKB WES data confirms the observation that a portion of the P/LP variants have far too high AF to be considered as genuine pathogenic variants (Fig. 3.8; Shah et al. 2018; Xiang et al. 2020). There are several plausible explanations for the presence of P/LP variants with incorrect pathogenicity labels in the ClinVar and HGMD databases. Entry variants in these databases reflect several decades of clinical genetics research and practice that has its origins previous to the onset of large genomics cohorts. In the seventies and early eighties, it was acceptable scientific practice to report a coding variant in a disease-implicated gene as pathogenic, if such allele differed from the published open-reading frame sequence. Especially if the variant was co-segregating with a constrained number of pedigrees. Furthermore, it could be argued that too much weight may have been given to the *in vitro* functional experiments (e.g. over-expression of the mutant open reading frames or inference of functional consequences from protein structure data). These practices were buffered with the establishment of the ExAc project (now subsumed in the gnomAD) which provided a reference for the MAF of variants in the healthy population.

The analysis performed with the WES genotyping results from the UKB participants, at the gene level (i.e burden test) and at the single variant level has revealed some interesting findings. Using the UKB WES data, I was able to calculate the effect sizes of P/LP variants, which to some extent recapitulate some previous GWAS findings of platelet-trait variants (e.g. chr3:129061921:A:G; chr20:59023753:G:A; Fig. 3.11; Fig. 3.14) and VTE genes (*PROC* and *PROS1*; William J. Astle et al. 2016; Klarin et al. 2017, 2019; Vuckovic et al. 2020). For example, BSS causing P/LP variants in GP1BA and GP9, in

heterozygosity, exert a significant effect on platelet count and the other platelet parameters, confirming earlier observations (Sivapalaratnam et al, Downes et al and Turro et al). In this chapter, I was able to estimate the effect size of these P/LP variants. Moreover, this observation refutes the assumption that there are no phenotypic consequences of being carrier of a P/LP variant in the genes implicated with BSS, an autosomal recessive disease. In fact, the burden test for G1BA, GP9 and TUBB1 shows that P/LP variant carriers may have a reduced count of about -0.7 sd (-70×10^9 plt/L). This reduction is a clinically relevant observation that can have direct application. For instance, in pregnancy or during myeloablative treatment, it might push the platelet count below the therapeutically actionable threshold. Furthermore, the cumulative effect of common variants (tested as PRS) could aggravate the phenotype that is caused by P/LP variants (Collins et al, BJH 2021).

Interestingly, some AR genes may have a different effect if carrying P/LP variants in heterozygosity, such as in the case of *MPL* (Fig. 3.11). According to the literature, It could have been reasoned that carriers of CAMT-causing LoF variants in *MPL* would have a reduced platelet count, similarly to what was seen for carriers of BSS-causing variants in GP1BA, GP1BB and GP9 (Plo et al. 2017; van den Oudenrijn et al. 2000; Tijssen et al. 2008; N. E. Fox et al. 2009). However, the results presented in this chapter (Fig. 3.11; Fig. 3.12) introduced a completely new scenario, which associates the heterozygous state of *MPL* LoF variants to an increased platelet count. This peculiarity has been already observed once, for the variant chr1:43338634:G:C, in a single pedigree (Bellanné-Chantelot et al. 2017) and my results provide a much-needed confirmation of these previous findings. Indeed, I expand this association from one isolated case to multiple, and not limited to a single variant, but to all the LoF variants in *MPL*. The possible mechanisms of this counterintuitive effect have been thought to be related to TPO residence time in the bloodstream (Bellanné-Chantelot et al. 2017). *MPL* protein product is responsible for the clearance of TPO from the bloodstream. A reduced amount of *MPL* on the membrane has a lower capacity to remove TPO, increasing the circulating concentration, which, as a reflection, increases the production of MKs and platelets. Whether this is the underlying explanation for the increased platelet count remains to be investigated. Another relevant question that has to be answered is whether this possible compensatory mechanism causes an increased risk for haematological malignancies, as these are observed in the CAMT cases themselves (Germeshausen and Ballmaier 2021).

VTE analysis confirmed the pathogenic role of known genes and rare variants (Fig. 3.14; Table 3.3), with OR of up to 17.42 for chr3:93877159:T:C. Indeed, both burden association test and single variant analyses identified OR similar to the ones previously

reported. For instance, F5 Leiden variant (chr1:169549811:C:A; rs6025; p.R506Q) and F2 variant rs1799963 (chr11:46739505:G:A; 3'UTR variant) have previously reported ORs for VTE that are comparable (or lower) to the VTE ORs for the other genes considered in this chapter, 2.97 and 2.61 respectively (Fig. 3.14; Table 3.3; Germain et al. 2015; Klarin et al. 2019; Lindström et al. 2019; Herrera-Rivero et al. 2021). However, these results have to be contextualised with the possibility that VTE phenotypes in UKB may be underestimated or epidemiologically underpowered. In fact, the number of total VTE cases in UKB is 14,222 (Klarin et al. 2019). One interesting consideration that emerged from the VTE analysis is in the role of PRS in trait manifestation (Fig. 3.15; Table 3.4), how important it is to consider all the variants (rare and common) in trait assessment (Table 3.4).

PRS aim to improve the diagnosis and prognosis of complex diseases (Mega et al. 2015; Khera et al. 2018; Green et al. 2020), however part of the scientific community is still reluctant on their application in the clinical practice because of limited accuracy (De La Vega and Bustamante 2018; Torkamani, Wineinger, and Topol 2018). Understanding the biology of diseases can be used to generate better PRS. Indeed, I combined PRS and protein interaction networks to show (Fig. 3.18; Fig. 3.20) that PRS are implicitly selecting for variants occurring in the networks that are crucial for the phenotypic manifestation, in line with the omnigenic model (Boyle, Li, and Pritchard 2017). However, some of the variants map to genes that seem not directly linked to any disease-relevant gene. This may be because the literature on diagnostic-grade genes and protein pathways is still not complete or because SNPs are associated with the incorrect genes. Non-coding and regulatory SNPs can exert their role on distant genes (see chapter 4.10), therefore a better mapping for these SNPs can improve the biological interpretation of the GWAS and PRS variants (Cano-Gamez and Trynka 2020). These PRS-protein-network links could be used to generate PRS that have less noise and possibly more accurate trait prediction.

Ultimately, further work is needed to understand the biochemistry of the association of the pathogenic variants. Indeed, the comparison between effect sizes and amino acid substitution showed a limited correlation (Fig. 3.21). There are a series of factors that can contribute to these differences. For instance, post-translational modifications cannot be easily predicted by the amino acid changes, therefore the lack of this information can contribute to the limited correlation. Moreover, the variant effect modelled on the deleterious effect on the protein is limited to protein-coding variants, while effect sizes of WGS and large cohorts can be utilised also to address the role of non-coding ones (discussed further in chapter 5). However, the fact that these two data do not correlate can also mean that they

describe different pieces of information that do not overlap, therefore, a third score that takes into consideration both of them may be a better predictor for the variant effect.

To conclude, in this chapter I confidently estimated the effect sizes of rare variants and investigated the consequences of being a carrier of one of these nucleotide changes in a healthy cohort. This analysis associated the status of the carrier of AR variants with discrete phenotypes, even opposite to the one expected. Moreover, this chapter emphasised the importance of integrating rare and common variants in the understanding of the phenotypes, adding more evidence to a growing literature (Kuchenbaecker et al. 2017; Fahed et al. 2020; I. S. Forrest et al. 2021; Goodrich et al. 2021). Possibly, in the future, genomic medicine diagnostics will use effect-sizes, or a derived score, as the central definition of diseases. It will not matter whether the contribution to the phenotype is coming from one rare variant with a large effect size or multiple commons with a smaller one, reducing the gap between rare Mendelian disorders and common complex ones.

Chapter 4

Identification of the regulatory regions relevant for BTPD genes

4.1 Introduction and aims of the chapter

GWAS are a powerful tool to determine the associations between common genetic variants and phenotypes/traits. They have been widely adopted in the haematological domain to associate thousands of genetic variants with phenotypes relevant for BTPD. This approach allowed the identification of biological pathways implicated in disease aetiology or in the regulation of quantitative traits and highlighted the major role of non-coding genetic variants in controlling quantitative traits and the risk of disease (Maurano et al. 2012; Astle et al. 2016b; Sabater-Lleal et al. 2019; Vuckovic et al. 2020; Nasser et al. 2021). Unfortunately, GWAS do not provide any mechanistic insights on the mode of actions of the variants. For coding non-synonymous (ns) variants, one can speculate that associated variants affect protein function, however, the approach of linking coding variants to protein function is not always accurate (Smemo et al. 2014). For non-coding variants, representing approximately 98% of GWAS variants, it is challenging to assign the mechanism because there is no obvious association to any cognate gene (Astle et al. 2016b). Furthermore, simple and widely used algorithmic approaches like variant effect predictor suggest erroneous candidate genes because the long-range interactions between regulatory elements and promoters is not being considered (Petersen et al. 2017). Therefore, alternative methods are needed to infer the most likely mode of action, for instance, to identify enhancer-promoter loops (Schoenfelder and Fraser 2019). In fact, given the variant enrichment in regulatory regions (Astle et al. 2016b), it is fair to assume that these SNPs alter the binding site of transcription factors and other DNA binding proteins, therefore influencing phenotypes via the regulation of the expression of the cognate genes. The systematic study of regulatory region targets is complex and a comprehensive interpretation has not been achieved yet. The complexity is due, at least, to two main reasons: (i) regulatory regions do not have a clear position in respect to the cognate gene and (ii) they are cell type-specific (Ong and Corces 2011; ENCODE Project Consortium 2012; Alberts 2014; Javierre et al. 2016; Lichou and Trynka 2020).

In most cases, regulatory regions control the expression of the nearest genes, however, several exceptions have been reported. Indeed, enhancers can regulate distal genes, even a megabase apart, and be inert towards the closest ones (see also chapter 1.3.5; Smemo et al. 2014; McGovern et al. 2016; Schoenfelder and Fraser 2019; Bick et al. 2020; Nasser et al. 2021; Bevacqua et al. 2021). Moreover, one enhancer can control multiple genes, therefore the regulatory effect on the closest gene does not exclude that the

same enhancer can also regulate other distal genes (Peter Hugo Lodewijk Krijger and de Laat 2016; Viñuela et al. 2021). Also, different cell types use different regulatory regions (Hnisz et al. 2013; Peter Hugo Lodewijk Krijger et al. 2016; Javierre et al. 2016). Furthermore, the cell-type specificity of regulatory regions can also be extended to cell state and age, therefore even the same biological cell type exposed to different stimuli or at different phases of its life can use a different set of regulatory regions (Petersen et al. 2017; Soskic et al. 2019; Guan et al. 2020).

For these reasons, the identification of loops between regulatory elements and cognate genes, in a cell type-specific manner, can only be obtained by extensive experimental work and functional validation. Every approach has its strengths and weaknesses and can successfully identify some regulatory regions while missing others (Kwasnieski et al. 2014; Bevacqua et al. 2021), possibly because several mechanisms of regulation co-exist at the same time (e.g.; promoter-enhancer loops or lncRNA; see chapter 1.3.5). In addition, not all the *bona fide* regulatory elements identified have a regulatory effect, meaning that there is still much to understand on the epigenetics modifications and motifs that make a DNA sequence a regulatory element (Kwasnieski et al. 2014). Indeed, the number of GWAS variants that are functionally validated is magnitudes lower than those identified (i.e. 1:46 validated SNPs to GWAS identified regions; Gallagher and Chen-Plotkin 2018).

Several *in silico* and high-throughput methods have been developed to define the relevant cell types and to assign non-coding variants to the cognate genes. *In silico* methods use genomic features (e.g. enhancers or insulators) and GWAS variants to identify the cell type, or types, responsible for the phenotype of interest (Xuanyao Liu et al. 2017; Soskic et al. 2019; Cano-Gamez and Trynka 2020). Other methods have been developed to predict interactions and assign regulatory regions to genes. For instance, DeepC (Schwessinger et al. 2020) and Akita (Fudenberg, Kelley, and Pollard 2020) are able to accurately predict the chromatin structure across several cell types and also foresee the effect of structural variants on the 3D structure of the chromatin.

Similarly, experimental approaches can be used to determine the chromatin structure in different cell types, and several studies used genome-wide chromosome conformation capture techniques (i.e. Hi-C; see chapter 1.3.5 and 1.4.3; Rao et al. 2014; Yardımcı et al. 2019). A more tailored approach focuses only on promoter interactions to maximize the number of regulatory loops identified in each experiment (i.e. pcHi-C; Mifsud et al. 2015; Javierre et al. 2016; Schoenfelder et al. 2018). However, both approaches have not been

applied to all BTPD-relevant cell types or have not been able to retrieve all the regulatory loops relevant to the BTPD genes. pcHi-C is a valid method to obtain information on the regulatory loops of the promoters, however, given the high complexity that still characterises these libraries, sequencing results may have reduced representation when it comes to gene-specific interactions (Fig. 4.4; see also chapter 1.3.5 and chapter 2.3; Mifsud et al. 2015; Javierre et al. 2016; Schoenfelder et al. 2018).

In this chapter, I describe my efforts to define more comprehensively the regulatory regions of BTPD genes in MK, EC and HEP (Fig. 4.1). This information can be used to improve our understanding of the role of common and rare variants in non-coding elements of the 93 BTPD genes in biology and disease aetiology. I first differentiated human induced pluripotent stem cells (hiPSCs) towards the principal cell types responsible for the coagulation process using previously published protocols (Cheung et al. 2012; Hannan et al. 2013; Moreau et al. 2016). I focused on these cell types because of the role they play in haemostasis, either because directly involved in it, i.e. platelets and endothelial cells, or because they produce the haemostasis proteins circulating in the bloodstream, i.e. hepatocytes (see chapter 1.1.2). Then, I generated high coverage cell type-specific chromatin conformation capture maps for the 131 diagnostic-grade genes assessed in the ThromboGenomics (TG) diagnostic panel (Simeoni et al. 2016, Downes et al. 2019, Megy et al. 2019). The chromatin conformation maps allowed the identification of thousands of *bona fide* regulatory loops relevant for the BTPD genes in MK, EC and HEP. These experiments confirmed previous descriptions of the length of the regulatory loops and highlighted the differential use of regulatory regions in different cell types (McCord, Kaplan, and Giorgetti 2020; Javierre et al. 2016; Petersen et al. 2017). I annotated the functions of the captured sequences with cell type-specific genomic features, such as enhancers or transcription factor binding sites, either obtained from the BLUEPRINT consortium (Stunnenberg et al. 2016, and Hirst 2016a) or from collaborators (see chapter 4.2).

Finally, I investigated the biological insights that can be derived from the regulatory loops. I examined the use of the regulatory space in the different cell types and correlated the type, number and length of interactions to transcription of BTPD genes. In the last part of this chapter, I explored the advantages that the integration of my data can bring to other high-throughput studies, such as GWAS and PRS. I was able to reassign 9.5% of the SNPs from a previous GWAS (Astle et al. 2016b) to what may be a more appropriate set of cognate genes.

To conclude, this approach generated, to the best of my knowledge, the most detailed cell type-specific interaction maps connecting the non-coding regions, with potential regulatory roles, to a set of genes with critical roles in haemostasis. Because of my interest and the expertise of the laboratory, most of the biological investigation in this and in the following chapter will be relevant for megakaryocyte biology, however, this strategy can be translated to other biological systems. Ultimately, these findings can be implemented in the analysis of whole-genome sequencing-based data from patients with unexplained inherited BTPDs, with the aim to improve the diagnostic yield and potentially accelerate access to effective treatments (Turro et al. 2020b).

Part of the work presented in this chapter has been a collaborative effort with colleagues in our group and colleagues from Prof. Ludovic Vallier laboratory (University of Cambridge). In particular, the EC cell generation has been performed by Dr. Matt Sims; HEP Hi-C and H3K27Ac ChIP-Seq libraries have been generated by Dr. Rute Tomaz and Dr. Jose Garcia-Bernardo both from Prof. Ludovic Vallier group. Also, the information on transcription levels and H3K27Ac for MKs derive from the BLUEPRINT consortium experiments (L. Chen et al. 2014; Grassi et al. 2020).

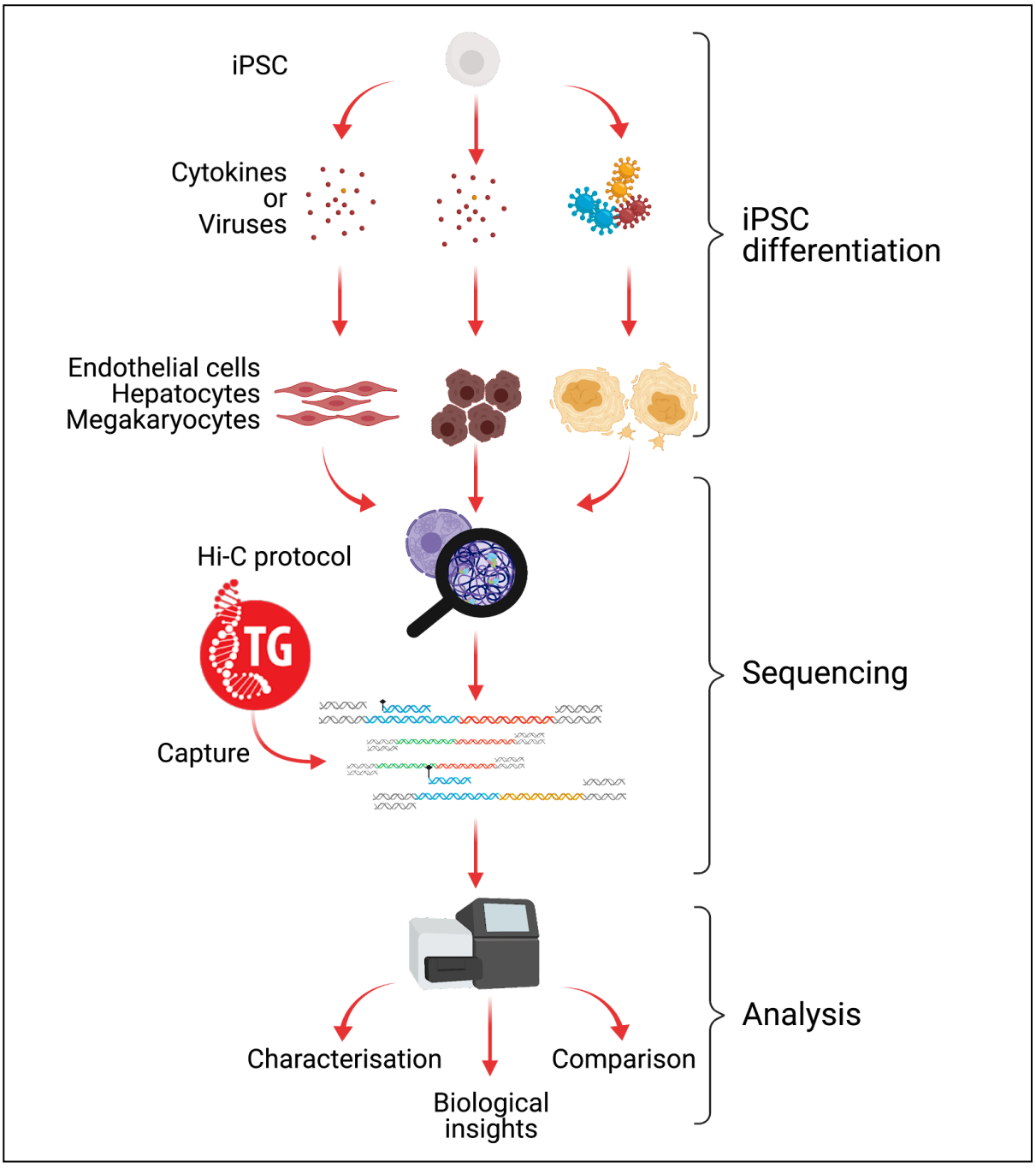


Fig. 4.1 | Schematic representation of the workflow adopted in the TG Hi-C experiments and analyses described in this chapter. *For simplicity, the workflow can be divided into three consecutive steps. Firstly, the iPSC are differentiated in the three cell types of interest. Secondly, the nuclei of these cells are used to produce TG Hi-C libraries. Lastly, the sequencing results are analysed.*

4.2 hiPSC as a biological source of BTPD relevant tissues

Since 2015, Hi-C protocols have been able to deal with small biological inputs, up to single-cell technologies (Rao et al. 2014; Furlan-Magaril et al. 2015; Nagano et al. 2015; Belaghzal, Dekker, and Gibcus 2017). However, I decided to start from a large number of cells (i.e. 10 Mln) in order to increase the complexity of the Hi-C libraries with the aim to obtain an interaction map at a higher resolution. Hi-C library complexity is a function of the initial number of cells that are used, as each cell contributes to a library with two alleles for each locus. Increasing the number of cells scales up linearly the amount of information for a specific locus. Moreover, the TG capture step requires micrograms of chromatin, therefore it was not feasible to reduce the number of cells in these experiments (Ilenia Simeoni et al. 2016).

BTPD relevant cell types cannot be obtained from primary tissues in the number needed for these experiments. There are protocols to differentiate MK and EC from circulating blood stem cells, however, these derived cells also have a limited expansion capacity (Zeigler et al. 1994; Martin-Ramirez et al. 2012; Perdomo et al. 2017). To overcome this limitation, I decided to *in vitro* differentiate hiPSC into the cell types of interest and to expand them to a number sufficient for the experiments. At the end of the differentiation protocol, I obtained homogeneous populations of MK, EC, and HEP (Fig. 4.2).

The differentiation stage of the cells was determined by flow cytometry, qPCR or enzymatic activity experiments (see chapter 2.2 for a detailed description). Briefly, after the twenty-day long differentiation protocol, MK maturity was assessed by flow cytometry using the cell surface markers CD41 and CD42b. CD41 is integrin-alpha-IIb that together with integrin-beta-3 forms the receptor for fibrinogen, and CD42b is glycoprotein (GP) Ib-alpha which with GPIb-beta, GPIX and GPV form the receptor for VWF. Both protein complexes are highly specific for the megakaryocyte-platelet lineage and essential for platelet function. CD42b, in particular, is considered to be a reliable late differentiation antigen that only appears on mature MK (Fig. 4.2.B). Brightfield images also show the floating MK that form clumps, as previously described in Moreau et. al 2016 (Fig. 4.2.A). HEPs were obtained following a thirty-day long cytokine differentiation protocol (Hannan et al. 2013). Brightfield images show a dense and tight monolayer of cells with the characteristics glycogen and cholesterol storage bodies (Fig. 4.2.C(Hannan et al. 2013). HEP maturity was also tested by

monitoring the enzymatic activity of CYP3A4 and the expression of key marker genes *NANOG, HNF4a and ALB* by qPCR (Fig. 4.2.D). These parameters indicate that the differentiated-hepatocytes have oxidase metabolic capacity and have lost pluripotency markers in favour of hepatocyte-specific ones (see chapter 2.2). ECs were generated following a nine-day long differentiation protocol. Brightfield microscopy images show a monolayer of cobble-stone shape *in vitro* differentiated endothelial cells (Fig. 4.2.E). The quality of the differentiated EC was checked, at the protein level, by measuring the expression of VE-Cadherin (VECAD; Fig. 4.2.F), which is a fundamental EC protein involved in cell adhesion and function (Crosby et al. 2005). The iPSC differentiation experiments produced the cell types (i.e. MK, EC and HEP) and number (i.e. 10^6 per library) required for the following capture TG Hi-C experiments.

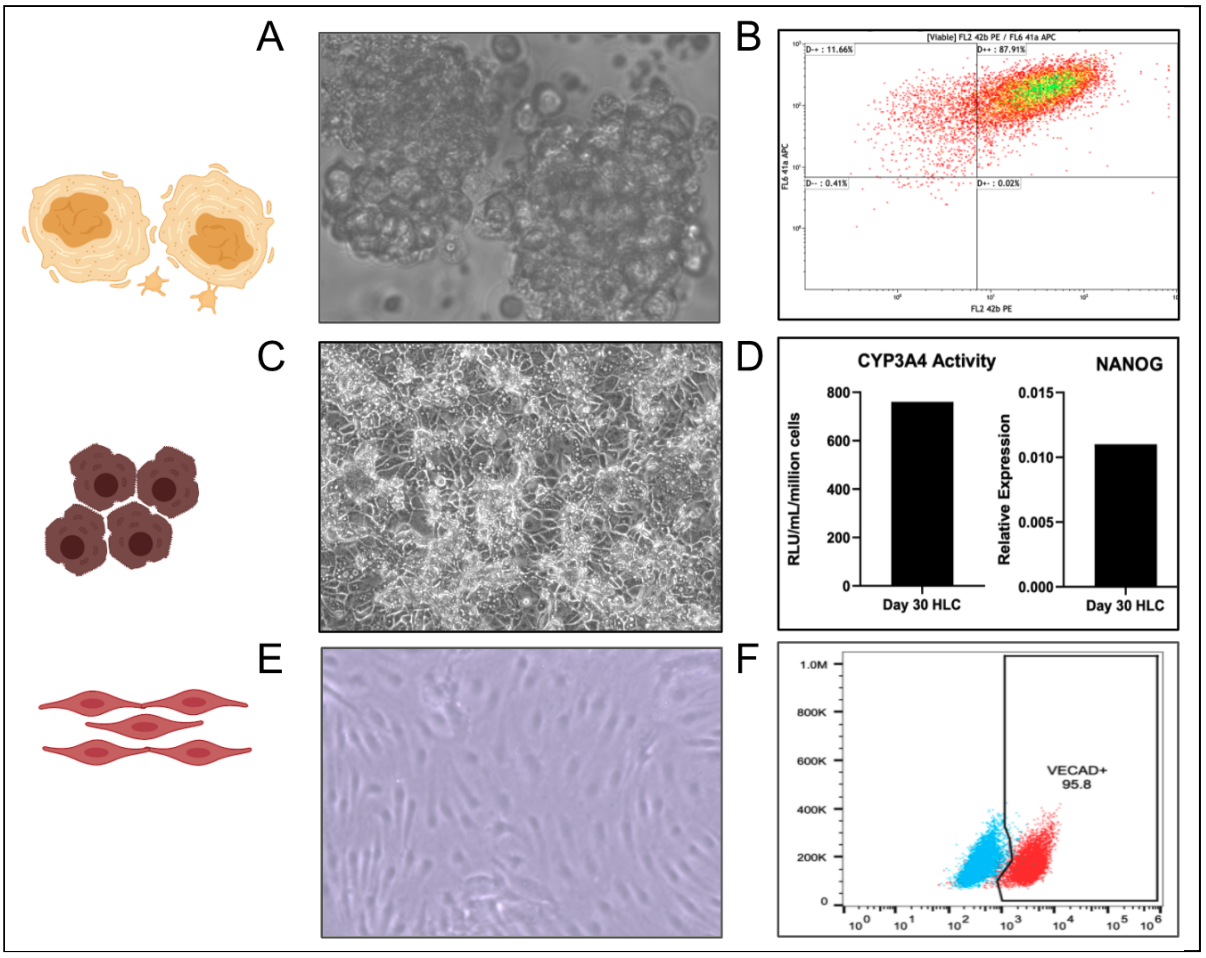


Fig. 4.2 | Cartoon of the differentiated cells using bright-field microscopy and other experiments to assess the effectiveness of the differentiation protocol*. A) A brightfield image of the differentiated MKs. B) An example of a flow cytometer experiment showing the percentage of CD41+ (y-axis) and CD42b+ (x-axis) for the megakaryocyte differentiation experiments. C) A brightfield image of the differentiated hepatocytes. D) level of CYP3A4 activity measured to assess the efficacy of the protocol for HEP differentiation. E) A brightfield image of the differentiated ECs. F) VECAD expression in the EC differentiated cells was measured via flow cytometry. See also chapter 2.2 for more details.*

4.3 TG Hi-C experiments produced a dense network of interactions

Fig. 4.3 shows an example of interactions identified with the TG Hi-C approach. In particular, the regulatory loops for the *GP1BA* gene. *GP1BA* gene encodes for one of the four proteins of the CD42 receptor complex for VWF and stimulates platelet adhesion at sites of vascular damage with disrupted EC architecture (see chapters 1.1.2 and 1.2.3; A. D. Michelson et al. 1986).

These chromatin interactions, which may be considered *bona fide* regulatory loops, are represented as green arches (Fig. 4.3). Interestingly, they colocalise with open chromatin regions (i.e. ATAC-Seq peaks), regulatory regions (i.e. H3K27Ac ChIP-Seq peaks) and binding sites for several of the key MK-specific transcription factors (Fig. 4.3). As discussed in the introduction, the regulatory elements may be far apart from the cognate gene. This is the case for the longest interaction reported for *GP1BA* (~280Kb on the 5' of the gene body) that binds to a regulatory region located within *NR_103482*, a lncRNA. This interaction bypasses 5 other genes and it would have been virtually impossible to predict its role in the regulation of *GP1BA* without the information coming from these kinds of experimental studies.

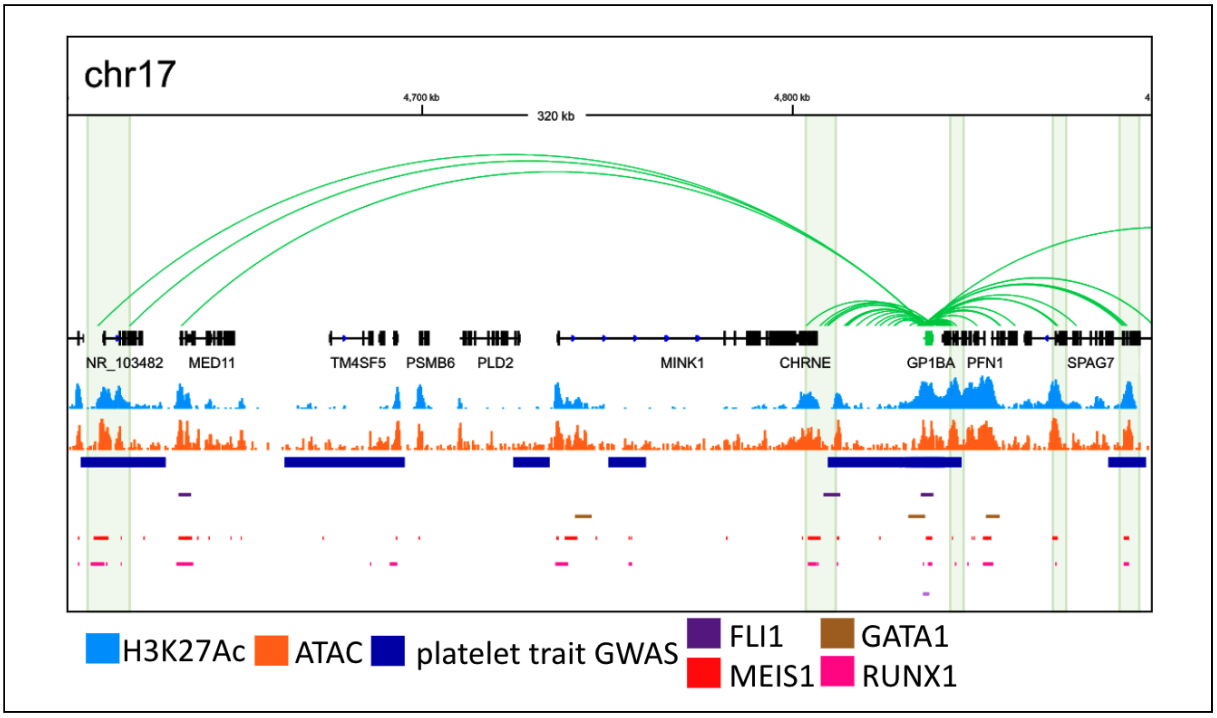


Fig. 4.3 | TG Hi-C MK interactions for GP1BA. *Bright green arches represent the interactions identified with the TG Hi-C experiments in MK. Black bars are the genes bodies. H3K27Ac and ATAC are the results of MK ChIP-Seq and ATAC-Seq, respectively. Platelet trait GWAS refers to the regions surrounding the SNPs identified in Astle et al. 2016. Vertical green bars highlight the regions of interest where arches overlap with MK regulatory regions.*

4.4 Comparison of the TG Hi-C results with pcHi-C

In order to test the improvement in resolution and depth of interactions, I compared my results with the data obtained with a similar workflow and cell type (i.e. MK, Fig. 4.4; Javierre et al. 2016). Even though the total number of interactions genome-wide is lower in my experiments, Fig. 4.4 indicates that the TG Hi-C increased the resolution of interactions for the 93 genes of interest. The number of statistically significant regions associated with a gene on average increased from 7 in Javierre et al. to 239 in my experiments (Fig 4.4.B).

Two examples of these differences are shown in Fig. 4.4. *MYH9* is a gene involved in macrothrombocytopenia, deafness and kidney failure. In the TG screening, only 21 out of 50 cases of suspected *MYH9* disorder have been fully explained (Downes et al. 2019). The TG platform mainly focused on the coding regions of the diagnostic grade genes, therefore it may be possible that a fraction of these unexplained cases is due to the presence of rare variants in non-coding regulatory regions. pcHi-C interactions from Javierre et al. have limited resolution for this gene and the previously published results were not sufficient to explore the MK regulatory landscape of *MYH9*. Instead, the TG Hi-C approach, focusing on a modest number of genes, expanded the number of cognate regulatory regions by several folds. Moreover, the TG Hi-C approach returns information on the structure of the gene body (Fig. 4.4, the dense network of interactions in the *MYH9* gene body). Exons and introns chromatin structure has been recently associated with an alternative transcription and therefore affecting translation, protein structure and function (Ruiz-Velasco et al. 2017). The second example concerns the *VWF* locus (Fig. 4.4.C). The plasma protein VWF, which is mainly synthesised by EC (but to a lesser extent also by MK) is a key protein in haemostasis, it is a chaperon for F8 and as mentioned is essential in platelet adhesion to the damaged vessel wall (see chapter 1.1.2; Sadler 1998). The *VWF* locus showed in the pcHi-C experiment (Javierre et al. 2016) good resolution, but the TG Hi-C experiment identified additional interactions (Fig. 4.4). The intricate network of interactions is a reflection of the genetic features that characterise the entire region. Indeed, this chromosomal area contains 4 strong enhancers that define a super-enhancer region (Petersen et al. 2017).

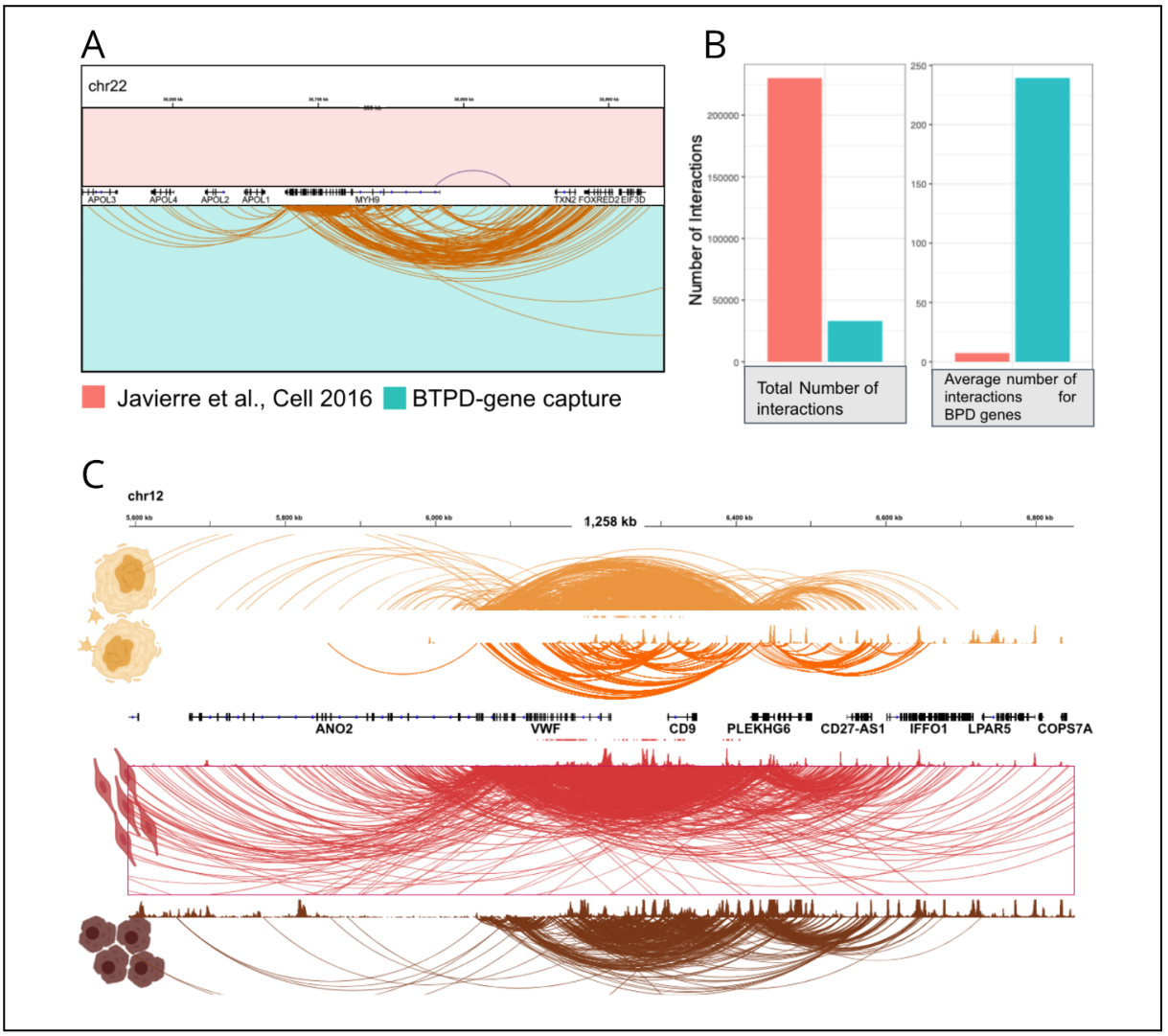


Fig. 4.4 | Comparison of the MK TG Hi-C data with the pcHi-C ones from Javierre et al. 2016. *A) Example of MYH9 interactions and comparison between pcHi-C and TG Hi-C. B) bar plots for the number of interactions identified genomewide and for the BTPD genes. C) VWF region chromatin interactions. Orange, red and brown lines refer to the interaction data identified in MK, EC and HEP, respectively. Light orange, the top track of the orange interactions, reports the interactions identified with the TG Hi-C approach. Dark orange, the bottom track of the MK interactions, reports the interactions identified in pcHi-C (Javierre et al. 2016).*

4.5 Description of the TG Hi-C interactions: regulatory space, length of the interactions and captured genomic features

After sequencing, analysis and quality controls (see chapter 2.3), the TG Hi-C experiments produced 62,027 different interactions for the three BTPD cell types, directed towards 17,228 regions (Table 4.1). The number of unique regions is lower than the number of unique interactions because these are redundant, indeed multiple viewpoints (i.e. baits) interact with the same region. These TG Hi-C interactions expanded the genomic regions defined in the TG bait set (i.e 3,092,645 bp) to about 19.5 Mb of regions with potential regulatory function, roughly a third of the size of the human exome (Clark et al. 2011).

	Unique interactions	**Unique regions**	**Prey-defined genomic space (bp)**	**Increase on TG bait space (bp) Overlap with bait removed**
MK	37,176	15,856	4,000,592	2,606,789
EC	18,429	13,874	5,153,562	3,249,987
HEP	9,809	4,983	17,669,470	16,485,090
Cumulative	62,027	17,228	24,277,022	19,653,540

Table 4.1 | Summary statistics of the TG Hi-C interactions. The differences in the number of interactions reported in Table 4.1 are due to overlapping regions having been counted only once.

Fig. 4.5 shows the summary statistics describing the width of the interactions (distance from the viewpoint) and the ratio of cis-to-trans interactions. When centred on the viewpoint, it is quite clear that the number of interactions is a function of the distance from the bait that decay exponentially (Fig. 4.5 A, D and G), the further away from the bait the lesser interactions are captured. The gap that surrounds the regions immediately after the baits is the result of the statistical processing used in the CHiCAGO algorithm, which removes the regions immediately flanking the viewpoint (evident in Fig. 4.3.G; see chapter 2.3). The plots in Fig. 4.5 also show that there is no directionality of interaction relative to the viewpoint position, as one would expect from the enhancer-promoter interactions. Indeed, enhancers can have any position with respect to the cognate genes (Alberts 2014).

Panel B, E, H of Fig. 4.5 and Table 4.2 show that the majority of interactions have distances in the order of tens to hundreds of thousands of bp. These are the lengths of the interactions that are most enriched for regulatory loops and consequently could be involved

in enhanceropathies (i.e. diseases caused by an altered function of a regulatory region; Sanyal et al. 2012; Jin et al. 2013; van Arensbergen, van Steensel, and Bussemaker 2014; Laugsch et al. 2019; McCord, Kaplan, and Giorgetti 2020; Turro et al. 2020b; Thaventhiran et al. 2020).

The ratio cis-to-trans interactions is a good proxy for the library quality, with the idea that the majority of interactions reside on the same chromosome, in cis. Indeed the CHiCAGO algorithm uses this ratio to normalise the Hi-C signal (Javierre et al. 2016; Cairns et al. 2016a). The majority of the statistically significant interactions in the TG Hi-C experiments are in cis (Fig4.5.C-F-I). Although some of the *trans* interactions are potentially interesting for their biological role, I did not consider them for the following experiments, analysis and discussions due to the difficulties in their interpretation (Noordermeer et al. 2011; Schoenfelder, Sugar, et al. 2015).

	Mean (bp)	Standard deviation (bp)	Median (bp)	Max distance (bp)
MK	81,947	350,753	6,387	45,572,825
EC	156,547	2,078,217	38,592	190,120,227
HEP	100,053	1,399,014	29,542	87,796,604

Table 4.2 | Summary statistics of the length of the TG Hi-C interactions

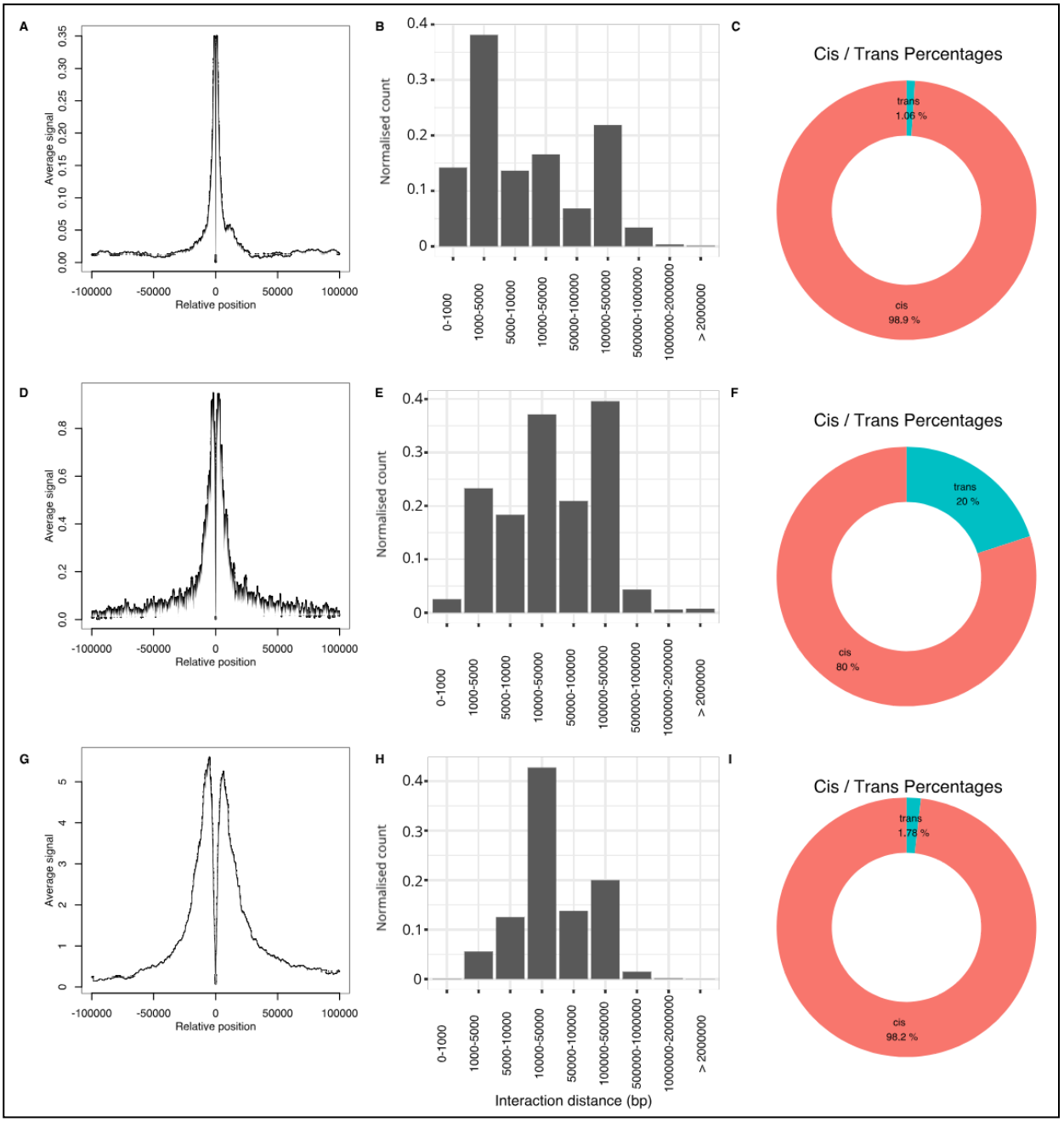


Fig. 4.5 | Summary description of the interactions in the three different cell types. *First, second and third row (i.e. A, D and G) refer to the results obtained with MK, EC and HEP, respectively A, D and G are density plots for the distribution of the length of TG Hi-C interactions centred on the viewpoints, distances are expressed in bp. B, E and H are barplot still describing the distances of interaction, distances are binned in 9 groups. C, F and I show the percentage of cis (red) versis trans (green) TG Hi-C interactions.*

Lastly, I annotated the interactions with the overlapping genomic features using ENSEMBL (v99) and BLUEPRINT data (Fig. 4.6; Adams et al. 2012; Aken et al. 2016; Stunnenberg et al. 2016, and Hirst 2016b). Theoretically, a genomic region could contain two or more genomic features, for instance, intron and enhancer. In order to assign only one

feature to each interacting region, I adopted prioritisation criteria. As a result, the regions overlapping with two or more features would be reported only for the highest-ranking one. The assigned features, ranked from the highest to the lowest, are: exons; introns; promoter; promoter flanking region; enhancer; enhancer flanking region and intergenic region.

The three cell types are quite heterogeneous in the features interacting with the baits (Fig. 4.6), indeed the profile of the interactions across the different genomic features (i.e. enhancers, exons, introns, intergenic regions and promoters) is not constant. This could be a reflection of the capture probe design, the differential transcription of the BTPD genes in three cell types and also of the endogenous biological differences between the three types of cells. For instance, EC has the least number of expressed genes in the TG panel and these genes will be silenced in endothelial cells. It is unlikely that these inactive genes interact with enhancers and therefore these interactions will be enriched for non-enhancers regions (i.e. intergenic regions; according to the definition I adopted in this analysis). Notably, the profile of the total interactions resemble the MK one (Fig. 4.6.A and D), this is possibly because of the high contribution of interactions coming from the MK compared to the whole number of interactions (Table 4.1).

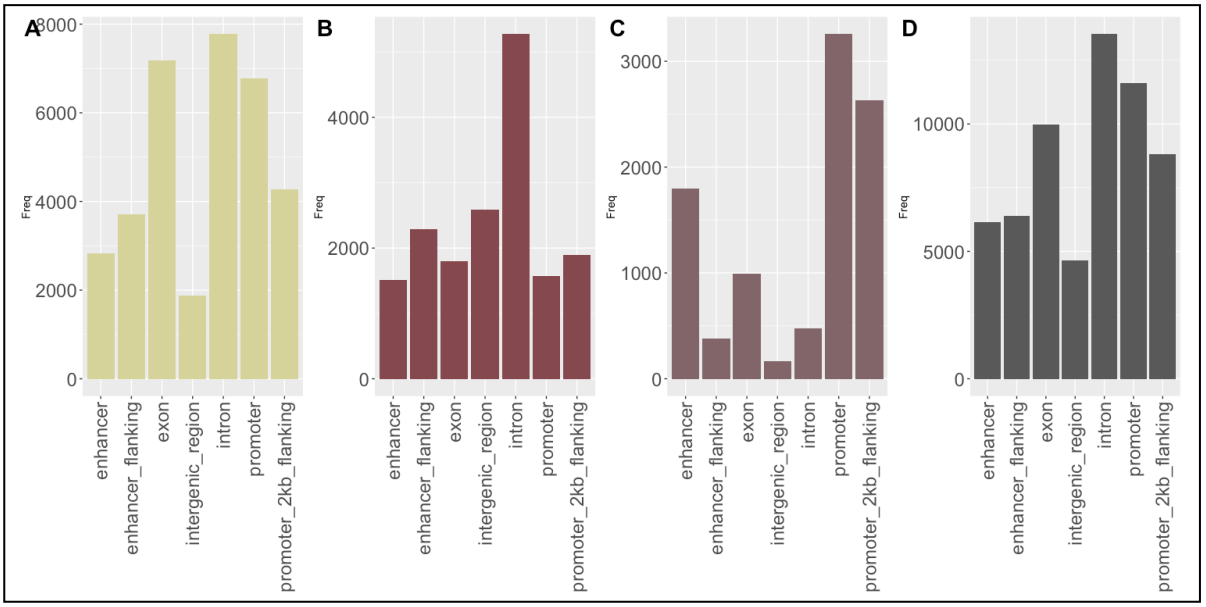


Fig. 4.6 | Number of interactions with preys overlapping the different genomic features in MK (A), EC (B), HEP (C) and cumulative (D)

This data demonstrated a good enrichment for prey regions containing enhancers (Fig. 4.6.D). The TG Hi-C approach identified 6,187 interactions of prey with enhancers, showing an enrichment for this regulatory feature (z-score=1087.098; p-value=0.009, permutation analysis, 100 permutations). Similar permutation analysis showed that MK

interaction data are also enriched for CTCF binding regions (z-score=77.443; p-value 0.0099, permutation analysis, 100 permutations).

Interestingly, the ranked score distribution for H3K27Ac, which is a proxy for enhancer, and CTCF signals is different between the captured regions and the genome-wide ones (Fig. 4.7). The interactions captured the entire spectrum of enhancers, from the weakest to the strongest. On the other hand, CTCF regions were captured only in the mid-to-low signal intensity, with a difference between the distribution of the whole genome and captured signals (p-value=2.2e-16, Kolmogorov-Smirnov test).

A possible interpretation for this difference is that CTCF has more dynamic and less stable interactions with the DNA when regulating gene expression and longer residence time when engaged with cohesin and involved in the definition of the chromatin structure. A shorter residence time on the DNA results in a weaker ChIP-Seq signal at the population level when CTCF works as a transcription factor. In fact, studies of the kinetics of CTCF residence time reported different behaviours of this protein, suggesting the existence of different functions of this molecule (Hansen et al. 2017; Agarwal et al. 2017).

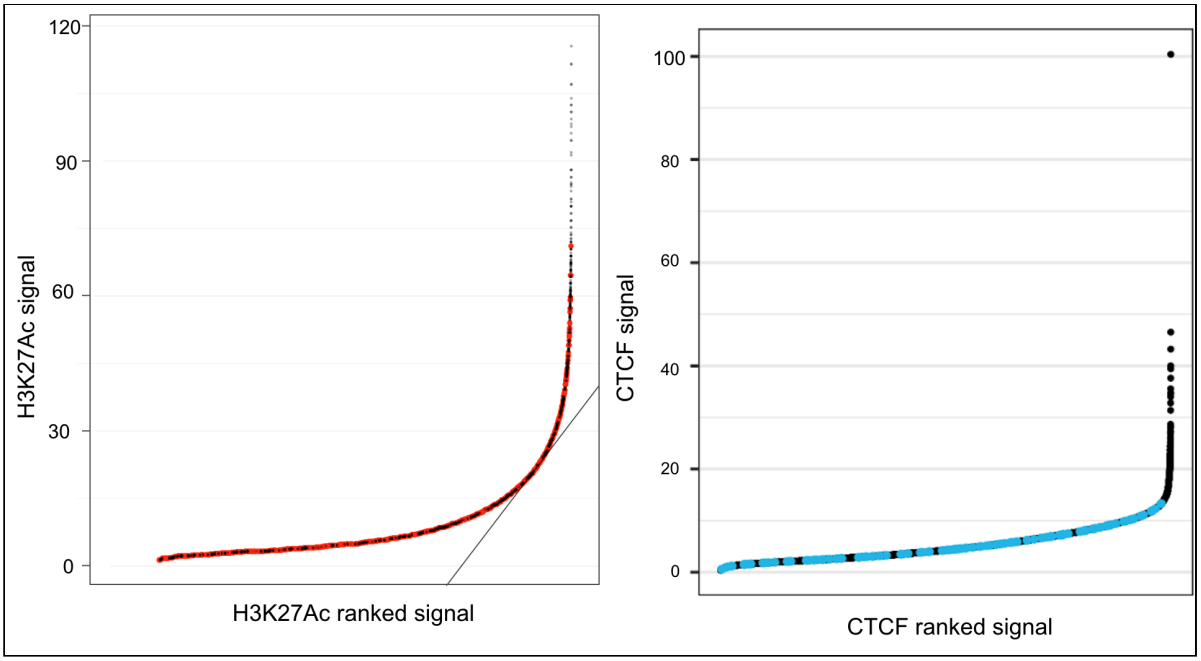


Fig. 4.7 | H3K27Ac and CTCF ranked signal for MK genomic features. *Red dots are the enhancers captured in the TG Hi-C experiments. Blue dots are the CTCF peaks overlapping with the regions captured by the TG Hi-C. Black dots are the genome-wide background.*

4.6 Promoter interactions

Non-coding regulatory regions (i.e. enhancers) exert their regulatory effect mainly via direct interaction with promoters (Javierre et al. 2016). Different cells make different use of regulatory regions (Bulger and Groudine 2011; Ong and Corces 2011; Spitz and Furlong 2012; Lelli, Slattery, and Mann 2012; Calo and Wysocka 2013; Hnisz et al. 2013; Petersen et al. 2017) and they also use alternative promoters, which allow for different transcripts to be expressed (Kimura et al. 2006; Xin, Hu, and Kong 2008). Indeed, resources such as ENSEMBL report lists of cell-type specific promoter regions, generally relying on the regions surrounding the transcription starting sites (TSS) for cell-specific transcripts (Zerbino et al. 2015, 2016).

H3K27Ac is an epigenetic modification that marks not only active enhancers but also promoters (Creyghton et al. 2010). Therefore, the acetylation chromatin state is a layer of information that can be used to better define cell-specific promoters. While RNA-Seq experiments can be used to position the TSS. Combining the information coming from these two high-throughput experiments allowed me to determine the promoters locations specific for the cell types used in the TG Hi-C experiments (Fig. 4.8). Results presented in Fig. 4.8 show the acetylation levels surrounding the TSS using experimental data and compare them with the acetylation on the standard promoter regions defined in ENSEMBL (Zerbino et al. 2015, 2016). This analysis gave better positioning of the promoters for the BTPD genes to use in mapping the promoter interactions.

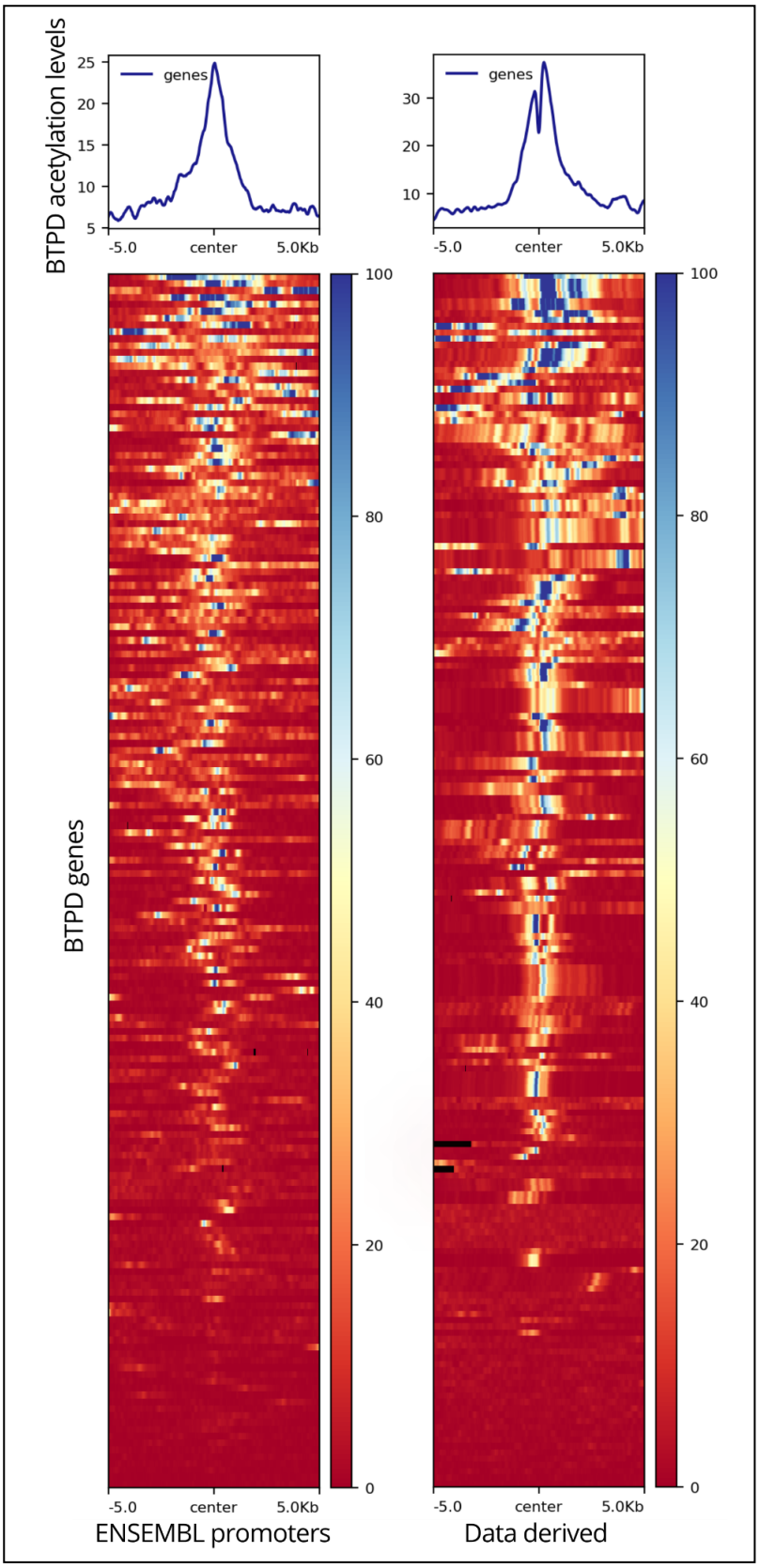


Fig. 4.8 | H3K27Ac peaks on the BTPD promoters, defined by ENSEMBL and in the promoters derived by experimental evidence. *The heatmap colour coded the normalised level of H3K27Ac surrounding the BTPD TSS.*

Fig. 4.9 shows promoter-specific MK interactions with other genomic features. The interaction length distribution is similar for almost all the features taken into account, namely exons, introns, CTCF regions and enhancers. Overall, the distances observed for these interactions confirm the lengths observed in Fig. 4.6. The majority of interactions occur in the order of tens to hundreds of Kb, meaning that we are selecting loops that are involved in gene regulation (McCord, Kaplan, and Giorgetti 2020). Their lengths relate to the ones observed in other studies that look at sub-TADs regulatory interactions (Sanyal et al. 2012; Jin et al. 2013; van Arensbergen, van Steensel, and Bussemaker 2014; McCord, Kaplan, and Giorgetti 2020).

Promoter-promoter interactions, which go from the promoter of one gene to the promoter of another gene, act differently. The majority of these interactions are in the order of tens of Kb, with a long tail in the distribution, in the order of hundreds of Kb. Promoter-promoter interactions have been observed in several other studies, however, their role still requires better characterisation (G. Li et al. 2012; Javierre et al. 2016; Dao et al. 2017; Jung et al. 2019; Qin et al. 2020). It is still debated if these interactions are genuine or biological artefacts. Co-expressed genes colocalise at transcription factories (i.e. speckles; Spector and Lamond 2011; Ha 2020) and the Hi-C may reflect their proximity. More interestingly, some promoters could have a function similar to the enhancers, promoting the expression of the genes that they are interacting with (G. Li et al. 2012; Schoenfelder, Furlan-Magaril, et al. 2015; Javierre et al. 2016; Dao et al. 2017; Jung et al. 2019; Qin et al. 2020).

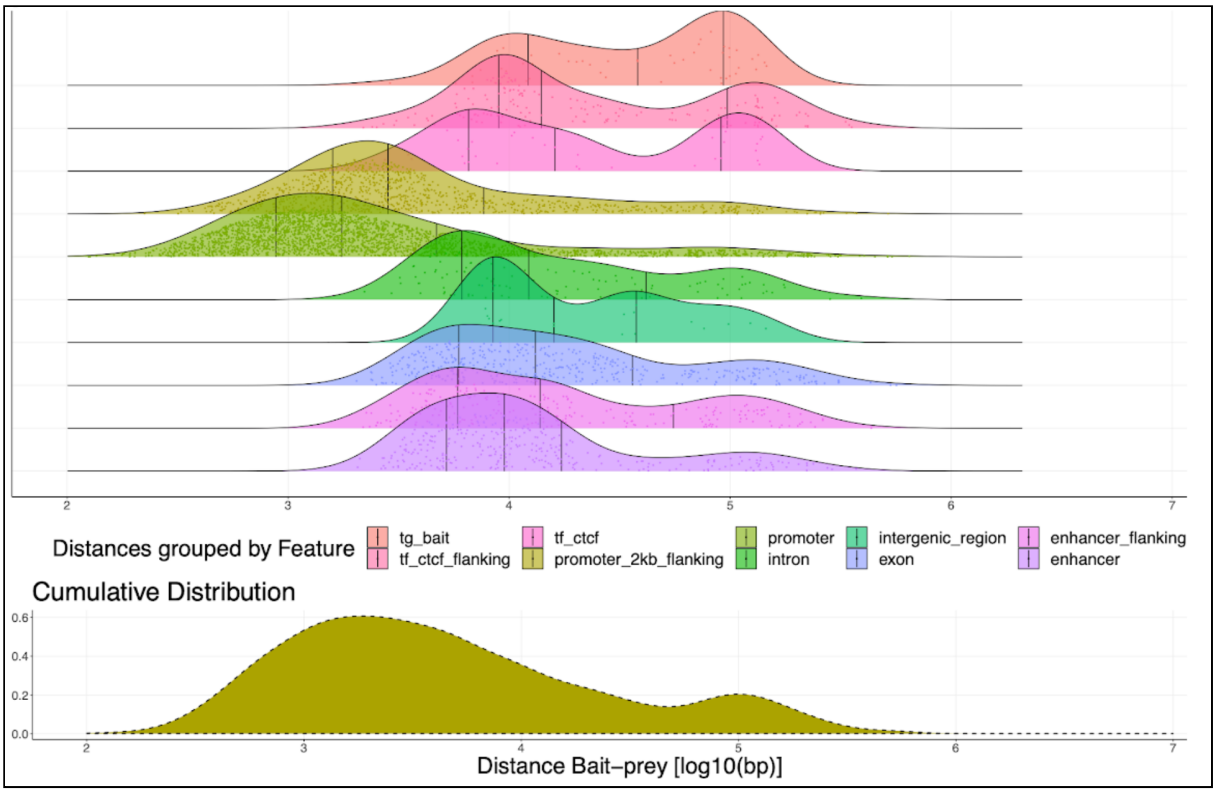


Fig. 4.9 | Distribution of the length of the promoter interactions identified in the TG Hi-C experiments for MK. *The x-axes are shared between the cumulative distribution and the detailed one. The flanking regions refer to a one-Kb region surrounding the relative genomic feature.*

4.7 Different cell types make different use of their regulatory space

The differential use of regulatory elements located mainly in the non-coding DNA (Fig. 4.6 and Fig. 4.9) drives development, cell differentiation and defines the identity of mature cells (Bulger and Groudine 2011; Ong and Corces 2011; Spitz and Furlong 2012; Lelli, Slattery, and Mann 2012; Calo and Wysocka 2013; Hnisz et al. 2013; Petersen et al. 2017). Furthermore, the epigenetic landscape allows cells to respond to different stimuli (Allis and Jenuwein 2016; Klemm, Shipony, and Greenleaf 2019; Soskic et al. 2019). These distinct biological characteristics are a reflection of the differential gene expression that is driven by the different use of regulatory space (Sanyal et al. 2012; ENCODE Project Consortium 2012). This differential use of the regulatory space, amongst different cell types, can be appreciated also in the patterns of interactions, as previously reported (Javierre et al. 2016).

The TG Hi-C interactions observed in the three different cell types corroborate these observations (Fig. 4.10). Indeed, using the regions captured with the TG Hi-C it is possible to separate the three cell types using the principal component analysis (PCA) and the Jaccard distance (Fig. 4.10, A and B). Interestingly, the dendrogram built on the Jaccard distance of the prey regions places MK and EC closer to each other in respect of the HEP. This similarity between EC and MK was suggested before, possibly because of their shared ontogenetic origin and both being mesoderm derived versus HEP being endoderm-derived. (see chapter 1.1.2; Choi et al. 1998; Rafii and Lyden 2003; Thiele et al. 2012; Malara et al. 2015) The similarities in the patterns of interaction in the BTPD genes confirmed the similitude between the cells. Zooming into the data and looking at the single promoters (Fig. 4.10.C), PCA again shows a separation between the three cell types, with EC and MK promoters closer to each other (physical proximity in PCA can be considered similarity). However, it is worth noting that, at least in part, these differences may be driven by the use of different restriction enzymes used in the production of the TG Hi-C libraries for HEP (see chapter 2.3).

Differences in the interaction amongst the cell type are visible in the *RUNX1* gene (Fig. 4.11). *RUNX1* is a transcription factor that is a master regulator of haematopoiesis (Sood, Kamikubo, and Liu 2017). Rare pathogenic and likely-pathogenic variants in this gene are causing familial thrombocytopenia accompanied by myeloid leukaemia (Sood, Kamikubo, and Liu 2017). *RUNX1* is highly transcribed in MK and to some extent in EC

(Grassi et al. 2020), while completely silenced in HEP (Blood Atlas and GTEx data accessed in September 2021). The chromatin interactions and H3K27Ac peaks show how the regulatory backgrounds are different in the different cells, as is also visible in the PCA.

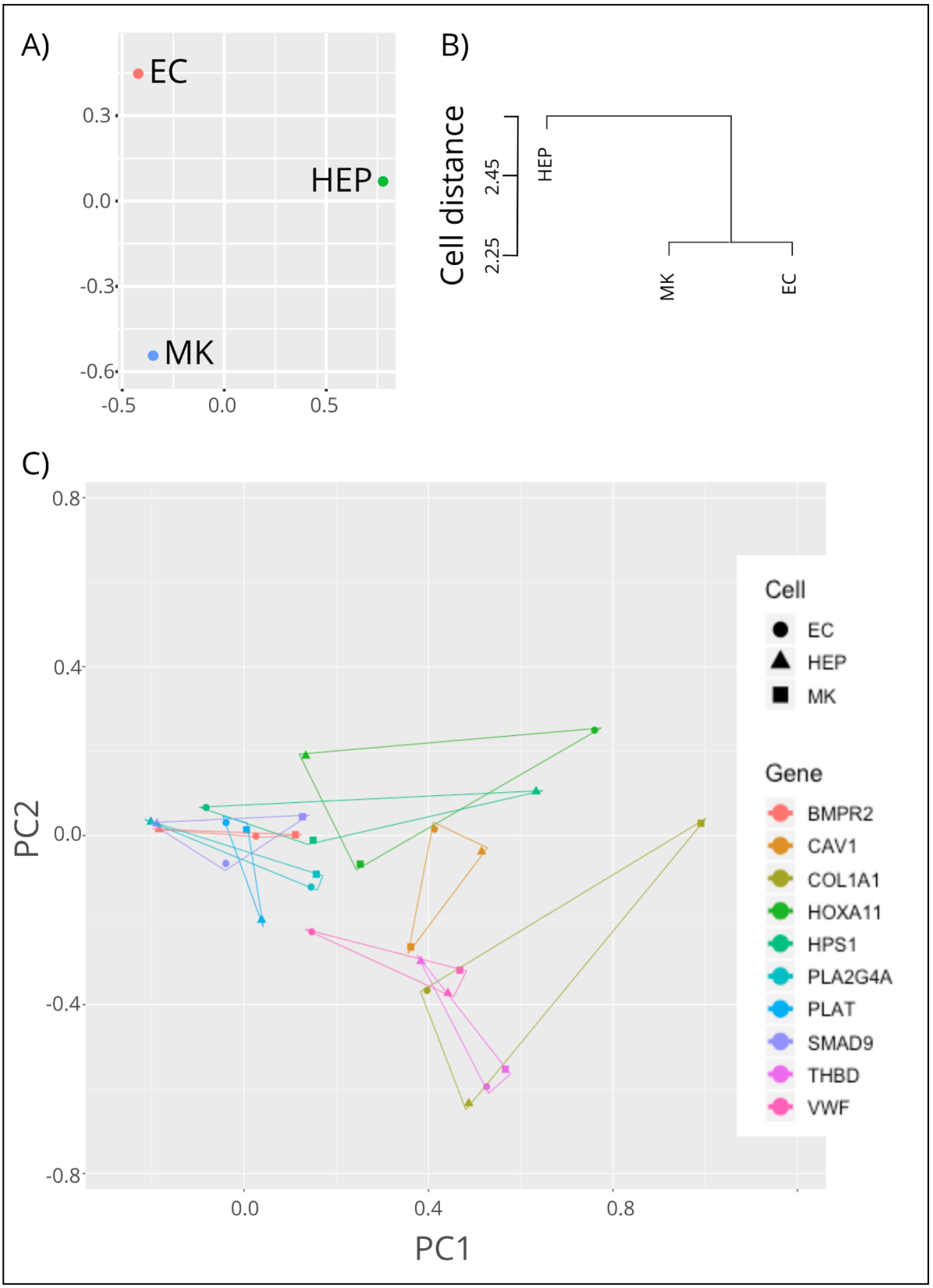


Fig. 4.10 | Differences in the interactions observed in the different cell types. A) PCA for the prey interactions identified in the TG Hi-C experiments. *B) Dendrogram reporting the calculated distance built on the Jaccard similarity coefficient score. C) PCA of promoter interactions for specific examples.*

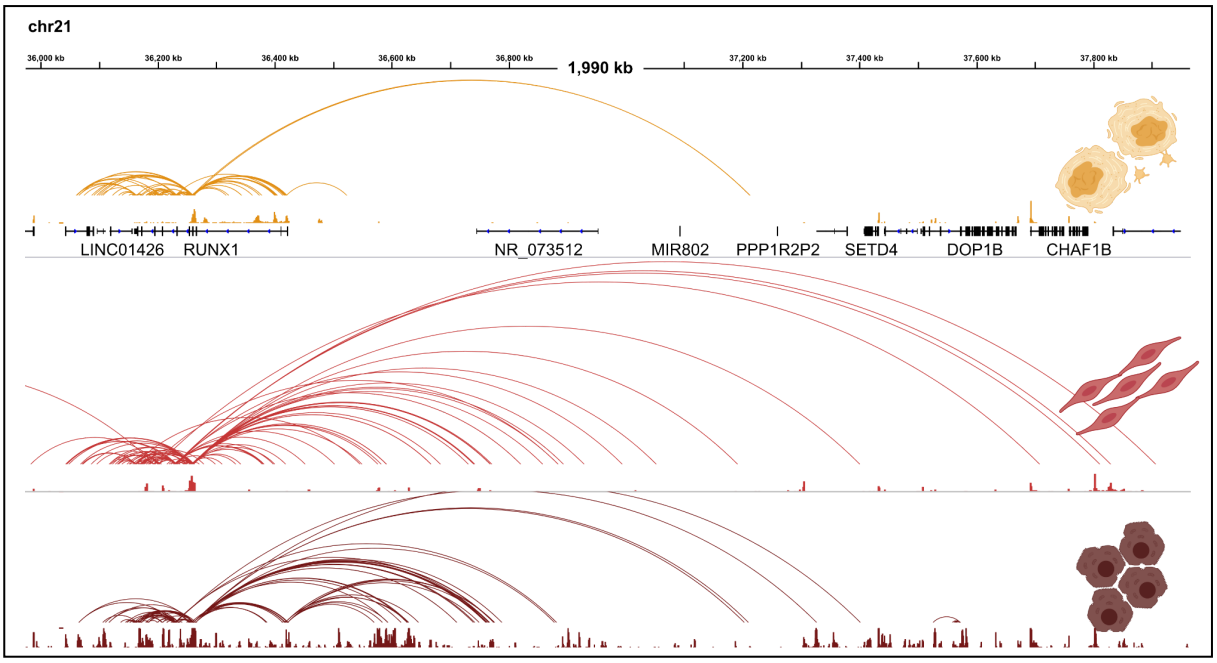


Fig 4.11 | Examples of the chromatin structures in the three different cell types. *RUNX1 gene chromatin interactions. Orange, red and brown lines refer to interactions identified in MK, EC and HEP. Each cell type has chromatin interactions (top track) and H3K27Ac (bottom track).*

4.8 Interactions regulate gene expression

Enhancer-promoter loops place in spatial proximity the transcription machinery components required to regulate and transcribe genes in a cell-type specific fashion (Spitz and Furlong 2012; Klemm, Shipony, and Greenleaf 2019). A limited number of studies have investigated the regulatory effect of multiple enhancers on the same gene and tried to generalise and define biological and statistical models to explain gene regulation (Schmidt, Kern, and Schulz 2020; Zrimec et al. 2020; Boettiger and Murphy 2020).

Several models for the enhancer effect on regulation co-exist (Buecker and Wysocka 2012; Dukler et al. 2016). For instance, multiple enhancers acting on the same gene can have an additive effect and these multiple interactions can linearly increase the transcription levels (Hay et al. 2016). Also, other non-additive models have been proposed, such as multiplicative or exponential, in which the effect on the transcription of the two enhancers co-regulating the same gene is more than the sum of the single enhancer (Bothma et al. 2015; Lam et al. 2015; H. Y. Shin et al. 2016). TG Hi-C interactions, intersected with the BLUEPRINT gene expression data for MK, could provide further evidence in support of the models mentioned above.

The correlation between the number of promoter interactions and the gene expression levels is plotted in Fig 4.12. Panel A shows that a linear model does not identify any significant correlation between the TG Hi-C interactions and the expression for the BTPD genes in MK (p-value=0.265). However, when limiting the interactions to the promoter-enhancer ones, a significant correlation emerges from the data (p-value=1.075e-04), explaining 20% of the variability observed. This experiment suggests that the number of promoter-enhancer interactions is a discrete predictor of the gene expression and, moreover, they also point towards an additive effect of multiple enhancer interactions to the same gene.

Fig. 4.12 also shows that some genes are highly expressed despite having only a few interactions and vice versa, indicating that genome regulation is too complex to be predicted with a single model and probably these genes undergo different mechanisms of regulation.

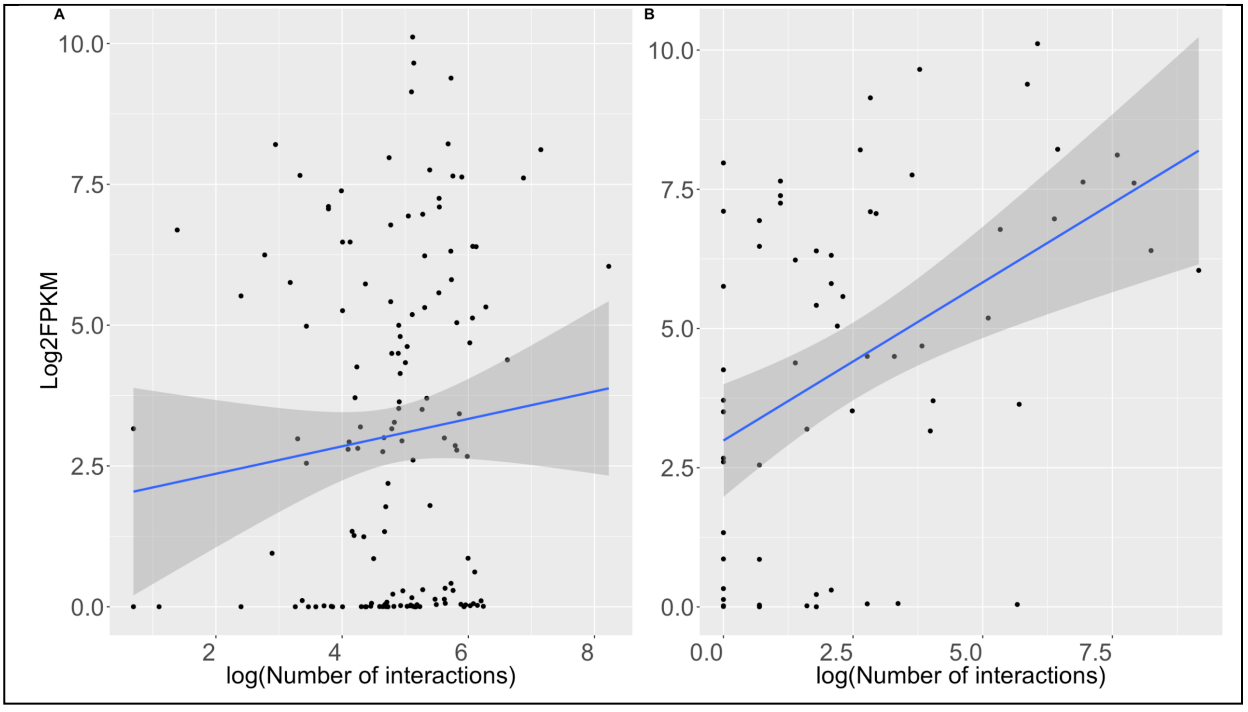


Fig. 4.12 | Correlation between the number of interactions and transcription levels for all MK interactions (A) and only MK promoter-enhancers interactions*. Distances and expressions are expressed in a log-log scale, which is commonly adopted for genomic studies (Zrimec et al. 2020).*

From the integration of the interaction and transcription data, there is another interesting consideration: the length of the interactions and the gene expression status are correlated (Fig. 4.13). Highly expressed genes (i.e. log2FPKM > 6) tend to have shorter interactions compared to genes which are not expressed ones (i.e. log2FPKM < 1). These

differences in length are statistically significant for promoters (Wilcoxon rank test, p-value=7.2e-103), promoter flanking regions (Wilcoxon rank test, p-value=1e-138), exons (Wilcoxon rank test, p-value=5.4e-120) and introns (Wilcoxon rank test, p-value=1.3e-114). This could be a reflection of the chromatin state. For example, a gene that is not expressed tends to be in chromatin dense regions, possibly heterochromatin (see chapter 1.3.5), and this location puts in closer proximity two regions far away from each other, hence longer interactions. Contrarily, euchromatin regions have DNA that is more accessible, resulting in a more untangled and spreaded string of nucleotides (Buenrostro et al. 2013). In fact, accessible chromatin is required to allow transcription factors and other proteins to access and bind the DNA. These physical characteristics may contribute to the observed differences in the interaction length distribution reported in Fig. 4.13.A. Interestingly, this figure is inverted when considering only promoter-enhancer interactions (Fig. 4.13.B), indeed these interactions are characterised by longer distances in active genes and shorter ones in repressed genes (Wilcoxon rank test, p-value=5.72e-05). This observation suggests that long-distance interactions, in transcriptionally active genes, are predominantly happening with enhancers, whilst they are more disordered in repressed genes.

Two illustrative loci that show the different lengths of interactions are *TUBB1* and *GATA1* (Fig 4.14). *TUBB1* is highly expressed in MK, while not expressed at all in EC (Grassi et al. 2020) and this molecular characteristic is associated with short interactions in MK and long interactions in EC (Fig 4.14.A), following the general observation reported before (Fig. 4.13). Similarly, *GATA1*, which has the same pattern of gene expression observed for *TUBB1,* also has longer interactions in EC. Interestingly, MK interactions, identified with the two different experiments (i.e. pcHi-C and TG Hi-C) show a similar range of interactions for these 2 genes (Fig. 4.14).

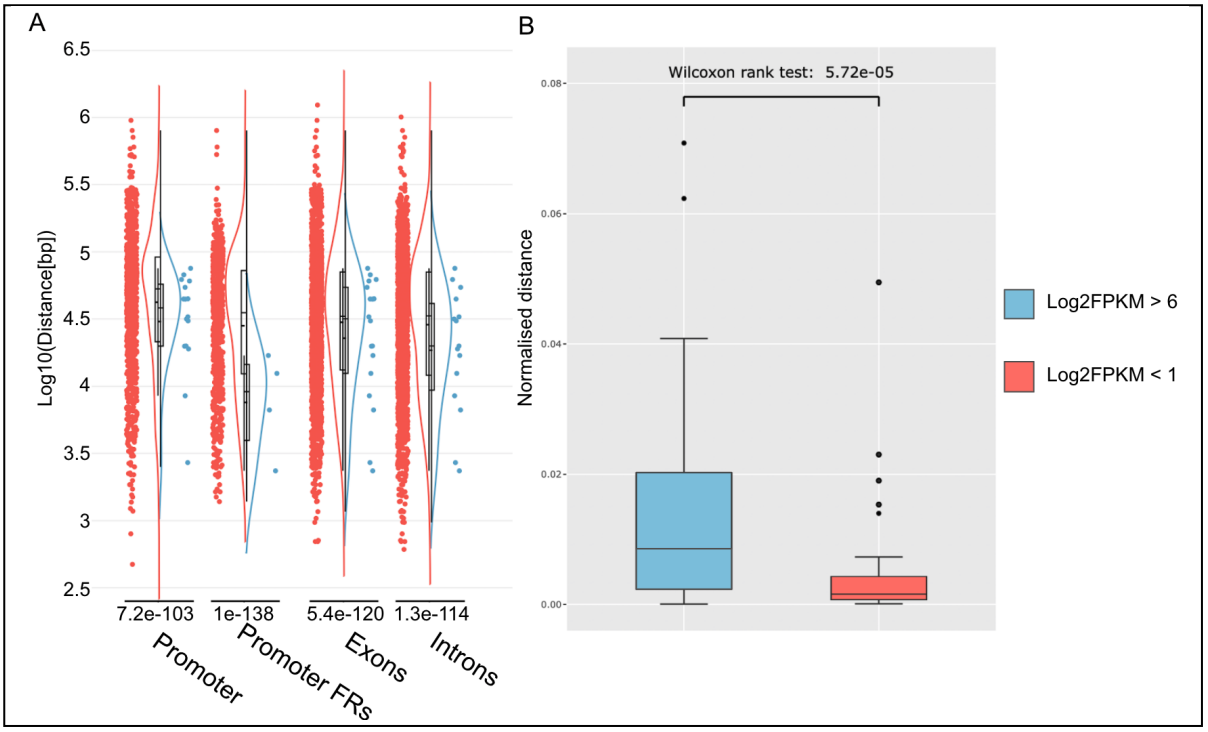


Fig. 4.13 | Differences in the length of interactions between highly expressed and repressed genes. *A) Length of interactions for all TG Hi-C interactions. Length is expressed as log10 bp. p-values are the result of the Wilcoxon rank test. Promoter flanking regions refers to the two-Kb region surrounding the promoters. B) Length of interactions for promoter-enhancers TG Hi-C interactions. Interactions are expressed distance relative to the longest one observed in MK. FRs = flanking regions.*

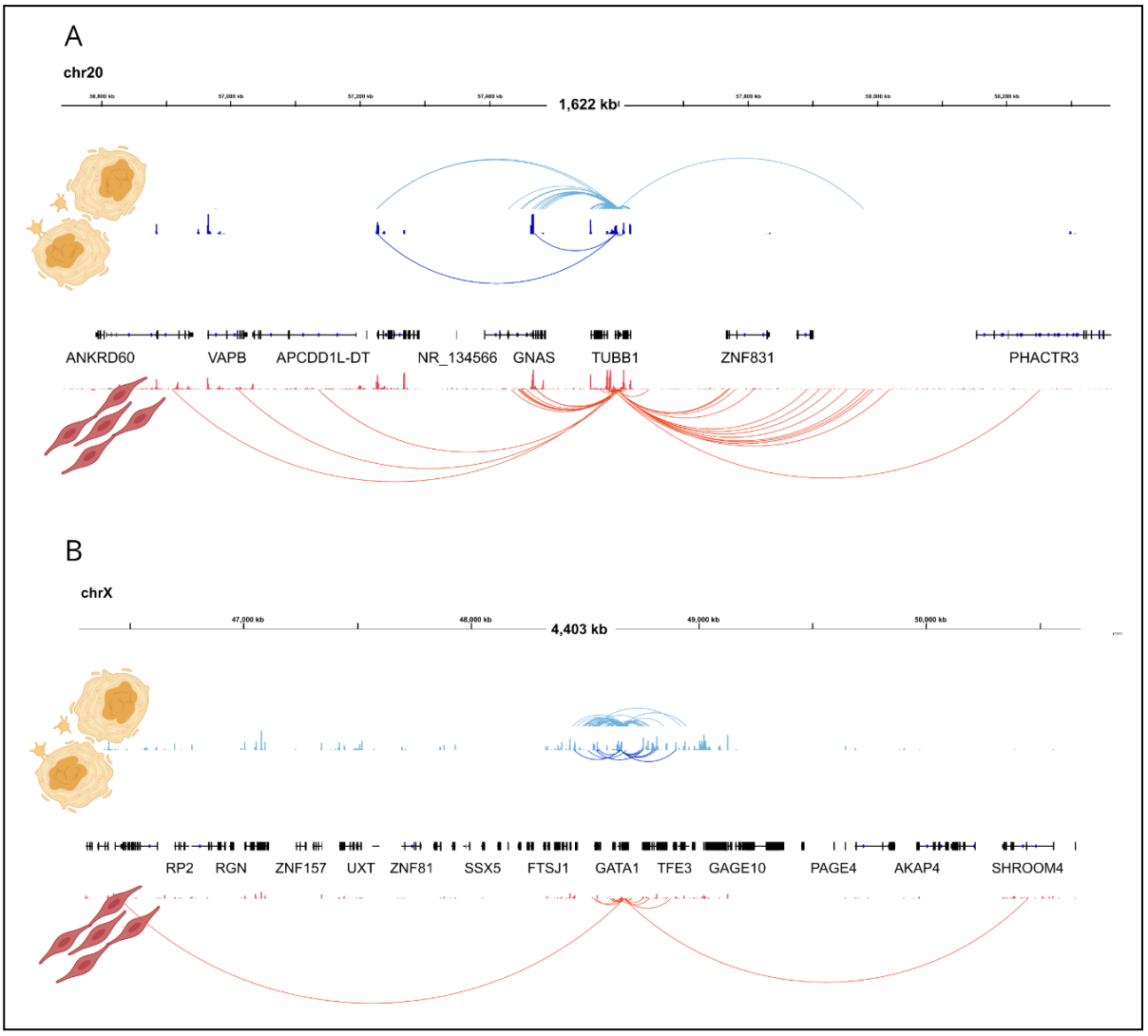


Fig 4.14 | Examples of the different interaction lengths in expressed (blue) and not expressed (red) genes. *A) TUBB1 and B) GATA1 gene and region. Respective cell types (HEP, MK, EC) are reported as cartoon images on the left of the panels. Light blue is the interactions identified in the TG Hi-C approach and dark blue are the interactions identified with the pcHi-C technique (Javierre et al. 2016).*

4.9 Interaction map in colocalisation studies can improve variant assignment

Biological interpretation of SNPs identified by GWAS is challenging and the variant-to-trait association should be considered in the context of cell-specific epigenetic information (Phylipsen et al. 2010; Pers et al. 2015; Lichou and Trynka 2020). The associated SNPs in the non-coding space are particularly challenging because they might exert their effect on genes located hundreds of thousands bp removed from the sentinel SNP (Phylipsen et al. 2010; Astle et al. 2016b; Javierre et al. 2016; Petersen et al. 2017). Moreover, some trait-associated variants can be embedded in the body of a gene but exert their effect on another one. A notable example of this is the *FTO-IRX3* regulatory loop. GWAS determined the association of rs9930506 as a sentinel SNP associated with body weight (Frayling et al. 2007; Gamazon et al. 2013). This variant lies within an intron of the *FTO* gene. Initial experiments on the function of the *FTO* failed to find a compelling explanation for the observed association. Years later it was reported that the *FTO* intron harbours an enhancer that regulates *IRX3* expression and silencing of *Irx3* in mice does modify bodyweight (Smemo et al. 2014). All together the initial assumption that the *FTO* gene was associated with bodyweight has shown to be false whilst the GWAS-identified SNP regulates the transcription of a gene localised 500 Kb downstream of the variant.

Colocalisation analyses can be used to infer the biological mechanism that is underlying the GWAS SNP associations, relying on the intersection of GWAS leading SNPs with evidence coming from other experimental sources (Wallace 2020). This approach can improve PRS and GWAS (Fig. 4.15) interpretation but needs support from other studies such as TG Hi-C. The regulatory loops can be used to link variants to genes, connecting a non-coding variant to a possible explanation.

F5 is a cofactor in the coagulation cascade and, together with *F10,* activates thrombin (see chapter 1.1.2). Its impairment is associated with both bleeding and thrombosis (Bertina et al. 1994; Vincent et al. 2013). The most notorious variant is rs6025 (i.e. Factor V Leiden) which reduces the binding capacity of the anticoagulant *PROC* and leads to a hypercoagulable state (see chapter 1.1.2). Several studies associated variants in *F5* with a higher risk of VTE and for this reason variants in this gene are often listed in the SNPs used for PRS calculation for thrombosis (Bertina et al. 1994; Klarin et al. 2017, 2019). The most recent PRS score for VTE uses a series of SNPs that are in the surrounding regions of F5

but map within other genes, such as *LINC00970* and *NR_135799* (Fig. 4.15, top; Klarin et al. 2019). With the TG Hi-C experiment I was able to reassign the SNPs occurring within these genes to *F5*; giving a more plausible biological explanation of the role of these SNPs in the PRS (Fig. 4.15). In fact, part of these variants seems to be in regulatory regions linked to *F5*, so it is possible to assume that their effect is to regulate the expression of this coagulation gene.

Similarly, it was possible to reassign a GWAS SNP to *MPIG6B* (Fig. 4.13 bottom). This gene encodes for a membrane receptor critical for myeloid commitment and platelet function (Ribas et al. 1999; de Vet, Aguado, and Duncan Campbell 2001; Newland et al. 2007). Indeed, variants in this gene are causative of a syndrome characterised by low platelet count (de Vet, Aguado, and Duncan Campbell 2001; Newland et al. 2007; Melhem et al. 2017). rs2269476 is a SNP reported to be associated with mean platelet volume and it maps within the *DDX39B* gene body (Fig. 4.15; William J. Astle et al. 2016). *DDX39B,* also known as *BAT1*, encodes for a protein involved in the splicing process and it has not been functionally associated with any platelet function so far (Peelman et al. 1995). The interactions defined in megakaryocytes with the TG Hi-C experiments show a connection between rs2269476 and *MPIG6B,* which provide evidence of a regulatory loop between the rs2269476 regulatory region and the cognate gene (i.e. *MPIG6B*). Also, in this case, the rs2269476 SNP is in a regulatory region that interacts with a gene known to play a role in a biological pathway more relevant to the trait in consideration.

Using the TG Hi-C results to inform the gene assignment of the 965 independent SNPs that have been linked to platelet traits (i.e. mean platelet volume, platelet count, plateletcrit and platelet distribution width) in the manuscript by Astle and colleagues (William J. Astle et al. 2016) I was able to re-assign 50 genes out of the 525 (i.e 9.5%). When considering that TG Hi-C only focuses on BTPD genes, this percentage is impressive. Moreover, the new and previous assignments are not mutually exclusive because enhancers can regulate multiple genes (Peter Hugo Lodewijk Krijger and de Laat 2016), so these data offer a complementary perspective that can be used to improve the biological interpretation of several SNPs.

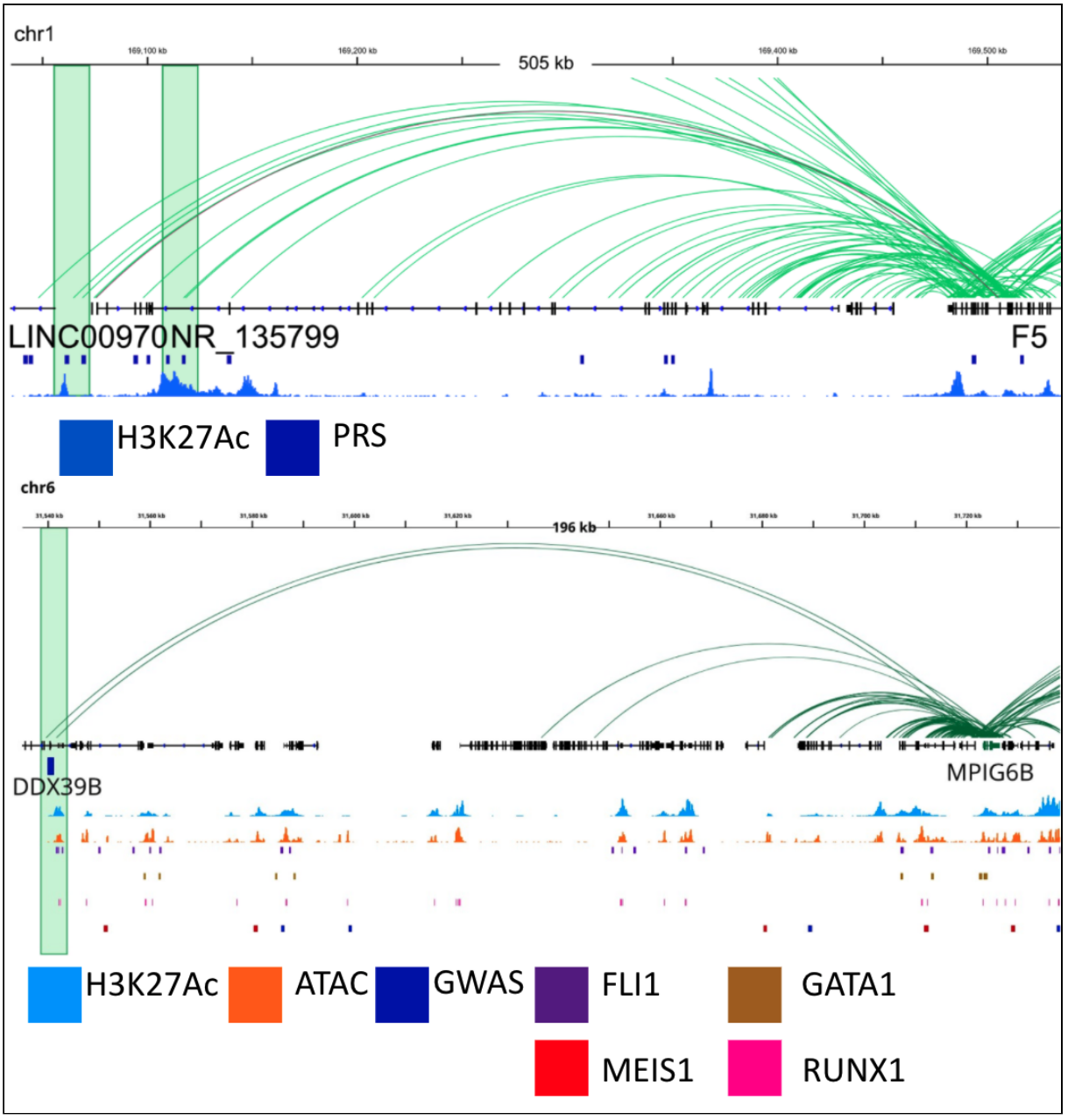


Fig. 4.15 | Reassignment of PRS and GWAS SNPs using TG Hi-C interactions. *The top and bottom panel are referring to the results obtained at the F5 and MPIG6B loci in HEP and MK, respectively. SNPs included in the calculation of the PRS for VTE or those found to be associated with platelet traits by GWAS are represented by short vertical dark-blue lines above and under the ATAC-Seq results track, respectively. Bright green arches are interactions identified in TG Hi-C experiments. The H3K27Ac peaks are referring to the results of ChIP-Seq experiments performed with a specific H3K27Ac antibody. Information about SNPs included in the PRS calculation have been taken from (Klarin et al. 2019) and for GWAS SNP for platelet traits have been retrieved from (Astle et al. 2016).*

4.10 Discussion

The experimental work described in this chapter successfully increases the resolution by which the regulatory elements have been identified for the 93 diagnostic grade BTPD genes. The TG Hi-C approach, limiting the number of genes in its libraries, has shown a 30-fold increase in the number of interactions observed per gene when compared to previous studies (Javierre et al. 2016) and it expanded the area of regulatory regions to a third of the size of the human exome (Niedringhaus et al. 2011). The lengths of the interactions are in line with the regulatory loops previously reported (Fig. 4.5;(Javierre et al. 2016; McCord, Kaplan, and Giorgetti 2020) and the genomic features annotations show a good representation for experimentally defined enhancers (Fig 4.6).

The number of interacting enhancers and the distance of the interactions to the gene body are positively correlated (R^2=0.47) with gene expression (Fig. 4.12). These observations are confirming the ones in the previous studies with a similar aim and structure (Javierre et al. 2016; Burren et al. 2017; Petersen et al. 2017). It would be interesting to implement other chromatin features (e.g. transcription factor binding sites) in the statistical model adopted and evaluate whether adding more information improves the capacity to predict gene expression levels (Shi, Fornes, and Wasserman 2019; Schmidt, Kern, and Schulz 2020). Schmidt and colleagues showed that the introduction of information on the TF binding in their machine learning model, in some cases, doubled the prediction accuracy of gene expression. Another study aims to integrate several different machine learning models to predict biological functions (Avsec et al. 2019).

Using the TG Hi-C interactions, It was possible to assign alternative genes to a series of SNPs that were identified in GWAS and used in the calculation of PRS for VTE (Fig. 4.13). PRS uses the genotype of an individual to predict the predisposition to certain traits or diseases, however, the molecular pathways underlying the PRS scores remain to be explored. For this to be achieved, the assignment of candidate genes to GWAS-associated SNPs must be improved (Klarin et al. 2019). The TG Hi-C experiments described in this thesis provides locus-specific examples of improved accuracy of candidate gene selection, with the *F5* and *MPIG6B* genes being far more plausible candidates for VTE and the formation of platelets by MKs. A careful inspection of these new data may even lead to the identification of alternative non-canonical pathways in VTE and megakaryopoiesis. However, the alternative associations to putative novel genes of unknown function require functional validation in the relevant cell and animal models before they can be considered for drug

development. For example, the Open Targets consortium (https://www.opentargets.org/) uses information from pcHi-C to assign variants to genes, in a cell-type specific manner, and inform their functional studies (Javierre et al. 2016; V2G pipeline, Open Targets).

Another observation from this study which merits further exploration is the dynamics of CTCF-DNA associations. The interactions observed in this study preferentially capture the weaker interactions of CTCF with its target regions; this could be a reflection of differences in the residence times between the two roles of CTCF (i.e. regulator of transcription and structural protein; Hansen et al. 2017; Agarwal et al. 2017). It is assumed that the reduced DNA residence time appears as weaker peaks, hence the enrichment for CTCF peaks scoring lower in the ranked signal (Fig. 4.7). Interestingly, a recent study suggested a similar behaviour for CTCF based on the degradation time of the CTCF (Luan et al. 2021).

Overall, the combined use of iPSC-derived differentiated cells and TG Hi-C allowed: i) for thousands of regulatory loops to be identified for the 93 BTPD genes and ii) for considerable differences in the regulatory landscape of genes between MK, EC and HEP cells to be mapped out. This new resource is of considerable value to drive forward research in haemostasis. First the detailed regulatory landscape of the core genes of the haemostasis canonical pathways provides an opportunity to explore to which extent rare variants in regulatory elements of these genes contribute to the number 1 killer in society, being thrombotic events in the venous (VTE) and arterial circulation (i.e. coronary artery disease, heart attacks, peripheral artery disease and thrombotic stroke). Second, a new genomic space has been defined which can be explored in cases of unexplained BTPDS (Downes et al. 2019; Turro et al. 2020b). Finally, by adding functional validation of the regulatory elements defined in this experiment, an additional quantitative attribute can be added for use in statistical analysis of whole genome sequencing data. However, the regulatory space of the 93 BTPD gene is too large (Table 4.1) to be included in association tests, like BeviMed, as putative association signals would escape identification because of multiple testing (Greene et al. 2017). Further analysis of the data is required to constrain the search space for rare non-coding variants causal of unexplained BTPDs, i.e. by applying the RedPop analysis approach to identify regulatory elements occupied by transcription factors (Turro et al. 2020b). These next steps in the analysis of the data generated in the experiments described in this chapter is discussed in chapter 5.

Chapter 5

The RONDA study: the Role Of Non-coding DNA in BTPD Aetiology

5.1 Introduction and aims of the chapter

HTS is the preferred method in screening cases with assumed inherited disorders. This technology allows to screen multiple loci, or the entire genome, in a single experiment, and for this, it advanced the understanding of the aetiology of several rare diseases, including BTPD ones (Albers et al. 2011, 2012; Cvejic et al. 2013; Westbury et al. 2015; Stritt, Nurden, Favier, et al. 2016; Turro et al. 2016; Stritt, Nurden, Turro, et al. 2016; Bariana et al. 2017; Lentaigne et al. 2019; Sivapalaratnam, Collins, and Gomez 2017; Sims et al. 2020; Turro et al. 2020b). HTS also improved the diagnostic rate of these rare diseases (Turro et al. 2020b; Taliun et al. 2021; Downes et al. 2019). Currently, the HTS diagnostic yield for BTPD cases with gene panel test is still approximately 50% if there is a strong prior belief of the condition being a rare inherited one (Simeoni et al. 2016; Downes et al. 2019). Notwithstanding the success of HTS in increasing the diagnostic yield, for the other half of cases, no molecular diagnosis has been reached.

There are several possible reasons why no genetic cause could be identified. First, the assumption that the BTPD condition was a monogenic inherited disorder may have been erroneous. Second, the condition is caused by a rare variant in a novel BTPD gene not yet identified. Based on the results of the pilot phase of the 100,000 genomes project, it is entirely plausible that a large number of BTPD genes remain to be identified (Lentaigne et al. 2016; Freson and Turro 2017). Indeed, over the past decade 18 new BTPD genes have been identified and more recently a rare variant in *MAST2* was identified as a novel cause of premature thrombotic events (Lentaigne et al. 2019; Morange et al. 2021). The identification of *MAST2* as a key regulator of the haemostasis pathways is a striking example of how studies of unexplained rare BTPD can lead to novel biological insights. Third, SVs cannot be identified with good sensitivity in short read HTS genotyping methods, whether it is WES or WGS. Therefore their role in disease aetiology is underestimated, although the sensitivity of the detection of deletions has improved with the introduction of WGS for rare diseases diagnosis (see chapter 1.3.6; Neerman et al. 2019; Mahmoud et al. 2019). This lack in sensitivity for the detection of SVs may in part be corrected by the application of long-read sequencing. For example, SVs causal of antithrombin deficiency, which resisted detection by a plethora of genotyping platforms, could be identified by analysis of DNA by Oxford Nanopore long read sequencing (de la Morena-Barrio et al. 2020). Lastly, in a fraction of cases, the aetiological variants may be in non-coding regulatory regions, either of a known BTPD gene or genes yet to be identified. Indeed, rare variants in the regulatory elements of the haemoglobin locus which are causal of the different forms of thalassemia are one of the

most studied examples of a rare disease causing dysregulation of gene expression (Higgs 2013; Hay et al. 2016). For instance, loss of CTCF binding sites has been shown to cause rare diseases in humans, such as polydactyly and branchio-oculo facial syndrome (Higgs 2013; Lupiáñez et al. 2015; Laugsch et al. 2019; Schoenfelder and Fraser 2019).

It is reasonable to argue that the role of regulatory variants as a cause of rare diseases may be under-appreciated because nearly all genetic investigations over the past 50 years have focussed on the coding space. Possibly, the most well-known example of haematological enhanceropathies is linked to the globin regulatory regions (Hay et al. 2016). There is an increasing number of examples of rare diseases caused by regulatory variants have been reported also in the haematological domain (De Gobbi et al. 2006; Phylipsen et al. 2010; Nicchia et al. 2016; Liang et al. 2020; Turro et al. 2020b; Thaventhiran et al. 2020).

In chapter 5, I describe the work I performed to prioritise, according to their regulatory potential, the regions defined in the previous chapter and assess the potential role of rare variants in these regions in the aetiology of some forms of BTPD. The variants were identified by whole genome sequencing of 13,037 DNA samples from rare disease patients and their close relatives of the NIHR BioResource cohort (Turro et al. 2020b). For the purpose of this study the 1,169 patients presenting a BTPD phenotype were defined as cases, whilst the remaining 11,868 participants of the cohort were used as reference (Fig. 5.1; Turro et al. 2020b). The Bayesian statistical framework, BeviMed, was applied to regulatory regions of the 93 BTPD genes to estimate the probability of a rare variant in a regulatory element being causal of a BTPD (Greene et al. 2017; Turro et al. 2020b). The BeviMed method models the different modes of inheritance and infers the posterior probability of a rare variant being causal. In a next analysis, the likelihood of rare variants in TG Hi-C regions altering the binding for CTCF was modelled, with the notion that non-coding variants can alter the binding site of this structural DNA-binding factor. Finally, two *in vitro* experimental methods were used to measure the functional effect of a set of selected rare variants on gene expression. The first method was a dual-luciferase reporter assay and the second one a CRISPR interference approach. The dual-luciferase reporter assay measures the effect of the rare DNA variant on transcription (Sherf et al. 1996; Grentzmann et al. 1998). The dCas9-KRAB system silences the entire region being targeted and therefore assesses the role of the region in a physiological context (Fulco et al. 2016). The combination of the two approaches provided a framework to test the role of specific rare variants in the regions of interest (i.e. dual-reporter assay) and to recapitulate the effect of of its loss of function in the local chromatin landscape (i.e. dCas9-KRAB).

The results of the above described experiments can be used to gain insight in the possible causal role of rare variants in regulatory elements in BTPD disease aetiology. Throughout the chapter, a few BTPD cases will be reviewed where the above analysis has identified putative causal variants in regulatory elements. The work presented in this chapter is the result of a collaborative effort with former colleagues from the Cambridge laboratory and colleagues from the University of Exeter. In particular, the BeviMed analysis (Fig. 5.6) has been executed by Dr. Daniel Greene and Prof. Ernest Turro, currently affiliated with the Icahn School of Medicine at Mount Sinai. The variant effect on CTCF binding motifs (Fig. 5.7) has been performed by Dr. Nick Owens from the University of Exeter. The imMKCL cell line used for the reporter assay experiments was made available by Prof. Koji Eto from the University of Kyoto.

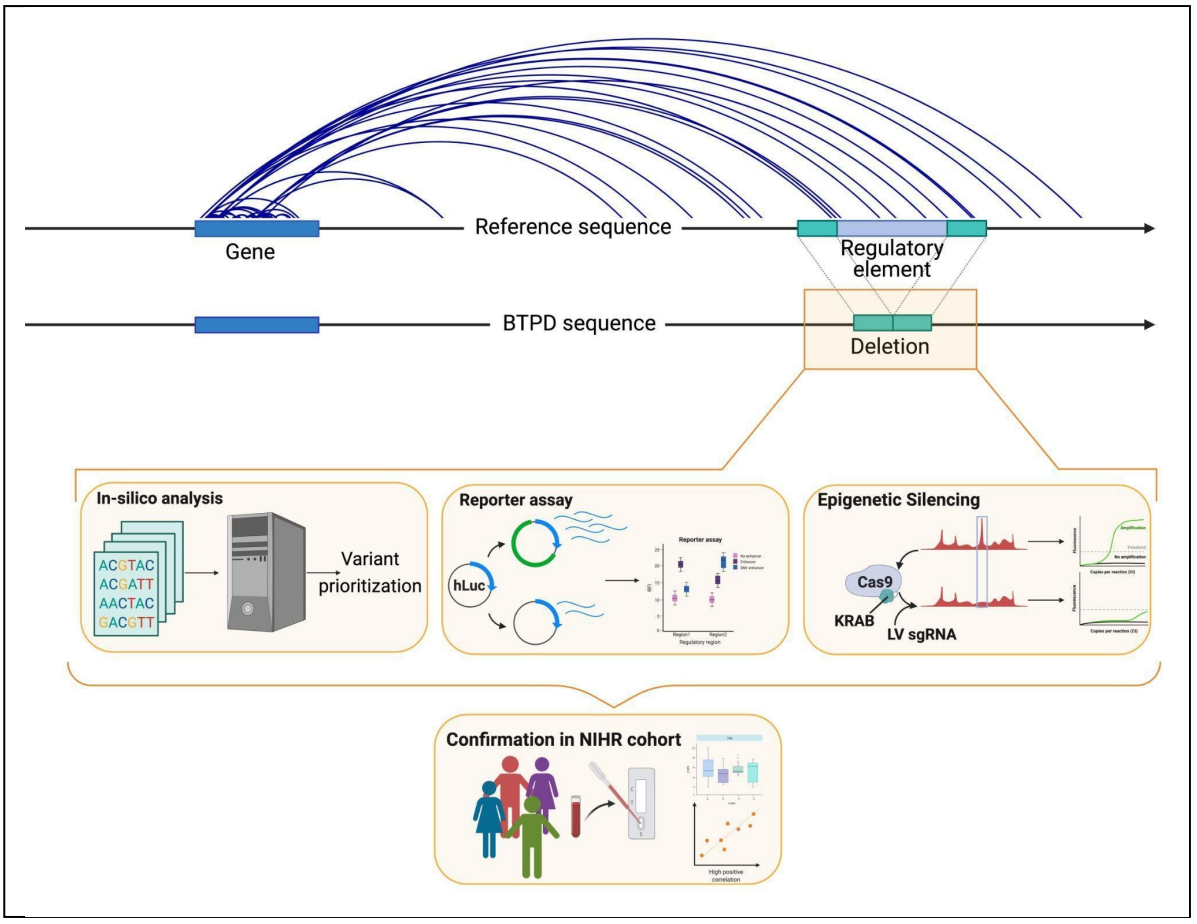


Fig. 5.1 | Cartoon describing the workflow adopted in this chapter and possible future experiments. *The first step requires the identification of rare variants in the TG Hi-C regulatory regions with a high likelihood of being functionally relevant for the rare BTPD conditions. Then the effect of these rare variants on gene function has been investigated with in silico and in vitro approaches. Ultimately, the rare non-coding variants with the highest probability of being pathogenic will be tested with functional studies in the individuals carrying these variants.*

5.2 The regions identified with TG Hi-C show potential regulatory capacity

The MK interactions defined in chapter 4 were ranked and prioritised based on their characteristics (e.g ATAC-Seq, H3K27Ac ChIP-Seq; see chapter 2.3 for a detailed description of the ranking algorithm). The ranking approach has been used to prioritise regions with a high probability of being functionally active MK regulatory elements and the seven highest-ranking ones have been cloned into a plasmid (Fig. 5.2.A and Fig. 5.2.C) and then their regulatory potential has been functionally validated with the reporter assay (Fig. 5.3).

The plasmids were transfected with the *bona fide* regulatory regions in a cell line that faithfully mimicked megakaryocyte biology, morphologically and functionally, namely imMKCL (Nakamura et al. 2014). imMKCLs grow in clumps of cells (Fig. 5.2.B) similarly to iMKs obtained by forward programming of iPSCs (Fig. 4.2). This molecular similarity is required so that the TFs and the transcription machinery are similar to those found in the primary MKs.

The regions to test with functional assays were selected to overlap H3K27Ac Chip-Seq peaks, CTCF Chip-Seq peaks and RedPop regions (Turro et al. 2020). The PCR fragments, which contain these regulatory regions, were cloned in the reporter plasmid respecting the 5'→3' directionality of the human genome (Fig. 5.2.C), and the correct insertion in the plasmid was confirmed by Sanger sequencing (Fig. 5.2.A).

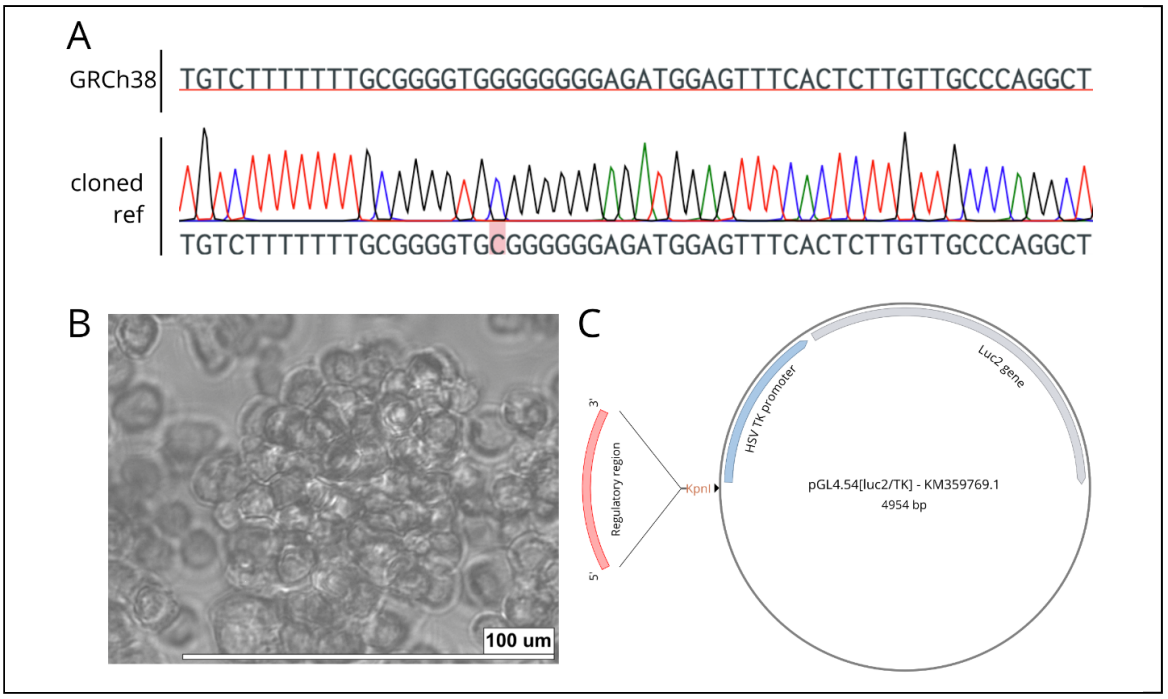


Fig. 5.2 | Reporter assay components and QC. *A) Example of a Sanger sequence performed to verify the correct insertion of the DNA sequence. B) Bright-field microscopy image showing the morphological characteristics of the imMKCL cell line used for the reporter assay experiments. C) Plasmid used for the reporter assay experiments. In grey, blue and pink are the reporter gene (either Luciferase or Renilla), the minimal promoter and the insertion site for the regulatory region, respectively.*

This reporter system showed that the seven highest scoring regions identified by the analysis of the TG Hi-C results significantly increase (p-value=1.4e-02 after Bonferroni correction) the expression of the reporter gene over the empty vector, supporting the original assumption of the regulatory function of these regions. Different regions show different levels of enhancing capacities, ranging from a two- to a 15-fold increase in the reporter gene expression (Fig. 5.3).

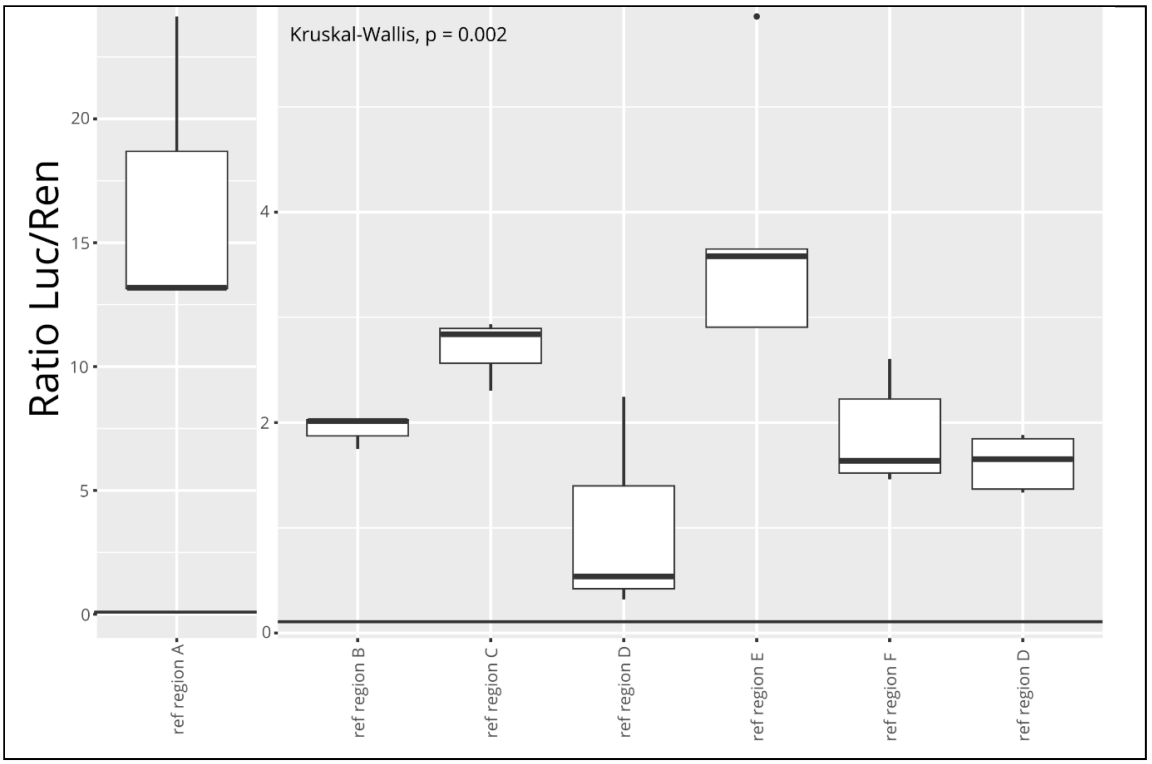


Fig. 5.3 | Effect of the selected TG Hi-C regions in the reporter assay. *Effect of the regions is reported as the expression of the reporter gene (i.e. Luciferase; Luc) normalised by the expression of Renilla (Ren). The horizontal black line is the Luc/Ren ratio for the empty vector.*

5.3 Identification of variants in TG Hi-C non-coding regions in participants of the NIHR BioResource Rare Diseases study

Non-coding variants occurring in the 2.5 Mb TG Hi-C regions identified in MK were selected from the NIHR BioResource Rare Diseases cohort and their effect on gene expression was estimated. It was postulated that deleterious non-coding variants may be better tolerated than deleterious coding ones and a relaxed AF threshold of ≤ 0.01 was set as filter for variant selection (see chapter2.3 for a description of the method). For instance, Albers and colleagues reported that the non-coding UTR variant rs139428292 (AF = 0.016) in *RBM8A*, in combination with a deletion of *RBM8A* on the alternate allele is the cause of the thrombocytopenia with absent radii (TAR) syndrome (Albers et al. 2012).

The initial search returned 63,529 variants in the MK interacting regions in the WGS genotype data of the BTPD cases (Fig. 5.4.A); 78% and 22% of the variants are SNPs, and INDELs, respectively (Fig. 5.4.A and Fig. 5.4.B). As expected, the vast majority of the single nucleotide variants were singletons (Fig. 5.4). The number of large deletions observed is in line with what one would expect from a recent study that aims to determine the number of these variants in a healthy population (Beyter et al. 2020).

If considering only SNPs and INDELs, the distribution of AFs for these variants observed in the TG Hi-C regions seemed lower than the AF in the remaining non-coding part of the genome. This observation is compatible with the notion that regulatory regions are under positive selective pressure (Fig. 5.4.C). To test this assumption, 25 rounds of random permutations of the genomic space were performed to select regions of equal length in bp to the ones defined in the TG capture Hi-C experiments. The number of variants in these ‘control’ regions were counted and their AF values calculated. Because these types of experiments are computationally demanding, the analysis was limited to the prey regions identified in chromosome 11. This chromosome was selected because it harboured the largest number of identified regulatory regions and the highest number of variants compared to the other chromosomes (Fig. 5.4.D). This analysis showed that the variants in TG captured Hi-C regions have a significantly lower AF, than the control regions confirming the hypothesis that the Hi-C identified regions are under purifying selection pressure and

therefore relatively devoid of rare variants compared to the other non-coding regions of the genome (p-value=2.2e-16, Kolmogorov-Smirnov test).

A

Type of variant	Number of observation
deletion	4824
insertion	9262
snp	49237
Large deletion	206

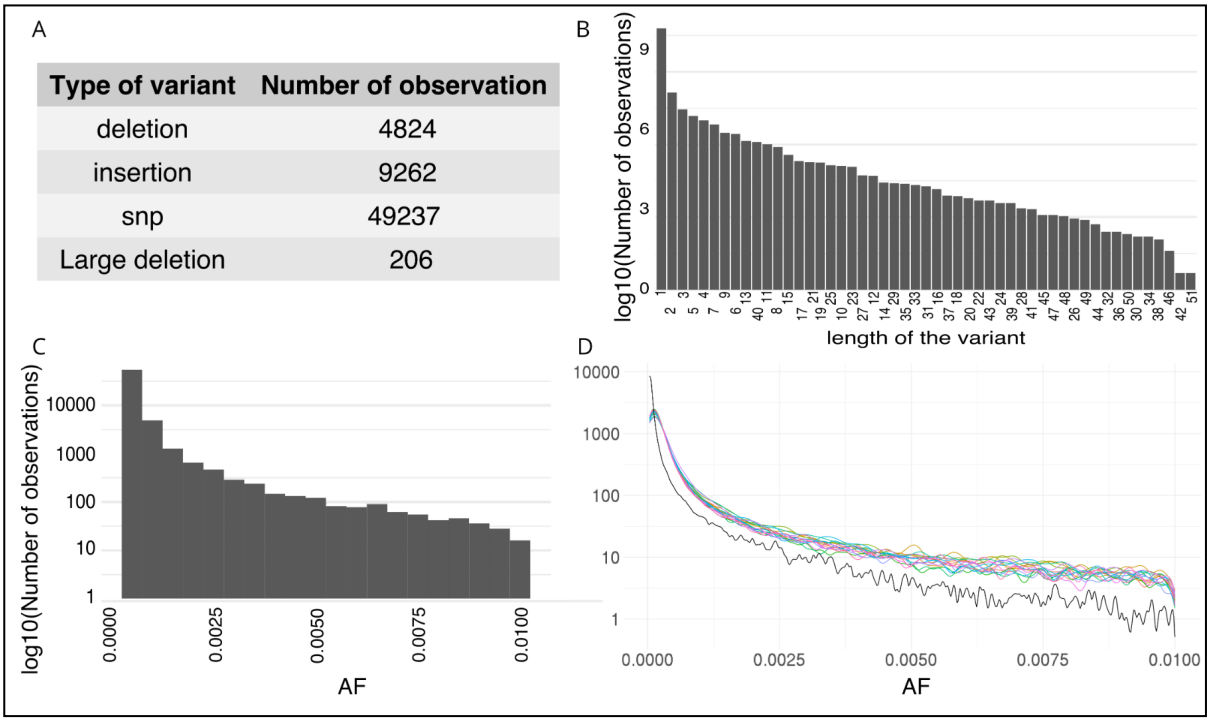


Fig. 5.4 | Variants from the NIHR BioResource Rare Diseases cohort overlapping TG Hi-C regions. *A) Summary statistics of the number of variants identified in the 2.5 Mb MKs regulatory regions. B) Number of observed variants categorised for the length of the variant in bp. C) Number of observed variants binned on the basis of their allele frequency. D) Number of observed variants in the TG Hi-C regions (black line) on chromosome 11 ranked according to their allele frequency and in randomly sampled control regions (n=25; multiple colours).*

There were only 206 large deletions, across all chromosomes, that overlapped the regulatory regions defined with the TG Hi-C experiments in MKs. These regions were manually reviewed by visual inspection of the sequence reads because applying statistical analysis methods on a small number of events may lead to biased results lacking robustness. One of the 206 SV overlapping the TG Hi-C defined regions localised in the *HDAC6-GATA1* locus. This interaction loop between *GATA1* and its enhancer, which is localised in one intron of *HDAC6,* was initially reported by (Fulco et al. 2016). More recently, the deletion of this regulatory circuit was reported to be the cause of a new syndrome of macrothrombocytopenia and autism-spectrum disorder in a young male patient (Turro et al. 2020b). Reassuringly, this variant was again identified in the analysis reported in this chapter using newly generated TG Hi-C data.

The visual inspection revealed another possible interesting SV, which warrants further investigation (Fig. 5.5). This variant is present in two siblings with an unexplained bleeding disorder. The sibs share a 51Kb heterozygous deletion that removes an entire copy

of *BCYRN1*, part of *EPCAM-DT* and a regulatory region, conserved across the three cell types (i.e. MK, HEP, EC) 500 Kb downstream the *MCFD2* gene. *BCYRN1*, also known as *BC200*, is a 200-nucleotide long non-coding RNA that is expressed predominantly in neurons and studied for its role in neurodegeneration and cancer (Mus, Hof, and Tiedge 2007; Su et al. 2020). Interestingly, several of the molecular pathways for dendrite formation are also relevant for platelet function (Padmakumar et al. 2019). The function of *EPCAM-DT* has not been identified to date and its transcript is lacking from MK (Grassi et al. 2020). Lastly, *MCFD2* is a BTPD diagnostic-grade gene, which plays a role in the F5 and F8 trafficking within cells, and rare P/LP variants in *MCFD2* are considered an important cause of of the combined deficiency of the coagulation factors V and VIII, which is an extremely rare disorder (Zhang et al. 2008). This variant ablates, in heterozygosity, 3 features, 2 non-coding RNA and a regulatory region, which could contribute to the phenotype observed in a pleiotropic way. Further studies in this pedigree are being planned to determine the segregation of the variant with the phenotype of bleeding.

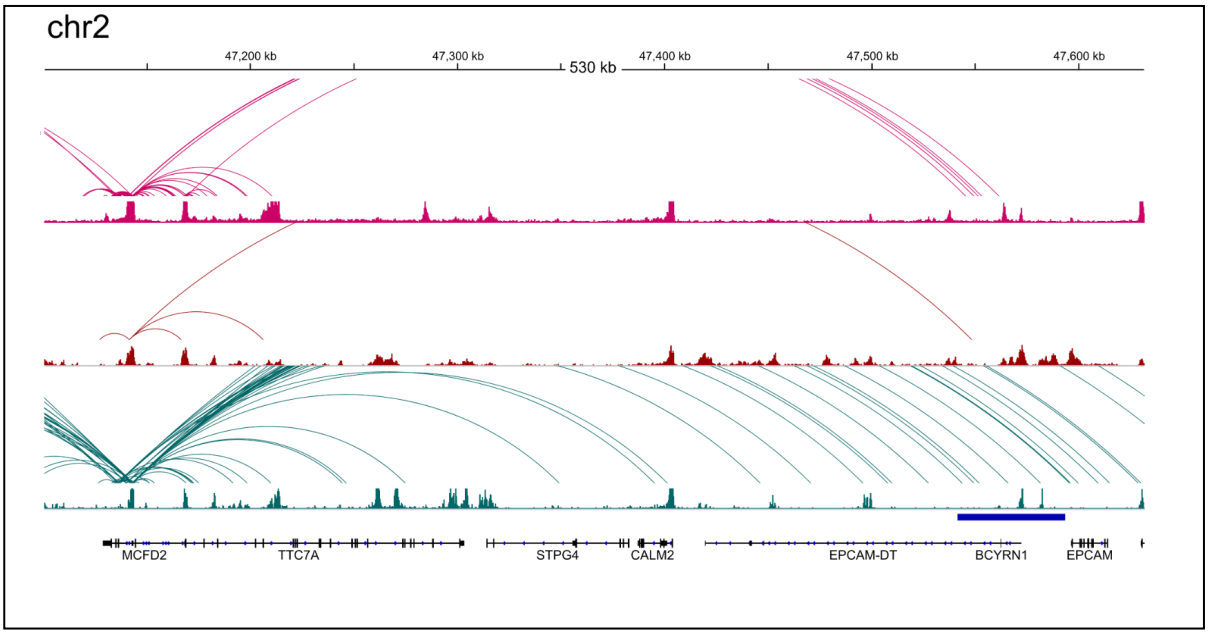


Fig. 5.5 | TG Hi-C interactions between MCFD2 promoter and a SV identified in the NIHR BioResource Rare Diseases. *MK (red), EC (brown) and HEP (green) interactions are all overlapping the deletion (blue horizontal bar) ~450Kb away.*

5.4 Statistical association to prioritise possible pathogenic variants in TG Hi-C

The power of the NIHR BioResource Rare Diseases study also resides in its 13,037 participants which provides reasonable power to detect rare non-coding variants associated with one of the rare diseases phenotypes in a constrained regulatory space. The Bayesian inference procedure BeviMed was applied to the WGS-generated genotype data of the 1,169 BTPD cases and 11,868 controls with the aim to identify putative causal regulatory loci and the variants therein associated with BTPDs phenotypes (Greene et al. 2017; Turro et al. 2020).

This analysis identified 61 associations at genetically independent loci with a posterior probability (PP) > 0.7 (Fig. 5.6). The associations were placed in three categories, being gene body to other gene body (n=13), gene body to the intergenic region (n=11) and within gene body (n=37). As expected, a large portion of the strongest associations identified occur over short distances that are mapping, because of their proximity, to the same gene (see chapter 2 for a better explanation of the method). The fact that the strongest associations are identified in the coding region or in regions in proximity to the coding ones gives more robustness to the results of this analysis. Indeed, one would expect that the majority of the disease-causing variants are coding. However, there are a few statistically significant associated variants in interactions between genes and intergenic regions (Fig. 5.6), for instance, *F7* (PP = 0.93). Although the expression of coagulation factors is very limited in platelets (log2(FPKM) for *F7* < 1), therefore it is difficult to interpret the biological effect of variants in these associations.

The first region with interactions outside the same gene body and high PP (i.e 0.98) is between the promoter of *BLOC1S6* and intron 2 of *GATM* (Fig. 5.6). The TG Hi-C experiment at this locus identified two independent strong interactions between the promoter and a distinct regulatory region 215Kb upstream (Fig. 5.6). Interestingly, the *GATM* region, that BeviMed associates to BTPD phenotypes, carries an H3K27Ac signal and is also a CTCF binding site.

A total of five BTPD patients have rare variants in these regions, with three having the same 9-bp insertion on one allele of chr15q21.1 g.45670316 G>AGACGCGCA (GRCh37). In the remaining two cases the following variants were observed in the same regulatory element. The first three cases were all characterised by an unexplained bleeding phenotype, which was accompanied by impaired platelet aggregation. BLOC1S6 encodes

pallidin one of the eight proteins of the ‘biogenesis of lysosome-related organelles complex-1 (BLOC1), and plays a role in lysosomal trafficking and possibly in platelet dense granules (Ciciotte et al. 2003; Li et al. 2003; Mao et al. 2017; Ambrosio and Di Pietro 2017). The gene is ubiquitously transcribed at high levels in haematopoiesis, including MK and platelets (log2-FPKM 6.4 and 9.4 respectively). P/LP variants in BLOC1S6, if present on both alleles are known to be causal of Hermansky-Pudlak syndrome-9, a rare multi-system disorder characterized by oculocutaneous albinism, bleeding diathesis and, in some cases, neutropenia with a proneness to pyogenic infections (Badolato et al. 2012). No rare coding variants with putative deleterious effects were identified in the BLOC1S6 gene of these cases. However, the observation that five patients with a bleeding phenotype have a rare variant in a regulatory element of BLOC1S6 warrants further studies on the presence of function-altering variants in interactors of pallidin in the BLOC1 complex and on the morphology and function of platelets of these cases.

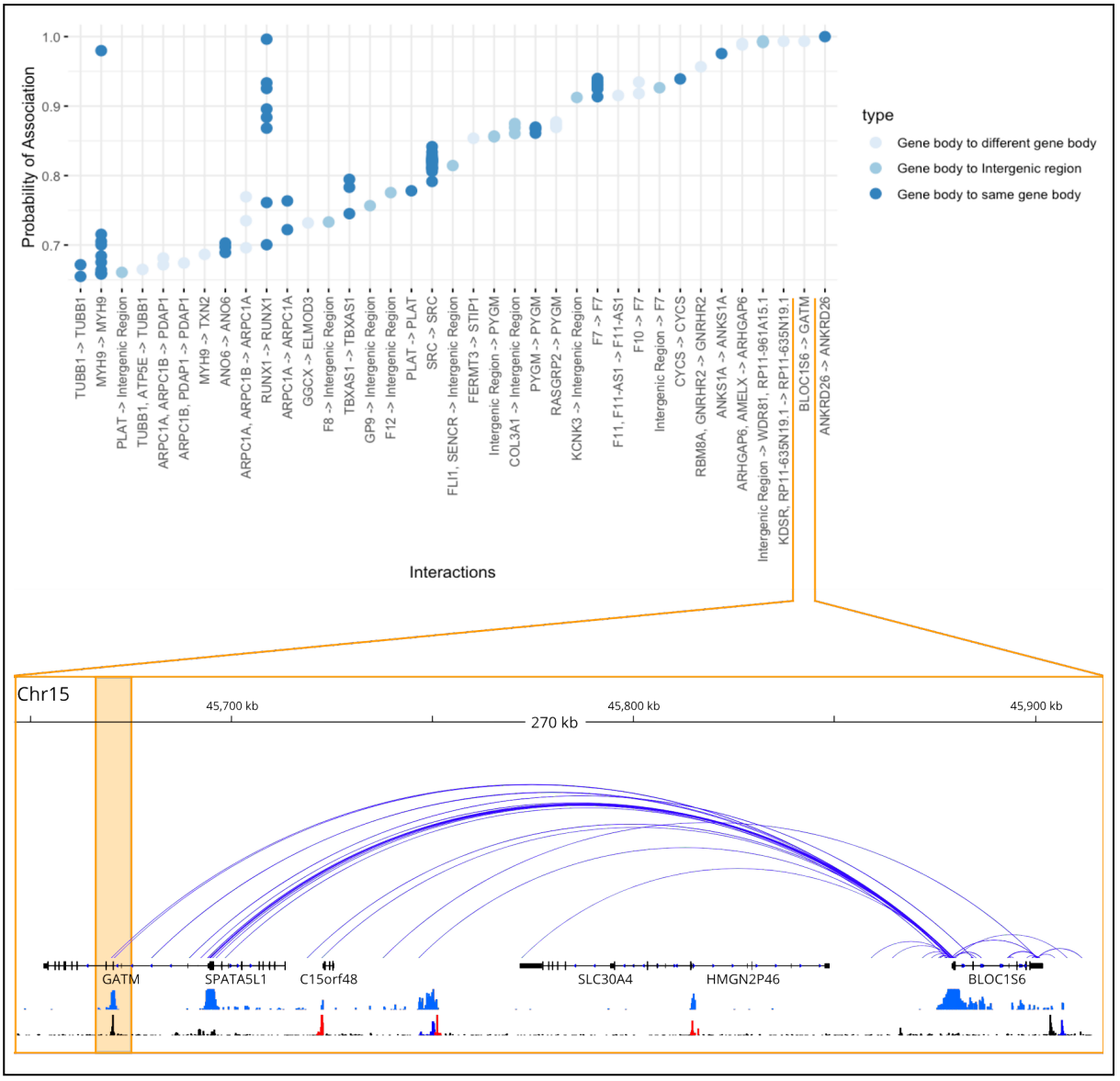


Fig. 5.6 | BeviMed data for variants in the TG Hi-C regions. *Upper panel: Blue dots refer to the posterior probability value returned by the BeviMed test. Locus names refer to the name of the gene which was used as prey. Lower panel: The TG Hi-C interactions for BLOC1S6 to its regulatory regions. The TG HiC interactions between the prey and the regulatory element are presented by blue arches. The orange shaded column at the left of the panel highlights the associated region identified by the BeviMed analysis. Results of ChIP-Seq experiments for H3K27Ac (light blue) and CTCF binding sites (red/blue) in MK are presented in the upper and lower tracks below the gene structures. For the CTCF track, red and blue indicate sense and antisense orientation of the CTCF motif, respectively. Black peaks in the lower track represent CTCF binding events in the absence of a canonical binding sequence motif for CTCF.*

5.5 *In silico* prediction of variant effects on transcription factor binding

An additional approach to investigate the possible pathogenicity role of non-coding variants is to infer their effect on the binding of transcription factors by analysing transcription factor binding motifs (TFBM; Fig. 5.7). The 63,323 variants (49,237 SNPs, 14,086 INDELs; see Fig. 5.4, A) were analysed for their effect on transcription factor binding by applying the Ensembl Regulatory Build method (Zerbino et al. 2015). In general, the effect of variants on the affinity of TF binding can be neutral or it may increase or reduce the avidity of TF binds binding to its TFBS. The distribution of the variant impact on the TFBMs shows a bell-shaped distribution (Fig. 5.7.B), slightly skewed towards the negative values (median density distribution = -0.03). The shape of this distribution suggests that the rare variants overall are more prone to reduce TF binding but the effect for most variants is very low. However, it is suggested that the impact of a few variants may be substantial (Fig. 5.7.B).

The effect on TF binding was ranked (Fig. 5.7A) according to the count of the affected motifs and the binding of TFAP2C and MAX was the most frequently altered, 269 unique rare variants across the 37,176 interactions in the BTPD loci (Fig. 5.7.A). MAX encodes for the MYC-associated factor X and in human MKs has been identified as one of the target genes of the key TFs for megakaryopoiesis (i.e. FLI1, GATA1/2, RUNX1 and TAL1). In fact, morpholino-mediated repression of the transcription of max resulted in a profound reduction of thrombocyte formation in zebrafish (Tijssen et al. 2011).

A

Transcription factor	number of transcription factors affected from variants
TFAP2C::MAX	269
GCM1::MAX	234
TEAD4::MAX	145
MYBL1::MAX	143
TEAD4::RFX5	114
HOXB2::ELF1	105
TEAD4::FOXI1	101
E2F1::ELK1	91
HNF4A	89
TEAD4::PITX1	86
CTCF	71
ETV2::RFX5	68
CLOCK::EVX1	63
TEAD4::CLOCK	60
TFAP2C::ELK1	59

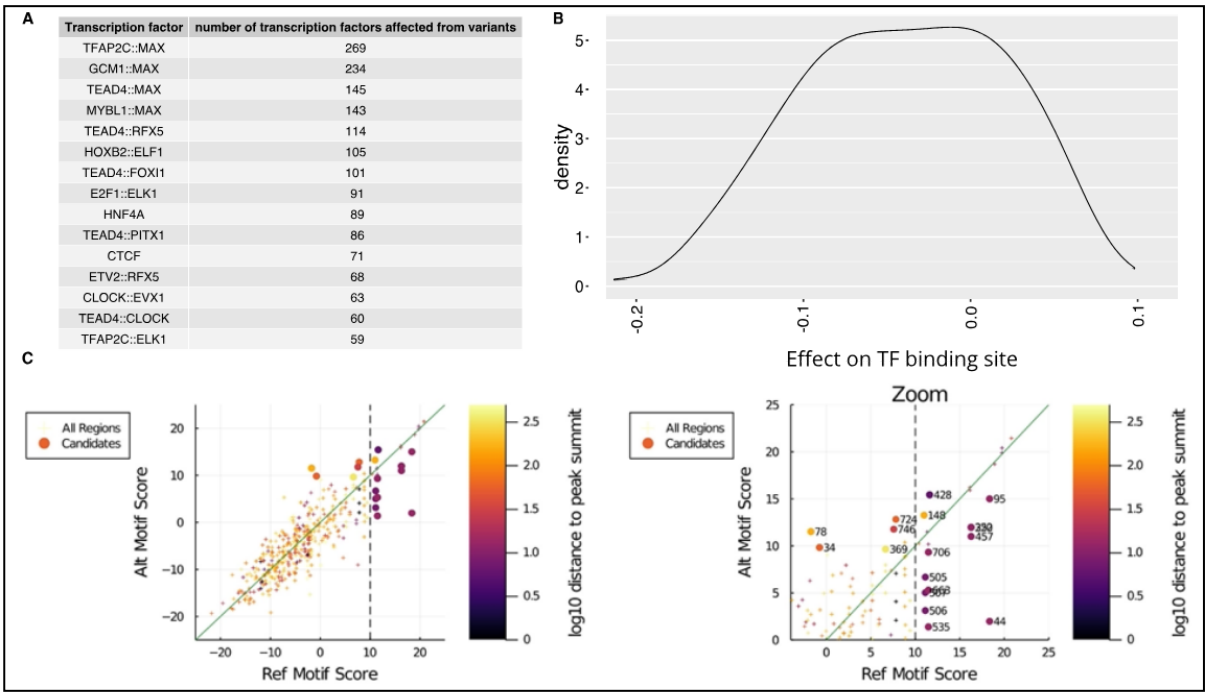


Fig. 5.7 | Effect of the rare variants in the TG Hi-C regions on the TFBS motifs. *A) Table listing the top TFBS affected by the variants identified in this analysis. The "::" indicates a complex of two transcription factors. B) predicted effect of the variant on the TFBS, ranging from negative to positive values to reflect a relative reduction and increase in the affinity of the TF binding. C) effect of the rare variants on the binding motif for CTCF. The left panel reports all the calculated CTCF motifs, the majority of the variants are not localised in a CTCF binding site (i.e. variants with values < 0). The shades of purple in the dots show the relative distance of the variant from the centre of the CTCF peak identified with the ChIP-Seq experiments in MK. The right panel in C is a zoom in on the variants that show alteration on the CTCF binding motif. The numbers on the right panels are the region IDs used in this experiment. Motif score = 10 (vertical dotted line) is the threshold adopted to reduce the number of false positives.*

Another interesting observation is the suggested effect of 89 rare variants on the binding site of the TFs HNF4A and RFX5 (Fig. 5.7.A). The former encodes hepatocyte nuclear factor 4-alpha. This transcription factor plays an important role in the development of the pancreas, kidney and liver. In adulthood, this transcription factor regulates the expression of several hepatic genes, including several coagulation factors (Inoue et al. 2006; see chapter 1.1.2). The role of RFX5 in the immune system has been well defined and variants in this locus are causal of severe combined immunodeficiency caused by a lack of expression of the HLA class II genes (Steimle et al. 1995). RFX5 is ubiquitously transcribed in haematopoiesis and whether it also plays a role in megakaryopoiesis and platelet formation remains to be investigated.

One of the other TFBS frequently affected by the rare variants identified in the NIHR BioResource cohort is the one for CTCF (Fig. 5.7,A). CTCF is expressed in all human cell types and has an important role in maintaining the structure of the DNA chromatin (see

chapter 1.3.5). The analysis of CTCF binding sites (Fig. 5.7,C) recapitulates the observation of the results for all the TF (Fig. 5.7.B) with 91% (n=781) having no discernable effect on CTCF binding. However, the results of the analysis suggest that the remaining ~70 (9%) variants may alter the binding of CTCF significantly (Fig. 5.5,A and Fig. 5.5,C).

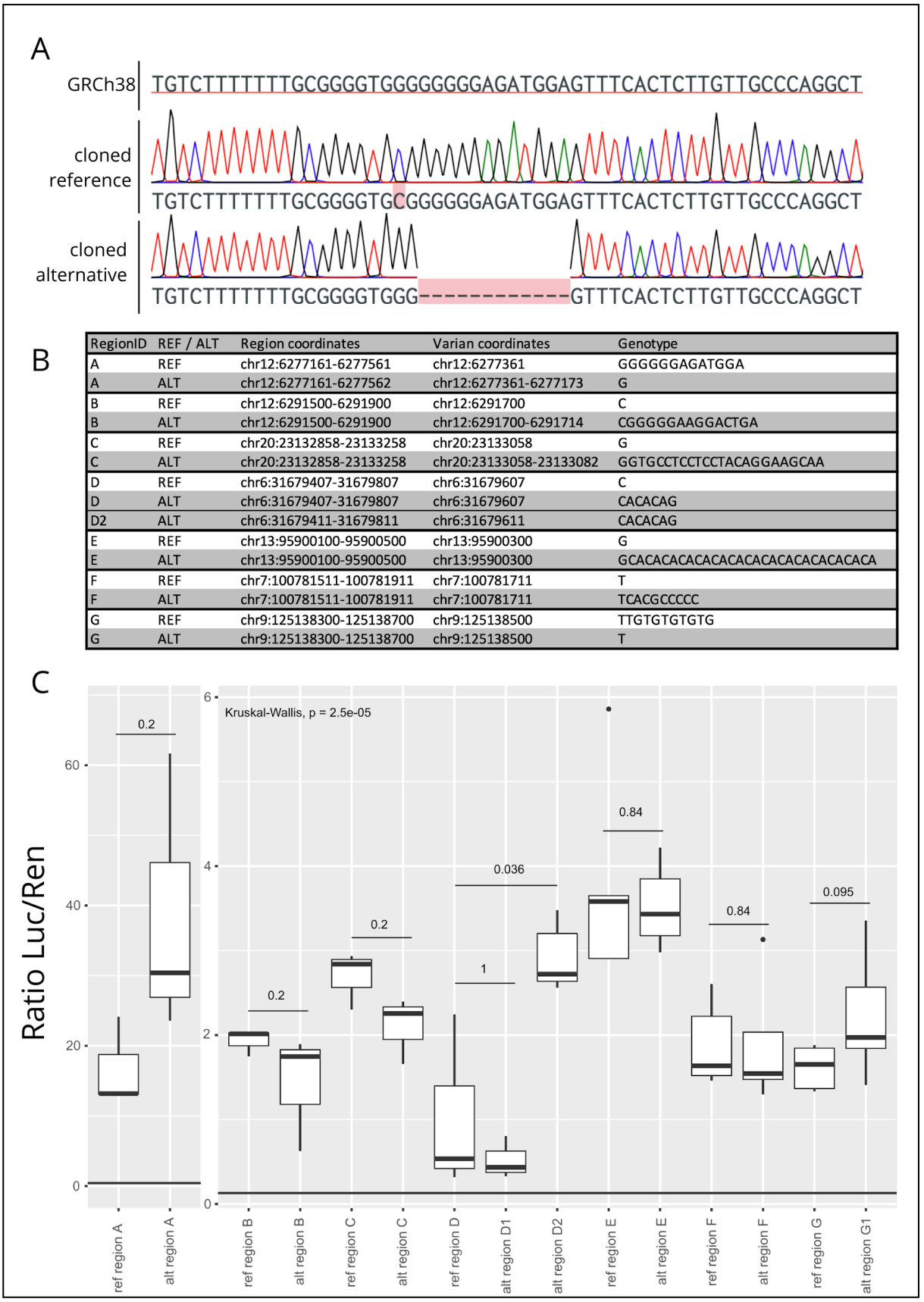
One of these rare variants with a significant negative effect is variant “44” in a CTCF binding site localised in the VWF locus in MK (chr12:6040204:AG:A GRCh38; gnomAD AF = 6.45828e-05). Three BTPD cases in the NIHR BioResource (Fig. 5.7,C right panel). The deletion of the G nucleotide is predicted to dramatically reduce the binding of CTCF. Whether this predicted reduction in CTCF binding at this position has consequences for gene regulation requires further experiments.

5.6 Functional validation of rare variants in TG Hi-C regions

The reporter gene experimental protocols described earlier in this chapter (Fig. 5.3) were applied to measure the effect of eight rare INDELs identified in the NIHR Rare Diseases Cohort on gene regulation. However, on this occasion, the regions of interest were retrieved from the DNA of the eight BioResource participants carrying these rare variants. The eight rare variants resided in seven unique genomic regions (Fig. 5.8,B). The reference regions are the same as used before (Fig. 5.3) and the complexity and repetitiveness of the eight INDELs is shown in Fig. 5.8,B.

As already shown in earlier experiments (Fig. 5.3), all the seven regulatory regions, including the alternative sequences, have the ability to alter the expression of the reporter gene (p-value=2.5e-05; Fig. 5.8,C). However, for seven of the eight rare variants, the difference between the reference and alternate sequences were not statistically significant (Fig. 5.8,C). The only result which reached the significance threshold was between reference regions D and its alternate rare allele D2 (p-value = 0.036), with the rare allele increasing the expression of the reporter gene by approximately three folds. For some of the remaining rare variants, non-significant trends were observed with different directionalities. Alternate rare alleles A, D2 and G showed an increased expression and B, C, D1, E and F a reduced expression of the reporter genes.

After a power calculation (see chapter 2), the number of replicates was increased to 15 for the regulatory regions A and C. These independent replication experiments confirmed the earlier results for these regions (Fig. 5.9). The rare allele for region A (chr12:6277361:GGGGGGAGATGGA:G) increases the expression of the reporter gene by 2.5 folds (p-value=1.1e-0.5; Fig. 5.9); and the rare allele for region C (chr20:23133058:GGTGCCTCCTCCTCCTACAGGAAGCAA:G) reduced expression by 1.5 folds (p-value=7.6e-0.5; Fig. 5.9).



RegionID	REF / ALT	Region coordinates	Varian coordinates	Genotype
A	REF	chr12:6277161-6277561	chr12:6277361	GGGGGGAGATGGA
A	ALT	chr12:6277161-6277562	chr12:6277361-6277173	G
B	REF	chr12:6291500-6291900	chr12:6291700	C
B	ALT	chr12:6291500-6291900	chr12:6291700-6291714	CGGGGGAAGGACTGA
C	REF	chr20:23132858-23133258	chr20:23133058	G
C	ALT	chr20:23132858-23133258	chr20:23133058-23133082	GGTGCCTCCTCCTACAGGAAGCAA
D	REF	chr6:31679407-31679807	chr6:31679607	C
D	ALT	chr6:31679407-31679807	chr6:31679607	CACACAG
D2	ALT	chr6:31679411-31679811	chr6:31679611	CACACAG
E	REF	chr13:95900100-95900500	chr13:95900300	G
E	ALT	chr13:95900100-95900500	chr13:95900300	GCACACACACACACACACACACACACACACA
F	REF	chr7:100781511-100781911	chr7:100781711	T
F	ALT	chr7:100781511-100781911	chr7:100781711	TCACGCCCCC
G	REF	chr9:125138300-125138700	chr9:125138500	TTGTGTGTGTG
G	ALT	chr9:125138300-125138700	chr9:125138500	T

Fig. 5.8 \| Reporter assay for the variants in the seven selected TG Hi-C regions. *A) Example of Sanger sequencing to confirm the introduction of the correct sequence. B) Table of the regulatory regions tested with the reporter assay. C) Effect of the rare variants in the TG regulatory regions*

compared to the reference sequence. The horizontal black line is the Luc/Ren ratio for the empty vector.

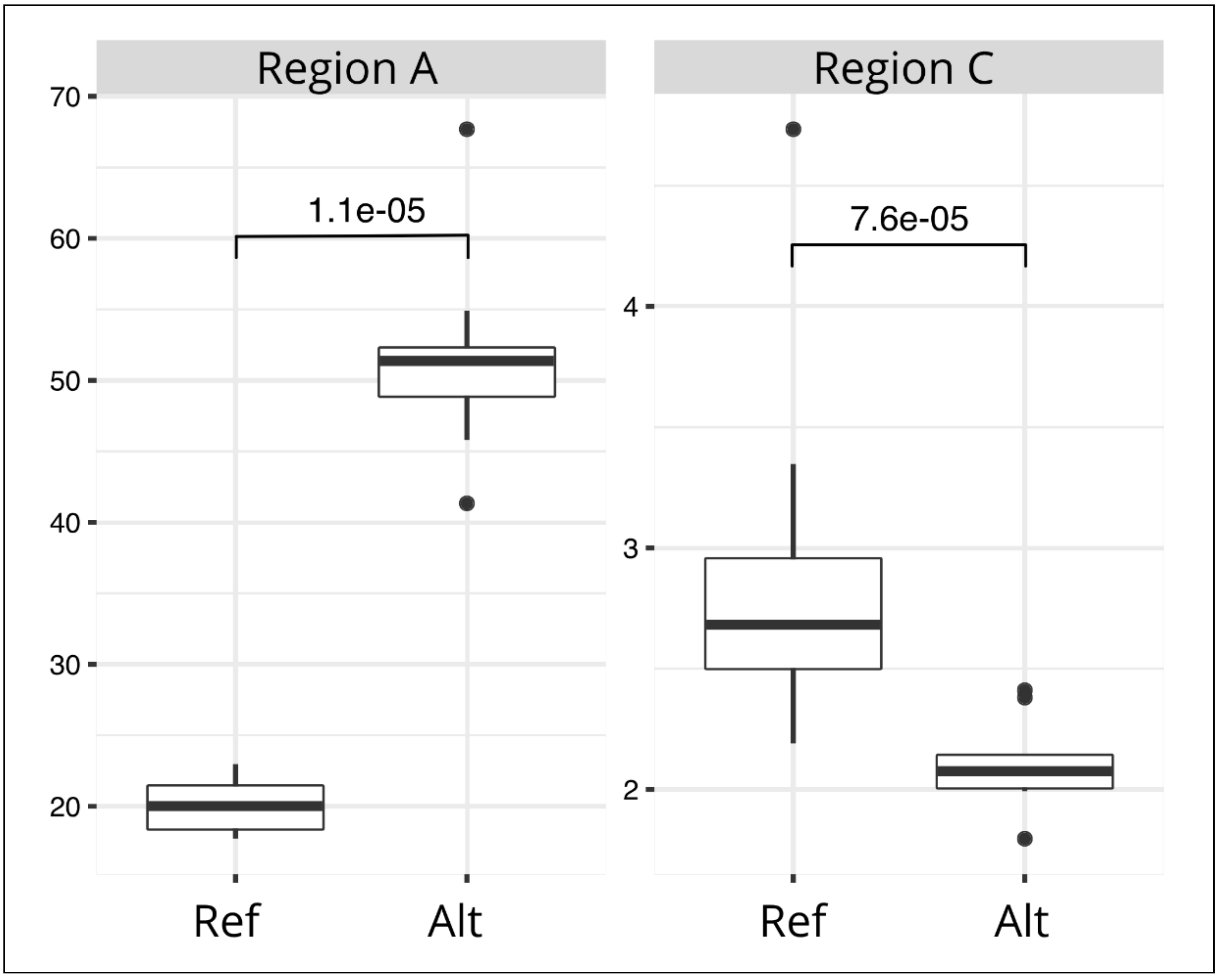


Fig. 5.9 | Replication of reporter assay for the rare variants in TG Hi-C regions *A and C. Same experiment as per Fig. 5.8,C but with an increased number of replicates, from 5 to 15.*

To investigate further the role of rare variants in TG Hi-C regulatory regions for BTPD genes the dCas-KRAB system was applied (Fulco et al. 2016; see chapter 2.4 for a detailed description of the method). The regulatory regions of *BLOC1S6* (Fig. 5.6)*, MCFD2* (Fig. 5.5) *and THBD* were investigated with this method. The dCas-KRAB system epigenetically silences the targeted regions via histone modifications. The KRAB domain decreases H3K27 acetylation and increases the tri-methylation at H3K9 (Table 1.6; Ying et al. 2015). To mimic the effect of germline LoF rare variants in these regulatory regions, repression was sustained, via the dCas-KRAB system, throughout the entire iMK differentiation protocol (see chapter 2.2).

The epigenetic silencing of the targeted regulatory regions should results in a significant reduction of the transcription of the cognate gene in comparison to the control gene (i.e. *GUSB*; Fig. 5.10). Indeed, the transcription levels of all cognate genes was

dramatically and significantly reduced, compared to the expression of the control gene (Fig. 5.10).

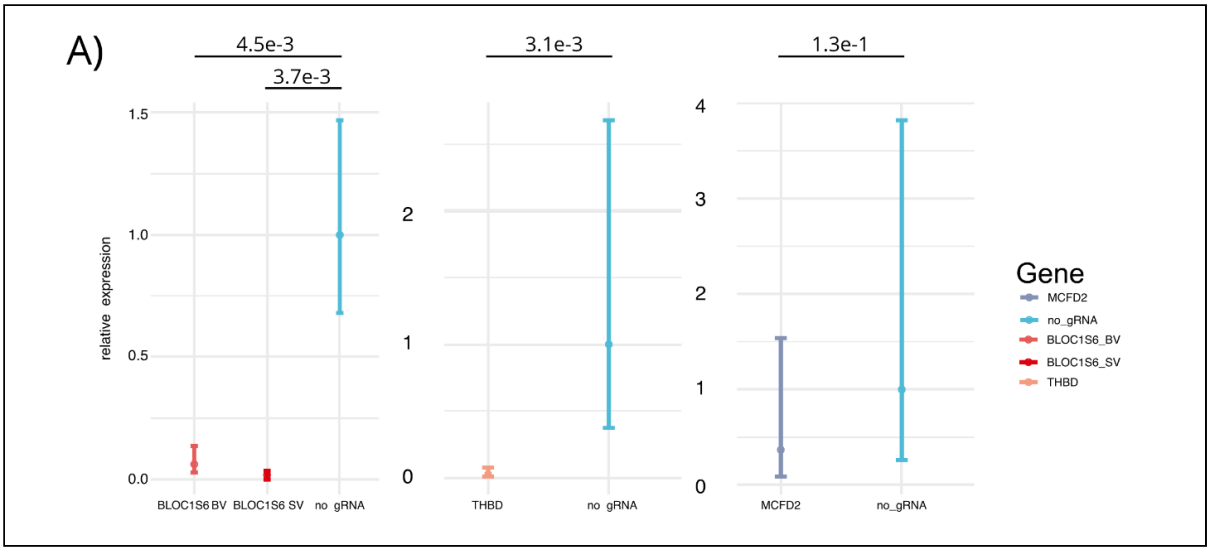


Fig. 5.10 | Expression levels of the cognate genes of the regions epigenetically silenced with the dCas-KRAB system. *p-values are calculated with the Wilcoxon test.*

However, the reduction in the transcript level was not specific for the cognate gene but affected all three genes, selected for the experiment, independently whether they were epigenetically silenced or not (Fig. 5.11). For instance, the experimental condition in which the regulatory region for the *MCDF2* gene was silenced also saw the reduction of the expression of *BLOC1S6* and *THBD*, but not the housekeeping genes used to normalise the transcription (Fig. 5.11, top panel). A similar pattern can also be observed for the condition used to repress the regulatory regions of *BLOC1S6* and *THBD*.

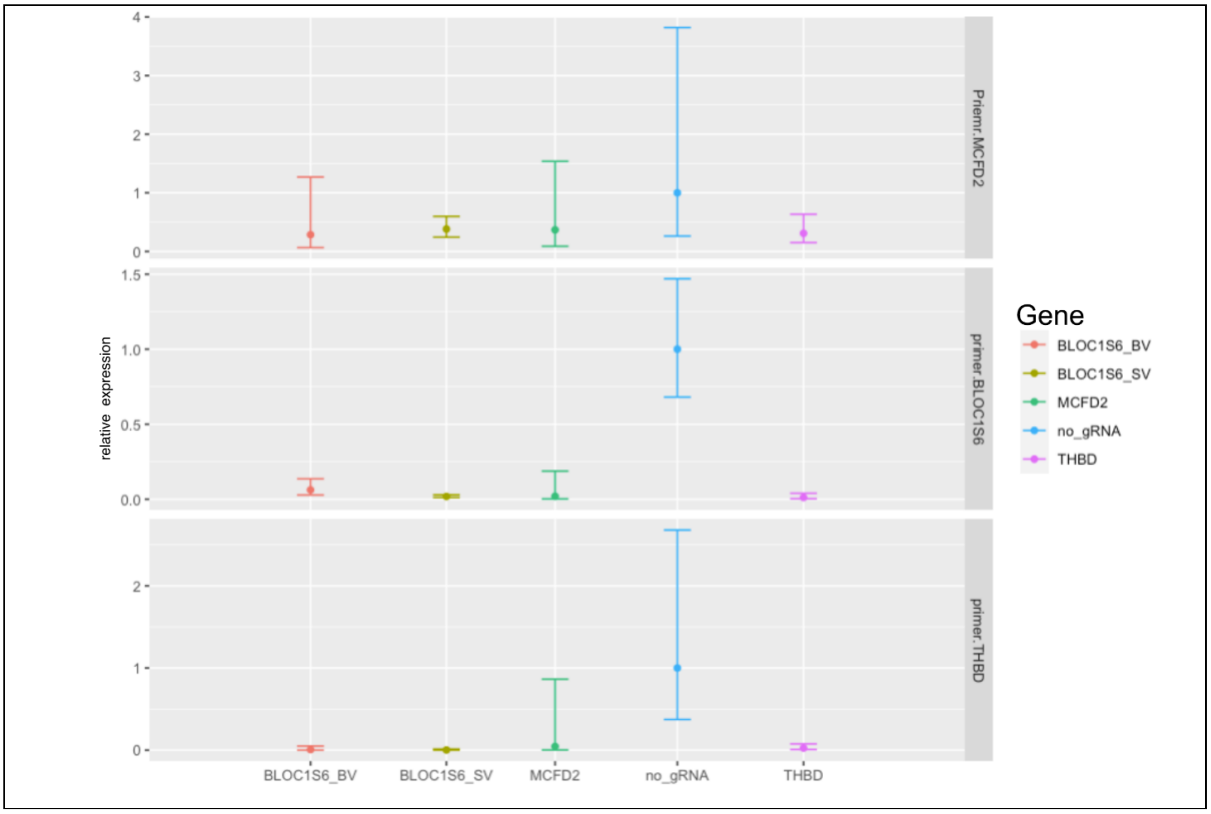


Fig. 5.11 | Transcript levels of the three test genes selected the dCas-KRAB mediated gene silencing experiment. *dCas-KRAB and gRNA against the regulatory regions of MCFD2 (top panel), BLOC1S6 (central panel), and THBD (bottom panel).*

5.7 Discussion

Decades of research in rare diseases led to a relatively good level of understanding of the relationship between rare coding non-synonymous DNA variants and their clinical sequelae (see chapter 3). In sharp contrast, there is very little knowledge on how the non-coding genome contributes to rare disease causality, with exception of a few loci (Lupiáñez et al. 2015; Schoenfelder and Fraser 2019; Higgs 2013; Liang et al. 2020; Turro et al. 2020). Thalassaemia remains the prototypic haematological inherited disorder that has regulatory nature. Although coding variants in the *HBA1*, *HBA2* and *HBB* genes can lead to thalassaemia, hundreds of different variants localised in the regulatory elements, particularly of the *HBB* gene can also be at the root of this inherited anaemia (Higgs 2013). In this chapter, I explored methods that can be applied to determine the functional role of regulatory regions and their likelihood of harbouring causal rare variants.

First, for a few selected putative regulatory regions, I showed that the sequences identified with the TG Hi-C approach have regulatory potential (Fig. 5.3), suggesting that TG Hi-C results can be used to expand the research of pathogenic variants for BTPDs. The regulatory role of these regions was also supported by the reduction in the number of variants observed when compared to reference random sampling (Fig. 5.4). Indeed, this relative depletion of rare variants has also been observed by others (Bejerano et al. 2004; Boffelli, Nobrega, and Rubin 2004).

Moreover, the reporter assay not only evaluated the regulatory capabilities of these regions but also assessed the effect on gene expression of 8 INDELs occurring within these sequences (Fig. 5.8 and Fig. 5.9). Unfortunately, the number of regions that could be tested was limited. The repetitiveness of the regulatory regions is incompatible with the DNA synthesis technologies (Hommelsheim et al. 2014), and the alternative approach which was used to overcome this challenge is far more laborious (i.e. PCR and cloning) and limited the number of regions that could have been tested to seven. The number of replicates needed to reach statistically significant outcomes, limited further the number of regions tested (Fig. 5.9). Possibly, high throughput approaches, such as massively parallel reporter assays (White 2015; Myint et al. 2019; Mulvey, Lagunas, and Dougherty 2020), could be used to obtain experimental validation of all the regions identified by the TG Hi-C experiment.

Second, a statistical approach was used to associate genotype to phenotype and to identify non-coding regions of interest, which are enriched for rare variants in cases versus

controls. For this, I applied the BeviMed method (Greene et al. 2017; Turro et al. 2020; Thaventhiran et al. 2020) with the intention of prioritising non-coding variants in the TG Hi-C defined regions (Fig. 5.6). This analysis showed 31 putative association signals with a PP>0.7 at 61 independent genetic loci. Further exploration of the associations at some of these loci is warranted, on a case-by-case basis, by a multidisciplinary team composed of BTPD experts.

Notwithstanding BeviMed test making a contribution to the prioritisation of interacting regions for further analysis, additional methods to determine the effect of rare variants on gene function could also make possible further contributions. The analysis of the variants in the TG Hi-C data has highlighted some possible functionally relevant findings for the transcription factors MAX and HNF4A and the possible loss of a CTCF binding site in the *VWF* locus (Fig. 5.7). The latter is a promising result that may be worth following up with *in vitro* investigation. In fact, others have shown that loss of CTCF binding sites can lead to syndromic conditions (Gregor et al. 2013; Konrad et al. 2019).

Finally, the dCas-KRAB method was used to investigate the role of the TG Hi-C regulatory regions in a more physiological system. In contrast to the reporter assay, this molecular tool maintains the regulatory regions embedded in their chromatin landscape (Thakore et al. 2015; Fulco et al. 2016), avoiding artefacts due to the episomal expression (Bevacqua et al. 2021). Initial analysis showed that the dCas-KRAB did in iMK achieve effective silencing of the cognate gene interacting with the targeted regulatory (Fig. 5.10). However, testing the expression of all the three cognate genes showed that this silencing was not specific (Fig. 5.11). This broad inhibitory effect can have several causes. First, the genes may have a shared regulatory mechanism (Chepelev et al. 2012), but this is unlikely because the genes reside on different chromosomes and the sharing of the same regulatory circuitry is not to be expected (Newton and Wernisch 2015; Cairns et al. 2016; Horta et al. 2018; Delaneau et al. 2019). Second, the co-repression may be an experimental artefact that has resulted from the extended period of the dCas-KRAB activity during the 20-day-long iMK forward programming. Further investigations are needed to determine the cause of the nonspecific inhibition of gene transcription caused by the introduction of the dCas-KRAB system.

The results presented in this chapter confirmed the regulatory potential of the regions identified in the TG Hi-C experiments and utilised in the experimental validation. The *in silico* analysis pinpointed several promising non-coding regions and variants therein which are deserving of additional investigation to evaluate their possible causality. It is hoped that

some of these variants and regions are confirmed to be pathogenic, thereby expanding the body of evidence that non-coding variants can contribute to inherited BTPD (Albers et al. 2012; Higgs 2013; Turro et al. 2020; Liang et al. 2020). It is also possible that some of the unexplained BTPD cases are caused by multiple coding and non-coding P/LP affecting one gene or multiple genes encoding functionally proximal neighbours in a canonical pathway relevant for the phenotype. Therefore, a multi-disciplinary team with BTPD expertise is an essential next step to review these unexplained cases carefully.

Chapter 6 Conclusion and future work

This final chapter aims to recapitulate, as a short summary, the main findings of my doctoral work and expand on the limitations of some of the approaches adopted in this research. Finally, it will conclude with remarks on how my research can be the springboard for future studies and how it can be applied to improve further our insights into the genetic architecture of unexplained inherited BTPDs.

6.1 Novel findings

Calculation of the effect sizes of rare variants for BTPD relevant phenotypes

In chapter 3, I calculated, for the first time, the effect sizes of rare P/LP variants reported to be causative of some BTPD phenotypes. The literature on such rare variants with relatively large effect sizes on phenotype has been rapidly expanding since the release of the UKB WES data (van Hout et al. 2020; Akbari et al. 2021; Goodrich et al. 2021; Forrest et al. 2021; Karczewski et al. 2021). Determining these effect sizes is fundamental as it will pave the way for the implementation of precision medicine into the clinics, where treatments and prevention are tailored for individuals based on their genotype. Particularly for rare diseases, having reliable estimates of effect sizes is clinically relevant to make the appropriate decisions about treatment (e.g. allogeneic stem cell transplantation for RUNX1-related thrombocytopenia), provide accurate prognostication, and support family planning decisions. Moreover, the results presented about the effect sizes of P/LP variants for BTPDs exemplify how the accuracy of the rare variant databases content, like ClinVar or HGMC, can be enhanced by systematic analysis of the genotype and phenotype associations in large population cohorts.

Besides, the estimates of effect sizes for P/LP variants, the statistical association analysis also revealed some novel findings. It is an assumption, in clinical genomics, that P/LP variants in carriers for disorders with an autosomal recessive mode of inheritance do not experience phenotypic consequences. However, case histories suggested that this assumption might have been incorrect. The analyses in this thesis provide further evidence to refute this assumption. First, carrying LoF P/LP variants in ITGB3 and ITGA2B (in homozygosity, these variants cause Glanzmann’s Thrombasthenia) leads to a modest reduction in platelet count. Second, the same phenomenon was observed in carriers of LoF P/LP variants in three of the four genes encoding the platelet receptor for von Willebrand Factor (GP1A, GP1B, GP9 but not GP5). Individuals with these LoF variants on both alleles

present with Bernard and Soulier syndrome, a severe form of macrothrombocytopaenia which is nearly always accompanied by bleeding. Finally, LoF variants in the *MPL* gene, when inherited on both alleles, cause early-onset chronic amegakaryocytic thrombocytopenia (van den Oudenrijn et al. 2000; Tijssen et al. 2008; Fox et al. 2009), which is an extremely rare and pre-leukaemic condition and warrants allogeneic stem cell transplantation. Surprisingly, carriers of these LoF variants have an increase in their platelet count compared to the controls. The mechanisms underlying this over-compensation remain to be determined. It might be possible that there is a regulatory circuit that leads to an excess in the transcription of the 'healthy' MPL reference allele.

Biological interpretation of PRS based on protein-protein interaction networks

I embedded the PRS for the four platelet traits (count, volume, crit and volume distribution width) and VTE in the context of protein-protein interactions (Fig. 3.18 and Fig. 3.20). The connection between large-effect PRS variants for platelet traits (e.g. $\beta > 0.5$ sd) and core genes for relevant traits were already observed based on gene co-expression (Vuckovic et al. 2020). The analysis of Vuckovic and colleagues confirmed the earlier observation (Gieger et al. 2011) that the core genes were strongly enriched for loci already known from rare inherited BTPDs and other Mendelian disorders. The research presented in this thesis is based on a protein-protein interaction network of 18,410 nodes (proteins) connected by 571,917 edges (interactions; Barrio-Hernandez et al. 2021; Schwartzentruber et al. 2021). In this thesis, the network analysis performed in platelet-PRS associated variants and the corresponding effect sizes also supports the omnigenic model proposed by Boyle and colleagues in 2017. In contrast, a similar analysis of the genes and proteins associated with the PRS for VTE did not align with the omnigenic model. There are several possible reasons for this observation: (i) the PRS for platelet traits are based on 2,648 genetically associated variants identified in a GWAS with 408,112 participants, whilst the one for VTE only uses 297 variants identified in a GWAS with only 26,066 cases of VTE. The lower number of associated variants for VTE versus platelet traits is in part explained by the GWAS for VTE being underpowered; (ii) the GWAS for full blood count parameters has highly accurate ascertainment of the quantitative traits whilst it is well documented that capturing of VTE in electronic health records is far from perfect; and (iii) as mentioned in chapter 4, mapping GWAS-associated variants to candidate genes by automated methods like Variant Effect Predictor is prone to substantial errors (Petersen et al. 2017, Fig. 4.15). This high error rate erodes the power for evaluating the omnigenic model, which is particularly problematic if the GWAS is underpowered, as is the case for VTE.

Contribution of the P/LP variants to VTE onset

A portion of cases with early-onset VTE have a monogenic cause, with P/LP variants in PROC, PROS1 and SERPINC1 being the main causal variants. Historically, ORs of such variants have been calculated based on relatively small population cohorts and tended to be enriched for cases being cared for at tertiary referral hospitals (Khan and Dickerman 2006; Konecny 2009; Kujovich 2011; Crous-Bou et al. 2016). The WES data from the UKB participants allowed for the first time to estimated the OR for a limited number of the P/LP variants in these three genes (Fig. 3.14 and Table 3.3). This showed that (i) the Burden test association, based on the aggregation of P/LP variant in one of these genes, confirmed that the OR for having a VTE event is increased by 2.76, 1.84 and 1.65 fold for PROC, PROS1 and SERPINC1, respectively; (ii) the P/LP variants in these three genes, that were present in enough carriers to perform single variant association analysis, returned ORs ranging between 17.42 and 4.22 (Table 3.3). These OR values are clinically relevant and are high in comparison to the ORs for other pathogenic variants, like those in cancer-causing genes such as *BRCA1* and *BRCA2.* In aggregate, UKB carriers of P/LP variants in either *BRCA1* or *BRCA2* have a cancer prevalence of 21.1% versus 6.6% for the non-carriers (OR 3.77; van Hout et al. 2020).

Altogether, my analysis estimated the ORs for P/LP variants implicated in venous thrombotic events for the first time. However, on average, these ORs are lower than what has been previously reported for VTE phenotypes in smaller population studies. These discrepancies in VTE ORs for individual variants can have several causes. First, most of the previous epidemiological studies have been relatively underpowered. Second, the association signals of these earlier studies may have been inflated by not taking into account the inflationary effect undetected population stratification can have on ORs. Third, the UKB may have been depleted for participants with severe early-onset pathologies, like recurring VTE at an early age, including those which lead to premature mortality. Finally, only a small portion of the known P/LP variants in the three prototypic VTE genes (*PROC*, *PROS1* and *SERPINC1*) could be analysed in this association study because 75% (~250,000) of P/LP variants have, so far, remained unobserved in UKB participants.

High-resolution chromatin structure maps for MK, EC, HEP; regulatory regions and associations with BTPD phenotypes

The technical workflow adopted to explore the regulatory regions of the BTPD genes in the three haemostasis-relevant cell types (i.e. TG capture Hi-C) allowed the generation of

high-resolution interaction maps. To the best of my knowledge, these maps are the highest resolution resource available so far for the regulatory elements of these 93 diagnostic-grade BTPD genes (Fig 4.4). Furthermore, the application of the BeviMed algorithm to the regulatory elements identified associations at posterior probabilities > 0.7 with BTPD phenotypes at 31 independent genetic loci. Several of these associations are worth further investigation to confirm or refute the rare variants in these associated elements being causal of diseases.

Furthermore, this thesis shows that a selected set of regulatory elements exert an effect on transcription in a reporter system, and in-depth investigations of two of these regions showed significant differences in transcript levels between the reference and the alternate sequence (Fig. 5.8 and Fig. 5.9). Ultimately, after some more validations (see chapter 6.3), this resource of regulatory elements could lead to the identification of novel mechanisms of BTPD pathogenicity and advise on the contribution of the non-coding regions to the onset of the Mendelian forms of BTPDs.

6.2 Limitations of the research

UKB cohort sampling may bias the effect sizes

All the population studies outcomes are affected by biases during cohort sampling (Heide-Jørgensen et al. 2018; Enzenbach et al. 2019). Despite being one of the least affected by this sampling bias, UKB is no exception (Swanson 2012; Fry et al. 2017). Indeed, this cohort has been depleted of severe clinical conditions (Fry et al. 2017). Therefore, the beta coefficients and ORs, including the one presented in this thesis, may be deflated because early-onset cases of VTE have not been enrolled in the cohort. I tried to control for this problem by excluding those variants with an insufficient number of carriers (i.e. < 15 participants). Still, the effect sizes of an unbiased population may be slightly different.

The limits of the Capture Hi-C resource

Developing the highest resolution interaction maps for the BTPD genes was feasible because of the small number of loci included in the experiment. This provides a limitation for the future because no information is available for novel BTPD genes, like the recently identified MAST2 gene and the LP variant therein which is deemed causal of early-onset VTE with an autosomal dominant mode of inheritance (Morange et al. 2021).

Technical limitations may hinder the identification of all the functions of regulatory elements

The strategy adopted in this study to functionally validate the regulatory regions (and the effect of variants in these regions) tested one region at a time. This approach may not be the optimal strategy to evaluate the function for all the regulatory regions. For instance, when multiple enhancers act synergistically (Junion et al. 2012; Hnisz et al. 2013; Shin et al. 2016) and only one component is tested, the effect of that component may be limited, and it is possible that the assay is not sensitive enough to detect the differences. Moreover, some elements of an enhancer may be required just to form the correct 3D structure of the regulatory region, while others may act as transcriptional regulators (Spivakov 2014). All these nuances of the enhancers may be missed in reporter assay approaches.

The BTPDs aetiology may lay in genomic features that have not been tested

The research workflow used in my doctoral work did not consider genomic features such as transposons, long non-coding RNAs and part of the complex structural variants, which are known to play a role in specific diseases (DiStefano 2018; Payer and Burns 2019). This is because these regions are difficult to map on the genome (i.e. transposons), or their biological function cannot be detected via physical interactions (i.e. long non-coding RNA) and is therefore missed by the TG Hi-C workflow. Furthermore, short-read sequencing used to genotype the NIHR BioResource participants has several limitations in detecting length-polymorphisms, particularly in repetitive sequence regions. Neither can copy number gain and complex structural variants being reliably identified. Other approaches, such as long-read sequencing complemented by RNA-sequencing of different types of blood cells, will provide further layers of information to define the role of these elements (Payer and Burns 2019; Cozzolino et al. 2021).

Technical limitations of the functional validation

The systems adopted to functionally validate the *bona fide* enhancers identified in chapter 4 have some limitations that are worthy of discussion. The reporter assay used nucleofection and ectopic expression from a plasmid. The copy number and the genetic context of the plasmid (e.g. histones and surrounding regions) are not resembling the physiological ones, which may introduce artefacts. Also, the dCas9-KRAB silenced regions, which in the case of enhancers, may act on multiple genes (Rosenbluh et al. 2017). This

approach could not discriminate whether the phenotypes were derived by silencing the tested gene of some other cis-regulated ones.

6.3 Future experiments

Are *MPL* findings also observed for other haematopoietic growth factor receptors?

The finding that carriers of LoF variants in *MPL* have an increased platelet count warrants further investigation. A first step, currently being undertaken, is to determine whether the observed effect can be replicated in the remaining UK Biobank participants (i.e. next release of WES data). However, it could be argued that a similar compensatory mechanism on the presence of LoF variants may also apply to other receptors for haematopoietic growth factors. P/LP variants of the LoF type, when present on both alleles of the *CSF3R* gene are causal of severe neonatal neutropenia, the CAMT-equivalent of the neutrophil lineage. Indeed, the variant rs138156467 introduces a premature stop codon in *CSF3R* and is causal of congenital neutropenia when present on both alleles (Klimiankou et al. 2015). In the carrier state, the variant increases the neutrophil count by 0.24 standard deviations (Vuckovic et al. 2020).

Recalling cases to validate the role of non-coding variants

The consent under which participants in the NIHR BioResource and the 100,000 Genomes Project have been engaged in the WGS study allows re-contact of participants with invites for follow-up studies. This study has identified rare variants in BTPD regulatory elements in several NIHR BioResource participants with unexplained BTPD phenotypes, and these variants may be causal of their inherited disorder. Inviting these participants and their close relatives is the first step to confirming or refuting these rare variants' possible causality. It also provides an opportunity for obtaining samples of blood of the proband and their relatives to perform additional cell biology and high-resolution imaging experiments (Turro et al. 2020; Thaventhiran et al. 2020). However, due to the COVID-19 pandemic, the recall of BTPD cases for these relevant variants had to be put on hold.

6.4 Direct application of this study to the clinical practice

Studies like this one, which aim to better understand the biology of diseases, can improve clinical practice in the foreseeable future. First, effect size and OR estimates for P/LP variants in BTPD genes will inform the reporting of their pathogenicity by the NHS Clinical Genomics Hub laboratories and beyond. Second, thrombotic events in venous circulation are among the leading causes of morbidity and mortality worldwide. It could be argued that low-frequency and rare variants in the regulatory elements of the BTPD genes may contribute to the high prevalence of VTE in the population at large. Nearly all BTPD genes are directly or indirectly implicated in the canonical pathways of thrombosis and haemostasis. This assumption could be tested imminently because the WGS data of the UKB participants will be released in 2022, and the high-resolution interaction maps for the BTPD genes will provide an excellent resource for a focused analysis.

Chapter 7 References

Abifadel, Marianne, Mathilde Varret, Jean-Pierre Rabès, Delphine Allard, Khadija Ouguerram, Martine Devillers, Corinne Cruaud, et al. 2003. “Mutations in PCSK9 Cause Autosomal Dominant Hypercholesterolemia.” *Nature Genetics* 34 (2): 154–56.

“ACGS -Association for Clinical Genomic Science.” n.d. Accessed May 5, 2021. https://www.acgs.uk.com/.

ACMG Board of Directors. 2015. “ACMG Policy Statement: Updated Recommendations Regarding Analysis and Reporting of Secondary Findings in Clinical Genome-Scale Sequencing.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 17 (1): 68–69.

Adams, David, Lucia Altucci, Stylianos E. Antonarakis, Juan Ballesteros, Stephan Beck, Adrian Bird, Christoph Bock, et al. 2012. “BLUEPRINT to Decode the Epigenetic Signature Written in Blood.” *Nature Biotechnology* 30 (3): 224–26.

Adli, Mazhar. 2018. “The CRISPR Tool Kit for Genome Editing and beyond.” *Nature Communications*. https://doi.org/10.1038/s41467-018-04252-2.

Agarwal, Harsha, Matthias Reisser, Celina Wortmann, and J. Christof M. Gebhardt. 2017. “Direct Observation of Cell-Cycle-Dependent Interactions between CTCF and Chromatin.” *Biophysical Journal* 112 (10): 2051–55.

Akashi, K., D. Traver, T. Miyamoto, and I. L. Weissman. 2000. “A Clonogenic Common Myeloid Progenitor That Gives Rise to All Myeloid Lineages.” *Nature* 404 (6774): 193–97.

Akbari, Parsa, Ankit Gilani, Olukayode Sosina, Jack A. Kosmicki, Lori Khrimian, Yi-Ya Fang, Trikaldarshi Persaud, et al. 2021. “Sequencing of 640,000 Exomes Identifies GPR75 Variants Associated with Protection from Obesity.” *Science* 373 (6550). https://doi.org/10.1126/science.abf8683.

Aken, Bronwen L., Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, et al. 2016. “The Ensembl Gene Annotation System.” *Database: The Journal of Biological Databases and Curation* 2016 (June). https://doi.org/10.1093/database/baw093.

Albers, Cornelis A., Ana Cvejic, Rémi Favier, Evelien E. Bouwmans, Marie-Christine Alessi, Paul Bertone, Gregory Jordan, et al. 2011. “Exome Sequencing Identifies NBEAL2 as the Causative Gene for Gray Platelet Syndrome.” *Nature Genetics* 43 (8): 735–37.

Albers, Cornelis A., Dirk S. Paul, Harald Schulze, Kathleen Freson, Jonathan C. Stephens, Peter A. Smethurst, Jennifer D. Jolley, et al. 2012. “Compound Inheritance of a Low-Frequency Regulatory SNP and a Rare Null Mutation in Exon-Junction Complex Subunit RBM8A Causes TAR Syndrome.” *Nature Genetics* 44 (4): 435–39, S1–2.

Alberts, Bruce. 2014. *Molecular Biology of the Cell*. 6th ed. New York, NY: Garland Publishing.

Albuschies, Jörg, and Viola Vogel. 2013. “The Role of Filopodia in the Recognition of Nanotopographies.” *Scientific Reports* 3: 1658.

Alexander, David H., John Novembre, and Kenneth Lange. 2009. “Fast Model-Based Estimation of Ancestry in Unrelated Individuals.” *Genome Research* 19 (9): 1655–64.

Alexandrov, Ludmil B., Serena Nik-Zainal, David C. Wedge, Samuel A. J. R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, et al. 2013. “Signatures of Mutational Processes in Human Cancer.” *Nature* 500 (7463): 415–21.

Allis, C. David, and Thomas Jenuwein. 2016. “The Molecular Hallmarks of Epigenetic Control.” *Nature Reviews. Genetics* 17 (8): 487–500.

All of Us Research Program Investigators, Joshua C. Denny, Joni L. Rutter, David B. Goldstein, Anthony Philippakis, Jordan W. Smoller, Gwynne Jenkins, and Eric Dishman. 2019. “The ‘All of Us’ Research Program.” *The New England Journal of Medicine* 381 (7): 668–76.

Althaus, Karina, and Andreas Greinacher. 2009. “MYH9-Related Platelet Disorders.”

Seminars in Thrombosis and Hemostasis. https://doi.org/10.1055/s-0029-1220327.

Amabile, Angelo, Alessandro Migliara, Paola Capasso, Mauro Biffi, Davide Cittaro, Luigi Naldini, and Angelo Lombardo. 2016. “Inheritable Silencing of Endogenous Genes by Hit-and-Run Targeted Epigenetic Editing.” *Cell* 167 (1): 219–32.e14.

Ambrosio, Andrea L., and Santiago M. Di Pietro. 2017. “Storage Pool Diseases Illuminate Platelet Dense Granule Biogenesis.” *Platelets* 28 (2): 138–46.

Amati, B., and H. Land. 1994. “Myc—Max—Mad: A Transcription Factor Network Controlling Cell Cycle Progression, Differentiation and Death.” *Current Opinion in Genetics & Development*. https://www.sciencedirect.com/science/article/pii/0959437X94900981.

Amberger, Joanna S., Carol A. Bocchini, Alan F. Scott, and Ada Hamosh. 2019. “OMIM.org: Leveraging Knowledge across Phenotype-Gene Relationships.” *Nucleic Acids Research* 47 (D1): D1038–43.

Amm, Ingo, Thomas Sommer, and Dieter H. Wolf. 2014. “Protein Quality Control and Elimination of Protein Waste: The Role of the Ubiquitin–proteasome System.” *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1843 (1): 182–96.

Andrews, Robert K., and Michael C. Berndt. 2013. “Bernard-Soulier Syndrome: An Update.” *Seminars in Thrombosis and Hemostasis* 39 (6): 656–62.

Antonarakis, Stylianos E., and Jacques S. Beckmann. 2006. “Mendelian Disorders Deserve More Attention.” *Nature Reviews. Genetics* 7 (4): 277–82.

Antuch, W., P. Güntert, M. Billeter, T. Hawthorne, H. Grossenbacher, and K. Wüthrich. 1994. “NMR Solution Structure of the Recombinant Tick Anticoagulant Protein (rTAP), a Factor Xa Inhibitor from the tickOrnithodoros Moubata.” *FEBS Letters*. https://doi.org/10.1016/0014-5793(94)00941-4.

Arensbergen, Joris van, Bas van Steensel, and Harmen J. Bussemaker. 2014. “In Search of the Determinants of Enhancer–promoter Interaction Specificity.” *Trends in Cell Biology* 24 (11): 695–702.

Ashuach, Tal, David S. Fischer, Anat Kreimer, Nadav Ahituv, Fabian J. Theis, and Nir Yosef. 2019. “MPRAnalyze: Statistical Framework for Massively Parallel Reporter Assays.” *Genome Biology* 20 (1): 183.

“Assigning Variants to Genes (V2G).” n.d. Accessed March 19, 2021. https://genetics-docs.opentargets.org/our-approach/data-pipeline.

Astle, William J., Heather Elding, Tao Jiang, Dave Allen, Dace Ruklisa, Alice L. Mann, Daniel Mead, et al. 2016. “The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease.” *Cell* 167 (5): 1415–29.e19.

Astle, William John, UK Blood Trait GWAS Team, and Cambridge Blueprint Epigenome. 2016. “A GWAS of 170,000 Individuals Identifies Thousands of Alleles Perturbing Blood Cell Traits, Many of Which Are in Super Enhancers Setting Cell Identity.” *Blood*. https://doi.org/10.1182/blood.v128.22.2652.2652.

Audano, Peter A., Arvis Sulovari, Tina A. Graves-Lindsay, Stuart Cantsilieris, Melanie Sorensen, Annemarie E. Welch, Max L. Dougherty, et al. 2019. “Characterizing the Major Structural Variant Alleles of the Human Genome.” *Cell* 176 (3): 663–75.e19.

Avsec, Žiga, Roman Kreuzhuber, Johnny Israeli, Nancy Xu, Jun Cheng, Avanti Shrikumar, Abhimanyu Banerjee, et al. 2019. “The Kipoi Repository Accelerates Community Exchange and Reuse of Predictive Models for Genomics.” *Nature Biotechnology* 37 (6): 592–600.

Badolato, Raffaele, Alberto Prandini, Sonia Caracciolo, Francesca Colombo, Giovanna Tabellini, Mauro Giacomelli, Maria E. Cantarini, et al. 2012. “Exome Sequencing Reveals a Pallidin Mutation in a Hermansky-Pudlak--like Primary Immunodeficiency Syndrome.” *Blood, The Journal of the American Society of Hematology* 119 (13): 3185–87.

Baird, Nicola, John Clancy, and Andrew McVicar. 2005. *Perioperative Practice:*

Fundamentals of Homeostasis. Routledge.

Bamforth, S. D., J. Bragança, J. J. Eloranta, J. N. Murdoch, F. I. Marques, K. R. Kranc, H. Farza, D. J. Henderson, H. C. Hurst, and S. Bhattacharya. 2001. “Cardiac Malformations, Adrenal Agenesis, Neural Crest Defects and Exencephaly in Mice Lacking Cited2, a New Tfap2 Co-Activator.” *Nature Genetics* 29 (4): 469–74.

Bariana, Tadbir K., Willem H. Ouwehand, Jose A. Guerrero, Keith Gomez, and BRIDGE Bleeding, Thrombotic and Platelet Disorders and ThromboGenomics Consortia. 2017. “Dawning of the Age of Genomics for Platelet Granule Disorders: Improving Insight, Diagnosis and Management.” *British Journal of Haematology* 176 (5): 705–20.

Barrell, B. G., A. T. Bankier, and J. Drouin. 1979. “A Different Genetic Code in Human Mitochondria.” *Nature* 282 (5735): 189–94.

Barrett, Rowan D. H., and Dolph Schluter. 2008. “Adaptation from Standing Genetic Variation.” *Trends in Ecology & Evolution* 23 (1): 38–44.

Barrio-Hernandez, Inigo, Jeremy Schwartzentruber, Anjali Shrivastava, Noemi del-Toro, Qian Zhang, Glyn Bradley, Henning Hermjakob, et al. n.d. “Network Expansion of Genetic Associations Defines a Pleiotropy Map of Human Cell Biology.” https://doi.org/10.1101/2021.07.19.452924.

Barski, Artem, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E. Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. 2007. “High-Resolution Profiling of Histone Methylations in the Human Genome.” *Cell* 129 (4): 823–37.

Bartman, Caroline R., Sarah C. Hsu, Chris C-S Hsiung, Arjun Raj, and Gerd A. Blobel. 2016. “Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping.” *Molecular Cell* 62 (2): 237–47.

Barton, Nick H., Alison M. Etheridge, and Amandine Véber. n.d. “The Infinitesimal Model.” https://doi.org/10.1101/039768.

Baryawno, Ninib, Nicolas Severe, and David T. Scadden. 2017. “Hematopoiesis: Reconciling Historic Controversies about the Niche.” *Cell Stem Cell* 20 (5): 590–92.

Battulin, Nariman, Veniamin S. Fishman, Alexander M. Mazur, Mikhail Pomaznoy, Anna A. Khabarova, Dmitry A. Afonnikov, Egor B. Prokhortchouk, and Oleg L. Serov. 2015. “Comparison of the Three-Dimensional Organization of Sperm and Fibroblast Genomes Using the Hi-C Approach.” *Genome Biology* 16 (April): 77.

Bearer, E. L., J. M. Prakash, and Z. Li. 2002. “Actin Dynamics in Platelets.” *International Review of Cytology* 217: 137–82.

Bedell, M. A., D. A. Largaespada, N. A. Jenkins, and N. G. Copeland. 1997. “Mouse Models of Human Disease. Part II: Recent Progress and Future Directions.” *Genes & Development* 11 (1): 11–43.

Bejerano, Gill, Michael Pheasant, Igor Makunin, Stuart Stephen, W. James Kent, John S. Mattick, and David Haussler. 2004. “Ultraconserved Elements in the Human Genome.” *Science* 304 (5675): 1321–25.

Belaghzal, Houda, Job Dekker, and Johan H. Gibcus. 2017. “Hi-C 2.0: An Optimized Hi-C Procedure for High-Resolution Genome-Wide Mapping of Chromosome Conformation.” *Methods* 123 (July): 56–65.

Bellanné-Chantelot, Christine, Matthieu Mosca, Caroline Marty, Rémi Favier, William Vainchenker, and Isabelle Plo. 2017. “Identification of MPL R102P Mutation in Hereditary Thrombocytosis.” *Frontiers in Endocrinology* 8 (September): 235.

Benabdallah, Nezha S., Iain Williamson, Robert S. Illingworth, Lauren Kane, Shelagh Boyle, Dipta Sengupta, Graeme R. Grimes, Pierre Therizols, and Wendy A. Bickmore. 2019. “Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation.” *Molecular Cell* 76 (3): 473–84.e7.

Bentley, D. R. 2000. “Decoding the Human Genome Sequence.” *Human Molecular Genetics* 9 (16): 2353–58.

Bernstein, Bradley E., John A. Stamatoyannopoulos, Joseph F. Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, et al. 2010. “The NIH Roadmap Epigenomics Mapping Consortium.” *Nature Biotechnology* 28 (10): 1045–48.

Bertina, R. M., B. P. Koeleman, T. Koster, F. R. Rosendaal, R. J. Dirven, H. de Ronde, P. A. van der Velden, and P. H. Reitsma. 1994. “Mutation in Blood Coagulation Factor V Associated with Resistance to Activated Protein C.” *Nature* 369 (6475): 64–67.

Bertozzi, Cara C., Paul R. Hess, and Mark L. Kahn. 2010. “Platelets: Covert Regulators of Lymphatic Development.” *Arteriosclerosis, Thrombosis, and Vascular Biology* 30 (12): 2368–71.

Bevacqua, Romina J., Xiaoqing Dai, Jonathan Y. Lam, Xueying Gu, Mollie S. H. Friedlander, Krissie Tellez, Irene Miguel-Escalada, et al. 2021. “CRISPR-Based Genome Editing in Primary Human Pancreatic Islet Cells.” *Nature Communications* 12 (1): 2397.

Beyter, Doruk, Helga Ingimundardottir, Asmundur Oddsson, Hannes P. Eggertsson, Eythor Bjornsson, Hakon Jonsson, Bjarni A. Atlason, et al. 2020. “Long Read Sequencing of 3,622 Icelanders Provides Insight into the Role of Structural Variants in Human Diseases and Other Traits.” *bioRxiv*. https://doi.org/10.1101/848366.

Bick, Alexander G., Joshua S. Weinstock, Satish K. Nandakumar, Charles P. Fulco, Erik L. Bao, Seyedeh M. Zekavat, Mindy D. Szeto, et al. 2020. “Inherited Causes of Clonal Haematopoiesis in 97,691 Whole Genomes.” *Nature* 586 (7831): 763–68.

Bicknell, Alicia A., Can Cenik, Hon N. Chua, Frederick P. Roth, and Melissa J. Moore. 2012. “Introns in UTRs: Why We Should Stop Ignoring Them.” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 34 (12): 1025–34.

Biémont, Christian, and Cristina Vieira. 2006. “Genetics: Junk DNA as an Evolutionary Force.” *Nature* 443 (7111): 521–24.

Bintu, Bogdan, Leslie J. Mateo, Jun-Han Su, Nicholas A. Sinnott-Armstrong, Mirae Parker, Seon Kinrot, Kei Yamaya, Alistair N. Boettiger, and Xiaowei Zhuang. 2018. “Super-Resolution Chromatin Tracing Reveals Domains and Cooperative Interactions in Single Cells.” *Science* 362 (6413). https://doi.org/10.1126/science.aau1783.

Blair, Price, and Robert Flaumenhaft. 2009. “Platelet α-Granules: Basic Biology and Clinical Correlates.” *Blood Reviews*. https://doi.org/10.1016/j.blre.2009.04.001.

“Blood Atlas.” n.d. Accessed August 8, 2021. https://blueprint.haem.cam.ac.uk/mRNA/.

Blundell, J. 1818. “Experiments on the Transfusion of Blood by the Syringe.” *Medico-Chirurgical Transactions* 9 (Pt 1): 56–92.

Bocher, Ozvan, and Emmanuelle Génin. 2020. “Rare Variant Association Testing in the Non-Coding Genome.” *Human Genetics* 139 (11): 1345–62.

Boettiger, Alistair, and Sedona Murphy. 2020. “Advances in Chromatin Imaging at Kilobase-Scale Resolution.” *Trends in Genetics: TIG* 36 (4): 273–87.

Boffelli, Dario, Marcelo A. Nobrega, and Edward M. Rubin. 2004. “Comparative Genomics at the Vertebrate Extremes.” *Nature Reviews. Genetics* 5 (6): 456–65.

Bomba, Lorenzo, Klaudia Walter, and Nicole Soranzo. 2017. “The Impact of Rare and Low-Frequency Genetic Variants in Common Disease.” *Genome Biology*. https://doi.org/10.1186/s13059-017-1212-4.

Bonev, Boyan, Netta Mendelson Cohen, Quentin Szabo, Lauriane Fritsch, Giorgio L. Papadopoulos, Yaniv Lubling, Xiaole Xu, et al. 2017. “Multiscale 3D Genome Rewiring during Mouse Neural Development.” *Cell* 171 (3): 557–72.e24.

Bothma, Jacques P., Hernan G. Garcia, Samuel Ng, Michael W. Perry, Thomas Gregor, and Michael Levine. 2015. “Enhancer Additivity and Non-Additivity Are Determined by Enhancer Strength in the Drosophila Embryo.” *eLife* 4 (August). https://doi.org/10.7554/eLife.07956.

Boycott, Kym M., Ana Rath, Jessica X. Chong, Taila Hartley, Fowzan S. Alkuraya, Gareth Baynam, Anthony J. Brookes, et al. 2017. “International Cooperation to Enable the

Diagnosis of All Rare Genetic Diseases." *American Journal of Human Genetics* 100 (5): 695–705.

Boyer, Thomas D., Michael Peter Manns, Arun J. Sanyal, and David Zakim. 2012. *Zakim and Boyer's Hepatology: A Textbook of Liver Disease*. Elsevier Health Sciences.

Boyle, Evan A., Yang I. Li, and Jonathan K. Pritchard. 2017. "An Expanded View of Complex Traits: From Polygenic to Omnigenic." *Cell* 169 (7): 1177–86.

Broze, George J., Jr, and Thomas J. Girard. 2012. "Tissue Factor Pathway Inhibitor: Structure-Function." *Frontiers in Bioscience* 17 (January): 262–80.

Brunt, Elizabeth M., Annette S. H. Gouw, Stefan G. Hubscher, Dina G. Tiniakos, Pierre Bedossa, Alastair D. Burt, Francesco Callea, et al. 2014. "Pathology of the Liver Sinusoids." *Histopathology* 64 (7): 907–20.

Budak, Yasemin Ustundag, Murat Polat, and Kagan Huysal. 2016. "The Use of Platelet Indices, Plateletcrit, Mean Platelet Volume and Platelet Distribution Width in Emergency Non-Traumatic Abdominal Surgery: A Systematic Review." *Biochemia Medica: Casopis Hrvatskoga Drustva Medicinskih Biokemicara / HDMB* 26 (2): 178–93.

Buecker, Christa, and Joanna Wysocka. 2012. "Enhancers as Information Integration Hubs in Development: Lessons from Genomics." *Trends in Genetics: TIG* 28 (6): 276–84.

Buenrostro, Jason D., Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. 2013. "Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position." *Nature Methods* 10 (12): 1213–18.

Buiting, Karin, Shinji Saitoh, Stephanie Gross, Bärbel Dittrich, Stuart Schwartz, Robert D. Nicholls, and Bernhard Horsthemke. 1995. "Inherited Microdeletions in the Angelman and Prader–Willi Syndromes Define an Imprinting Centre on Human Chromosome 15." *Nature Genetics* 9 (4): 395–400.

Bulato, Cristiana, Elena Campello, Sabrina Gavasso, Sara Maggiolo, Daniela Tormene, and Paolo Simioni. 2018. "Peculiar Laboratory Phenotype/ Genotype Relationship due to Compound Inherited Protein C Defects in a Child with Severe Venous Thromboembolism." *Hämostaseologie*. https://doi.org/10.5482/hamo-17-03-0013.

Bulger, Michael, and Mark Groudine. 2011. "Functional and Mechanistic Diversity of Distal Transcription Enhancers." *Cell* 144 (3): 327–39.

Buniello, Annalisa, Jacqueline A. L. MacArthur, Maria Cerezo, Laura W. Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, et al. 2019. "The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019." *Nucleic Acids Research* 47 (D1): D1005–12.

Burren, Oliver S., Arcadio Rubio García, Biola-Maria Javierre, Daniel B. Rainbow, Jonathan Cairns, Nicholas J. Cooper, John J. Lambourne, et al. 2017. "Chromosome Contacts in Activated T Cells Identify Autoimmune Disease Candidate Genes." *Genome Biology* 18 (1): 165.

Burt, Alastair D., Linda D. Ferrell, and Stefan G. Hubscher. 2017. *MacSween's Pathology of the Liver*. Elsevier.

Bush, William S., and Jason H. Moore. 2012. "Chapter 11: Genome-Wide Association Studies." *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1002822.

Butenas, Saulius, Thomas Orfeo, and Kenneth G. Mann. 2009. "Tissue Factor in Coagulation: Which? Where? When?" *Arteriosclerosis, Thrombosis, and Vascular Biology* 29 (12): 1989–96.

Butterworth, P. J. 2005. "Lehninger: Principles of Biochemistry (4th Edn) D. L. Nelson and M. C. Cox, W. H. Freeman & Co., New York, 1119 Pp (plus 17 Pp Glossary), ISBN 0-7167-4339-6 (2004)." *Cell Biochemistry and Function*. https://doi.org/10.1002/cbf.1216.

Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp,

Allan Motyer, et al. 2018. “The UK Biobank Resource with Deep Phenotyping and Genomic Data.” *Nature* 562 (7726): 203–9.

Byeon, Gun Woo, Elif Sarinay Cenik, Lihua Jiang, Hua Tang, Rhiju Das, and Maria Barna. 2021. “Functional and Structural Basis of Extreme Conservation in Vertebrate 5’ Untranslated Regions.” *Nature Genetics*, April. https://doi.org/10.1038/s41588-021-00830-1.

Cairns, Jonathan, Paula Freire-Pritchett, Steven W. Wingett, Csilla Várnai, Andrew Dimond, Vincent Plagnol, Daniel Zerbino, et al. 2016a. “CHiCAGO: Robust Detection of DNA Looping Interactions in Capture Hi-C Data.” *Genome Biology*. https://doi.org/10.1186/s13059-016-0992-2.

———. 2016b. “CHiCAGO: Robust Detection of DNA Looping Interactions in Capture Hi-C Data.” *Genome Biology* 17 (1): 127.

Calo, Eliezer, and Joanna Wysocka. 2013. “Modification of Enhancer Chromatin: What, How, and Why?” *Molecular Cell* 49 (5): 825–37.

Cambon-Thomsen, A., E. Rial-Sebbag, and B. M. Knoppers. 2007. “Trends in Ethical and Legal Frameworks for the Use of Human Biobanks.” *The European Respiratory Journal: Official Journal of the European Society for Clinical Respiratory Physiology* 30 (2): 373–82.

Cano-Gamez, Eddie, and Gosia Trynka. 2020. “From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases.” *Frontiers in Genetics* 11 (May): 424.

Capecchi, M. R. 1989. “Altering the Genome by Homologous Recombination.” *Science* 244 (4910): 1288–92.

Capra, John A., Roman A. Laskowski, Janet M. Thornton, Mona Singh, and Thomas A. Funkhouser. 2009. “Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure.” *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1000585.

Carlson, Robert. 2003. “The Pace and Proliferation of Biological Technologies.” *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science* 1 (3): 203–14.

Carvalho-Silva, Denise, Andrea Pierleoni, Miguel Pignatelli, Chuangkee Ong, Luca Fumis, Nikiforos Karamanis, Miguel Carmona, et al. 2019. “Open Targets Platform: New Developments and Updates Two Years on.” *Nucleic Acids Research* 47 (D1): D1056–65.

Castaldo, Giuseppe, Valeria D’Argenio, Paola Nardiello, Federica Zarrilli, Veronica Sanna, Angiola Rocino, Antonio Coppola, Giovanni Di Minno, and Francesco Salvatore. 2007. “Haemophilia A: Molecular Insights.” *Clinical Chemical Laboratory Medicine*. https://doi.org/10.1515/cclm.2007.093.

Castaman, Giancarlo, and Silvia Linari. 2017. “Diagnosis and Treatment of von Willebrand Disease and Rare Bleeding Disorders.” *Journal of Clinical Medicine Research* 6 (4). https://doi.org/10.3390/jcm6040045.

Cavazza, Marianna, BURQOL-RD Research Network, Yllka Kodra, Patrizio Armeni, Marta De Santis, Julio López-Bastida, Renata Linertová, et al. 2016. “Social/economic Costs and Quality of Life in Patients with Haemophilia in Europe.” *The European Journal of Health Economics*. https://doi.org/10.1007/s10198-016-0785-2.

Cenik, Can, Hon Nian Chua, Hui Zhang, Stefan P. Tarnawsky, Abdalla Akef, Adnan Derti, Murat Tasan, Melissa J. Moore, Alexander F. Palazzo, and Frederick P. Roth. 2011. “Genome Analysis Reveals Interplay between 5’UTR Introns and Nuclear mRNA Export for Secretory and Mitochondrial Genes.” *PLoS Genetics* 7 (4): e1001366.

Chaisson, Mark J. P., John Huddleston, Megan Y. Dennis, Peter H. Sudmant, Maika Malig, Fereydoun Hormozdiari, Francesca Antonacci, et al. 2015. “Resolving the Complexity of the Human Genome Using Single-Molecule Sequencing.” *Nature* 517 (7536):

608–11.
Chapin, John C., and Katherine A. Hajjar. 2015. “Fibrinolysis and the Control of Blood Coagulation.” *Blood Reviews*. https://doi.org/10.1016/j.blre.2014.09.003.
Chaudhry, Raheel, Syed Muhammad Usama, and Hani M. Babiker. 2020. “Physiology, Coagulation Pathways.” In *StatPearls*. Treasure Island (FL): StatPearls Publishing.
Chen, Baohui, Luke A. Gilbert, Beth A. Cimini, Joerg Schnitzbauer, Wei Zhang, Gene-Wei Li, Jason Park, et al. 2013. “Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System.” *Cell* 155 (7): 1479–91.
Cheng, Hui, Zhaofeng Zheng, and Tao Cheng. 2020. “New Paradigms on Hematopoietic Stem Cell Differentiation.” *Protein & Cell* 11 (1): 34–44.
Chen, Hongtao, Michal Levo, Lev Barinov, Miki Fujioka, James B. Jaynes, and Thomas Gregor. 2018. “Dynamic Interplay between Enhancer–promoter Topology and Gene Activity.” *Nature Genetics* 50 (9): 1296–1303.
Chen, Lu, Bing Ge, Francesco Paolo Casale, Louella Vasquez, Tony Kwan, Diego Garrido-Martín, Stephen Watt, et al. 2016. “Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells.” *Cell* 167 (5): 1398–1414.e24.
Chen, Lu, Myrto Kostadima, Joost H. A. Martens, Giovanni Canu, Sara P. Garcia, Ernest Turro, Kate Downes, et al. 2014. “Transcriptional Diversity during Lineage Commitment of Human Blood Progenitors.” *Science* 345 (6204): 1251033.
Chen, Ming-Huei, Laura M. Raffield, Abdou Mousas, Saori Sakaue, Jennifer E. Huffman, Arden Moscati, Bhavi Trivedi, et al. 2020. “Trans-Ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations.” *Cell* 182 (5): 1198–1213.e14.
Chepelev, Iouri, Gang Wei, Dara Wangsa, Qingsong Tang, and Keji Zhao. 2012. “Characterization of Genome-Wide Enhancer-Promoter Interactions Reveals Co-Expression of Interacting Genes and Modes of Higher Order Chromatin Organization.” *Cell Research* 22 (3): 490–503.
Cheshier, S. H., S. J. Morrison, X. Liao, and I. L. Weissman. 1999. “In Vivo Proliferation and Cell Cycle Kinetics of Long-Term Self-Renewing Hematopoietic Stem Cells.” *Proceedings of the National Academy of Sciences of the United States of America* 96 (6): 3120–25.
Chesterman, C. N. 1988. “Vascular Endothelium, Haemostasis and Thrombosis.” *Blood Reviews* 2 (2): 88–94.
Cheung, Christine, Andreia S. Bernardo, Matthew W. B. Trotter, Roger A. Pedersen, and Sanjay Sinha. 2012. “Generation of Human Vascular Smooth Muscle Subtypes Provides Insight into Embryological Origin-Dependent Disease Susceptibility.” *Nature Biotechnology* 30 (2): 165–73.
Chien, Shu. 2008. “Effects of Disturbed Flow on Endothelial Cells.” *Annals of Biomedical Engineering* 36 (4): 554–62.
Chimpanzee Sequencing and Analysis Consortium. 2005. “Initial Sequence of the Chimpanzee Genome and Comparison with the Human Genome.” *Nature* 437 (7055): 69–87.
Choi, K., M. Kennedy, A. Kazarov, J. C. Papadimitriou, and G. Keller. 1998. “A Common Precursor for Hematopoietic and Endothelial Cells.” *Development* 125 (4): 725–32.
Choobdar, Sarvenaz, Mehmet E. Ahsen, Jake Crawford, Mattia Tomasoni, Tao Fang, David Lamparter, Junyuan Lin, et al. 2019. “Assessment of Network Module Identification across Complex Diseases.” *Nature Methods* 16 (9): 843–52.
Cho, Seung Woo, Sojung Kim, Yongsub Kim, Jiyeon Kweon, Heon Seok Kim, Sangsu Bae, and Jin-Soo Kim. 2014. “Analysis of off-Target Effects of CRISPR/Cas-Derived RNA-Guided Endonucleases and Nickases.” *Genome Research* 24 (1): 132–41.
Choudhury, Ananyo, Shaun Aron, Laura R. Botigué, Dhriti Sengupta, Gerrit Botha, Taoufik

Bensellak, Gordon Wells, et al. 2020. “High-Depth African Genomes Inform Human Migration and Health.” *Nature* 586 (7831): 741–48.

Ciciotte, Steven L., Babette Gwynn, Kengo Moriyama, Marjan Huizing, William A. Gahl, Juan S. Bonifacino, and Luanne L. Peters. 2003. “Cappuccino, a Mouse Model of Hermansky-Pudlak Syndrome, Encodes a Novel Protein That Is Part of the Pallidin-Muted Complex (BLOC-1).” *Blood* 101 (11): 4402–7.

Clamp, Michele, Ben Fry, Mike Kamal, Xiaohui Xie, James Cuff, Michael F. Lin, Manolis Kellis, Kerstin Lindblad-Toh, and Eric S. Lander. 2007. “Distinguishing Protein-Coding and Noncoding Genes in the Human Genome.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (49): 19428–33.

Clark, Michael J., Rui Chen, Hugo Y. K. Lam, Konrad J. Karczewski, Rong Chen, Ghia Euskirchen, Atul J. Butte, and Michael Snyder. 2011. “Performance Comparison of Exome DNA Sequencing Technologies.” *Nature Biotechnology* 29 (10): 908–14.

“Classification of Diseases (ICD).” n.d. Accessed May 15, 2021. https://www.who.int/standards/classifications/classification-of-diseases.

Claussnitzer, Melina, Judy H. Cho, Rory Collins, Nancy J. Cox, Emmanouil T. Dermitzakis, Matthew E. Hurles, Sekar Kathiresan, et al. 2020. “A Brief History of Human Disease Genetics.” *Nature* 577 (7789): 179–89.

Collier, Dami A., Anna De Marco, Isabella A. T. M. Ferreira, Bo Meng, Rawlings P. Datir, Alexandra C. Walls, Steven A. Kemp, et al. 2021. “Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA Vaccine-Elicited Antibodies.” *Nature* 593 (7857): 136–41.

Collins, Francis S., Eric D. Green, Alan E. Guttmacher, Mark S. Guyer, and US National Human Genome Research Institute. 2003. “A Vision for the Future of Genomics Research.” *Nature* 422 (6934): 835–47.

Collins, Francis S., Michael Morgan, and Aristides Patrinos. 2003. “The Human Genome Project: Lessons from Large-Scale Biology.” *Science* 300 (5617): 286–90.

Collins, Janine, William J. Astle, Karyn Megy, Andrew D. Mumford, and Dragana Vuckovic. 2021. “Advances in Understanding the Pathogenesis of Hereditary Macrothrombocytopenia.” *British Journal of Haematology*, March. https://doi.org/10.1111/bjh.17409.

Collins, Ryan L., Harrison Brand, Konrad J. Karczewski, Xuefang Zhao, Jessica Alföldi, Laurent C. Francioli, Amit V. Khera, et al. 2020. “A Structural Variation Reference for Medical and Population Genetics.” *Nature* 581 (7809): 444–51.

Coppola, Candice J., Ryne C Ramaker, and Eric M. Mendenhall. 2016. “Identification and Function of Enhancers in the Human Genome.” *Human Molecular Genetics* 25 (R2): R190–97.

Corbin, Laura J., Vanessa Y. Tan, David A. Hughes, Kaitlin H. Wade, Dirk S. Paul, Katherine E. Tansey, Frances Butcher, et al. 2018. “Formalising Recall by Genotype as an Efficient Approach to Detailed Phenotyping and Causal Inference.” *Nature Communications* 9 (1): 711.

Corces, M. Ryan, Jason D. Buenrostro, Beijing Wu, Peyton G. Greenside, Steven M. Chan, Julie L. Koenig, Michael P. Snyder, et al. 2016. “Lineage-Specific and Single-Cell Chromatin Accessibility Charts Human Hematopoiesis and Leukemia Evolution.” *Nature Genetics* 48 (10): 1193–1203.

Cortes, Adrian, Patrick K. Albers, Calliope A. Dendrou, Lars Fugger, and Gil McVean. 2020. “Identifying Cross-Disease Components of Genetic Risk across Hospital Data in the UK Biobank.” *Nature Genetics* 52 (1): 126–34.

Cortes, Adrian, Calliope A. Dendrou, Allan Motyer, Luke Jostins, Damjan Vukcevic, Alexander Dilthey, Peter Donnelly, Stephen Leslie, Lars Fugger, and Gil McVean. 2017. “Bayesian Analysis of Genetic Association across Tree-Structured Routine Healthcare Data in the UK Biobank.” *Nature Genetics* 49 (9): 1311.

Cos, Hippocrates of, and Hippocrates of Cos. 1931. “Nature of Man.” *Digital Loeb Classical Library*. https://doi.org/10.4159/dlcl.hippocrates_cos-nature_man.1931.

Coster, Wouter De, Wouter De Coster, and Christine Van Broeckhoven. 2019. “Newest Methods for Detecting Structural Variations.” *Trends in Biotechnology*. https://doi.org/10.1016/j.tibtech.2019.02.003.

Cozzolino, Flora, Ilaria Iacobucci, Vittoria Monaco, and Maria Monti. 2021. “Protein–DNA/RNA Interactions: An Overview of Investigation Methods in the -Omics Era.” *Journal of Proteome Research*. https://doi.org/10.1021/acs.jproteome.1c00074.

Creyghton, Menno P., Albert W. Cheng, G. Grant Welstead, Tristan Kooistra, Bryce W. Carey, Eveline J. Steine, Jacob Hanna, et al. 2010. “Histone H3K27ac Separates Active from Poised Enhancers and Predicts Developmental State.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (50): 21931–36.

Crick, F. 1970. “Central Dogma of Molecular Biology.” *Nature* 227 (5258): 561–63.

Crick, F. H. 1966a. “The Genetic Code--Yesterday, Today, and Tomorrow.” *Cold Spring Harbor Symposia on Quantitative Biology* 31: 1–9.

———. 1966b. “Codon--Anticodon Pairing: The Wobble Hypothesis.” *Journal of Molecular Biology* 19 (2): 548–55.

Crick, F. H. C. 1958. “On Protein Synthesis. InSymposia of the Society for Experimental Biology; Number XII: The Biological Replication of Macromolecules.” Cambridge University Press: Cambridge, UK.

Crosby, Christopher V., Paul A. Fleming, W. Scott Argraves, Monica Corada, Lucia Zanetta, Elisabetta Dejana, and Christopher J. Drake. 2005. “VE-Cadherin Is Not Required for the Formation of Nascent Blood Vessels but Acts to Prevent Their Disassembly.” *Blood* 105 (7): 2771–76.

Crous-Bou, Marta, Laura B. Harrington, and Christopher Kabrhel. 2016. “Environmental and Genetic Risk Factors Associated with Venous Thromboembolism.” *Seminars in Thrombosis and Hemostasis* 42 (8): 808–20.

Cutting, Garry R. 2010. “Modifier Genes in Mendelian Disorders: The Example of Cystic Fibrosis.” *Annals of the New York Academy of Sciences* 1214 (December): 57–69.

Cvejic, Ana, Lonneke Haer-Wigman, Jonathan C. Stephens, Myrto Kostadima, Peter A. Smethurst, Mattia Frontini, Emile van den Akker, et al. 2013. “SMIM1 Underlies the Vel Blood Group and Influences Red Blood Cell Traits.” *Nature Genetics* 45 (5): 542–45.

Dahlbäck, Björn, and Bruno O. Villoutreix. 2005. “Regulation of Blood Coagulation by the Protein C Anticoagulant Pathway: Novel Insights into Structure-Function Relationships and Molecular Recognition.” *Arteriosclerosis, Thrombosis, and Vascular Biology* 25 (7): 1311–20.

Danecek, Petr, James K. Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O. Pollard, Andrew Whitwham, et al. 2021. “Twelve Years of SAMtools and BCFtools.” *GigaScience* 10 (2). https://doi.org/10.1093/gigascience/giab008.

Dao, Lan T. M., Ariel O. Galindo-Albarrán, Jaime A. Castro-Mondragon, Charlotte Andrieu-Soler, Alejandra Medina-Rivera, Charbel Souaid, Guillaume Charbonnier, et al. 2017. “Genome-Wide Characterization of Mammalian Promoters with Distal Enhancer Functions.” *Nature Genetics* 49 (7): 1073–81.

Darby, Sarah C., Sau Wan Kan, Rosemary J. Spooner, Paul L. F. Giangrande, Frank G. H. Hill, Charles R. M. Hay, Christine A. Lee, Christopher A. Ludlam, and Michael Williams. 2007. “Mortality Rates, Life Expectancy, and Causes of Death in People with Hemophilia A or B in the United Kingdom Who Were Not Infected with HIV.” *Blood*. https://doi.org/10.1182/blood-2006-10-050435.

Darby, S. C., D. M. Keeling, R. J. D. Spooner, S. Wan Kan, P. L. F. Giangrande, P. W. Collins, F. G. H. Hill, C. R. M. Hay, and UK Haemophilia Centre Doctors’ Organisation. 2004. “The Incidence of Factor VIII and Factor IX Inhibitors in the Hemophilia

Population of the UK and Their Effect on Subsequent Mortality, 1977-99." *Journal of Thrombosis and Haemostasis: JTH* 2 (7): 1047–54.

Davie, E. W., K. Fujikawa, and W. Kisiel. 1991. "The Coagulation Cascade: Initiation, Maintenance, and Regulation." *Biochemistry* 30 (43): 10363–70.

Davie, E. W., and O. D. Ratnoff. 1964. "WATERFALL SEQUENCE FOR INTRINSIC BLOOD CLOTTING." *Science* 145 (3638): 1310–12.

Davis, Carrie A., Benjamin C. Hitz, Cricket A. Sloan, Esther T. Chan, Jean M. Davidson, Idan Gabdank, Jason A. Hilton, et al. 2018. "The Encyclopedia of DNA Elements (ENCODE): Data Portal Update." *Nucleic Acids Research* 46 (D1): D794–801.

Deaton, A. M., and A. Bird. 2011. "CpG Islands and the Regulation of Transcription." *Genes & Development*. https://doi.org/10.1101/gad.2037511.

De Gobbi, Marco, Vip Viprakasit, Jim R. Hughes, Chris Fisher, Veronica J. Buckle, Helena Ayyub, Richard J. Gibbons, et al. 2006. "A Regulatory SNP Causes a Human Genetic Disease by Creating a New Transcriptional Promoter." *Science* 312 (5777): 1215–17.

Dekker, Job, Andrew S. Belmont, Mitchell Guttman, Victor O. Leshyk, John T. Lis, Stavros Lomvardas, Leonid A. Mirny, et al. 2017. "The 4D Nucleome Project." *Nature* 549 (7671): 219–26.

Dekker, Job, Marc A. Marti-Renom, and Leonid A. Mirny. 2013. "Exploring the Three-Dimensional Organization of Genomes: Interpreting Chromatin Interaction Data." *Nature Reviews. Genetics* 14 (6): 390–403.

DeLaForest, Ann, Masato Nagaoka, Karim Si-Tayeb, Fallon K. Noto, Genevieve Konopka, Michele A. Battle, and Stephen A. Duncan. 2011. "HNF4A Is Essential for Specification of Hepatic Progenitors from Human Pluripotent Stem Cells." *Development* 138 (19): 4143–53.

Delaneau, O., M. Zazhytska, C. Borel, G. Giannuzzi, G. Rey, C. Howald, S. Kumar, et al. 2019. "Chromatin Three-Dimensional Interactions Mediate Genetic Effects on Gene Expression." *Science* 364 (6439). https://doi.org/10.1126/science.aat8266.

De La Vega, Francisco M., and Carlos D. Bustamante. 2018. "Polygenic Risk Scores: A Biased Prediction?" *Genome Medicine* 10 (1): 100.

Dempster, E. R., and I. M. Lerner. 1950. "Heritability of Threshold Characters." *Genetics* 35 (2): 212–36.

Dendrou, Calliope A., Vincent Plagnol, Erik Fung, Jennie H. M. Yang, Kate Downes, Jason D. Cooper, Sarah Nutland, et al. 2009. "Cell-Specific Protein Phenotypes for the Autoimmune Locus IL2RA Using a Genotype-Selectable Human Bioresource." *Nature Genetics* 41 (9): 1011–15.

Deng, Wulan, Jongjoo Lee, Hongxin Wang, Jeff Miller, Andreas Reik, Philip D. Gregory, Ann Dean, and Gerd A. Blobel. 2012. "Controlling Long-Range Genomic Interactions at a Native Locus by Targeted Tethering of a Looping Factor." *Cell* 149 (6): 1233–44.

Despang, Alexandra, Robert Schöpflin, Martin Franke, Salaheddine Ali, Ivana Jerković, Christina Paliou, Wing-Lee Chan, et al. 2019. "Functional Dissection of the Sox9--Kcnj2 Locus Identifies Nonessential and Instructive Roles of TAD Architecture." *Nature Genetics* 51 (8): 1263–71.

Diamondstone, L. S., L. M. Aledort, J. J. Goedert, and MULTICENTRE HEMOPHILIA COHORT STUDY. 2002. "Factors Predictive of Death among HIV-Uninfected Persons with Haemophilia and Other Congenital Coagulation Disorders." *Haemophilia*. https://doi.org/10.1046/j.1365-2516.2002.00651.x.

Di Angelantonio, Emanuele, Simon G. Thompson, Stephen Kaptoge, Carmel Moore, Matthew Walker, Jane Armitage, Willem H. Ouwehand, David J. Roberts, John Danesh, and INTERVAL Trial Group. 2017. "Efficiency and Safety of Varying the Frequency of Whole Blood Donation (INTERVAL): A Randomised Trial of 45 000 Donors." *The Lancet* 390 (10110): 2360–71.

Dickel, Diane E., Athena R. Ypsilanti, Ramón Pla, Yiwen Zhu, Iros Barozzi, Brandon J. Mannion, Yupar S. Khin, et al. 2018. “Ultraconserved Enhancers Are Required for Normal Development.” *Cell* 172 (3): 491–99.e15.

DiStefano, Johanna K. 2018. “The Emerging Role of Long Noncoding RNAs in Human Disease.” *Methods in Molecular Biology* 1706: 91–110.

Dixon, Jesse R., Inkyung Jung, Siddarth Selvaraj, Yin Shen, Jessica E. Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, et al. 2015. “Chromatin Architecture Reorganization during Stem Cell Differentiation.” *Nature* 518 (7539): 331–36.

Dixon, Jesse R., Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. 2012. “Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions.” *Nature* 485 (7398): 376–80.

“DNA Sequencing Costs: Data.” n.d. Accessed August 13, 2021. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.

Doherty, Tara M., and Ashley Kelley. 2020. “Bleeding Disorders.” In *StatPearls*. Treasure Island (FL): StatPearls Publishing.

Dohm, Juliane C., Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. 2008. “Substantial Biases in Ultra-Short Read Data Sets from High-Throughput DNA Sequencing.” *Nucleic Acids Research* 36 (16): e105.

Doulatov, Sergei, Faiyaz Notta, Kolja Eppert, Linh T. Nguyen, Pamela S. Ohashi, and John E. Dick. 2010. “Revised Map of the Human Progenitor Hierarchy Shows the Origin of Macrophages and Dendritic Cells in Early Lymphoid Development.” *Nature Immunology* 11 (7): 585–93.

Downes, Kate, Karyn Megy, Daniel Duarte, Minka Vries, Johanna Gebhart, Stefanie Hofer, Olga Shamardina, et al. 2019. “Diagnostic High-Throughput Sequencing of 2396 Patients with Bleeding, Thrombotic, and Platelet Disorders.” *Blood* 134 (23): 2082–91.

Dukler, Noah, Brad Gulko, Yi-Fei Huang, and Adam Siepel. 2016. “Is a Super-Enhancer Greater than the Sum of Its Parts?” *Nature Genetics* 49 (1): 2–3.

EBI Gene Expression Team – https://www. ebi. ac. uk/about/people/irene-papatheodorou. n.d. “Expression Atlas.” Accessed May 16, 2021. https://www.ebi.ac.uk/gxa/about.html.

Ecker, Joseph R., Wendy A. Bickmore, Inês Barroso, Jonathan K. Pritchard, Yoav Gilad, and Eran Segal. 2012. “Genomics: ENCODE Explained.” *Nature* 489 (7414): 52–55.

Edwards, A. W. F. 2008. “G. H. Hardy (1908) and Hardy-Weinberg Equilibrium.” *Genetics* 179 (3): 1143–50.

Edwards, R., and L. Glass. 2000. “Combinatorial Explosion in Model Gene Networks.” *Chaos* 10 (3): 691–704.

Eeftens, Jorine, and Cees Dekker. 2017. “Catching DNA with Hoops—biophysical Approaches to Clarify the Mechanism of SMC Proteins.” *Nature Structural & Molecular Biology*. https://doi.org/10.1038/nsmb.3507.

Ehret, C. F., and G. de Haller. 1963. “ORIGIN, DEVELOPMENT AND MATURATION OF ORGANELLES AND ORGANELLE SYSTEMS OF THE CELL SURFACE IN PARAMECIUM.” *Journal of Ultrastructure Research* 23 (October): SUPPL6:1–42.

Ehrlich, Melanie, Miguel A. Gama-Sosa, Lan-Hsiang Huang, Rose Marie Midgett, Kenneth C. Kuo, Roy A. McCune, and Charles Gehrke. 1982. “Amount and Distribution of 5-Methylcytosine in Human DNA from Different Types of Tissues or Cells.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/10.8.2709.

Elston, Robert C., Jaya M. Satagopan, and Shuying Sun. 2012. *Statistical Human Genetics: Methods and Protocols*. Humana Press.

ENCODE Project Consortium. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489 (7414): 57–74.

ENCODE Project Consortium, Ewan Birney, John A. Stamatoyannopoulos, Anindya Dutta, Roderic Guigó, Thomas R. Gingeras, Elliott H. Margulies, et al. 2007. “Identification and

Analysis of Functional Elements in 1% of the Human Genome by the ENCODE Pilot Project." *Nature* 447 (7146): 799–816.

ENCODE Project Consortium, Jill E. Moore, Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shoresh, Jessika Adrian, et al. 2020. "Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes." *Nature* 583 (7818): 699–710.

"Ensembl Variant Effect Predictor (VEP)." n.d. Accessed August 11, 2021. https://www.ensembl.org/info/docs/tools/vep.

Enzenbach, Cornelia, Barbara Wicklein, Kerstin Wirkner, and Markus Loeffler. 2019. "Evaluating Selection Bias in a Population-Based Cohort Study with Low Baseline Participation: The LIFE-Adult-Study." *BMC Medical Research Methodology* 19 (1): 135.

Escolar, G., and J. G. White. 1991. "The Platelet Open Canalicular System: A Final Common Pathway." *Blood Cells* 17 (3): 467–85; discussion 486–95.

Eurice GmbH. n.d. "Welcome to IHEC · IHEC." Accessed May 16, 2021. http://ihec-epigenomes.org/.

Eyre-Walker, Adam, and Peter D. Keightley. 2007. "The Distribution of Fitness Effects of New Mutations." *Nature Reviews. Genetics* 8 (8): 610–18.

Fachinetti, Daniele, Joo Seok Han, Moira A. McMahon, Peter Ly, Amira Abdullah, Alex J. Wong, and Don W. Cleveland. 2015. "DNA Sequence-Specific Binding of CENP-B Enhances the Fidelity of Human Centromere Function." *Developmental Cell* 33 (3): 314–27.

Fahed, Akl C., Minxian Wang, Julian R. Homburger, Aniruddh P. Patel, Alexander G. Bick, Cynthia L. Neben, Carmen Lai, et al. 2020. "Polygenic Background Modifies Penetrance of Monogenic Variants for Tier 1 Genomic Conditions." *Nature Communications* 11 (1): 3635.

Fairley, Susan, Ernesto Lowy-Gallego, Emily Perry, and Paul Flicek. 2020. "The International Genome Sample Resource (IGSR) Collection of Open Human Genomic Variation Resources." *Nucleic Acids Research* 48 (D1): D941–47.

Fang, Gang, Diana Munera, David I. Friedman, Anjali Mandlik, Michael C. Chao, Onureena Banerjee, Zhixing Feng, et al. 2012. "Genome-Wide Mapping of Methylated Adenine Residues in Pathogenic Escherichia Coli Using Single-Molecule Real-Time Sequencing." *Nature Biotechnology* 30 (12): 1232–39.

Fanucchi, Stephanie, Ezio T. Fok, Emiliano Dalla, Youtaro Shibayama, Kathleen Börner, Erin Y. Chang, Stoyan Stoychev, et al. 2019. "Immune Genes Are Primed for Robust Transcription by Proximal Long Noncoding RNAs Located in Nuclear Compartments." *Nature Genetics* 51 (1): 138–50.

Feldmann, G., J. Penaud-Laurencin, J. Crassous, and J. P. Benhamou. 1972. "Albumin Synthesis by Human Liver Cells: Its Morphological Demonstration." *Gastroenterology* 63 (6): 1036–48.

Feng, W., M. Madajka, and B. A. Kerr. 2011. "A Novel Role for Platelet Secretion in Angiogenesis: Mediating Bone Marrow–derived Cell Mobilization and Homing." *Blood, The Journal*. https://ashpublications.org/blood/article-abstract/117/14/3893/20504.

Firth, Helen V., Caroline F. Wright, and DDD Study. 2011. "The Deciphering Developmental Disorders (DDD) Study." *Developmental Medicine and Child Neurology* 53 (8): 702–3.

Fischer, Doris, Luciana Porto, Hildegard Stoll, Christof Geisen, and Rolf L. Schloesser. 2010. "Intracerebral Mass Bleeding in a Term Neonate: Manifestation of Hereditary Protein S Deficiency with a New Mutation in the PROS1 Gene." *Neonatology* 98 (4): 337–40.

Fisher, R. A. 1919. "XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance." *Transactions of the Royal Society of Edinburgh*. https://doi.org/10.1017/s0080456800012163.

Fleming, I., J. Bauersachs, and R. Busse. 1996. "Paracrine Functions of the Coronary Vascular Endothelium." *Molecular and Cellular Biochemistry* 157 (1-2): 137–45.

Forrest, Iain S., Kumardeep Chaudhary, Ha My T. Vy, Shantanu Bafna, Daniel M. Jordan, Ghislain Rocheleau, Ruth J. F. Loos, Judy H. Cho, and Ron Do. 2021. “Ancestrally and Temporally Diverse Analysis of Penetrance of Clinical Variants in 72,434 Individuals.” *medRxiv*, March, 2021.03.11.21253430.

Forrest, Lucy R., Christopher L. Tang, and Barry Honig. 2006. “On the Accuracy of Homology Modeling and Sequence Alignment Methods Applied to Membrane Proteins.” *Biophysical Journal* 91 (2): 508–17.

Forsgren, Margaretha, Benny Råden, Marianne Israelsson, Kerstin Larsson, and Lars-Olof Hedén. 1987. “Molecular Cloning and Characterization of a Full-Length cDNA Clone for Human Plasminogen.” *FEBS Letters*. https://doi.org/10.1016/0014-5793(87)81501-6.

Foudi, Adlen, Konrad Hochedlinger, Denille Van Buren, Jeffrey W. Schindler, Rudolf Jaenisch, Vincent Carey, and Hanno Hock. 2009. “Analysis of Histone 2B-GFP Retention Reveals Slowly Cycling Hematopoietic Stem Cells.” *Nature Biotechnology* 27 (1): 84–90.

Fox, Joan E. B. 1996. “Platelet Activation: New Aspects.” *Pathophysiology of Haemostasis and Thrombosis*. https://doi.org/10.1159/000217291.

Fox, Norma E., Rose Chen, Ian Hitchcock, Jennifer Keates-Baleeiro, Haydar Frangoul, and Amy E. Geddis. 2009. “Compound Heterozygous c-Mpl Mutations in a Child with Congenital Amegakaryocytic Thrombocytopenia: Functional Characterization and a Review of the Literature.” *Experimental Hematology*. https://doi.org/10.1016/j.exphem.2009.01.001.

Fox, P. T., K. L. Miller, D. C. Glahn, and P. M. Fox. 2009. “Correspondence of the Brain’s Functional Architecture during Activation and Rest.” *Proceedings of the*. https://www.pnas.org/content/106/31/13040.short.

Frayling, Timothy M., Nicholas J. Timpson, Michael N. Weedon, Eleftheria Zeggini, Rachel M. Freathy, Cecilia M. Lindgren, John R. B. Perry, et al. 2007. “A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity.” *Science* 316 (5826): 889–94.

Freire-Pritchett, Paula, Helen Ray-Jones, Monica Della Rosa, Chris Q. Eijsbouts, William R. Orchard, Steven W. Wingett, Chris Wallace, Jonathan Cairns, Mikhail Spivakov, and Valeriya Malysheva. 2021. “Detecting Chromosomal Interactions in Capture Hi-C Data with CHiCAGO and Companion Tools.” *Nature Protocols*, August, 1–39.

Freire-Pritchett, Paula, Stefan Schoenfelder, Csilla Várnai, Steven W. Wingett, Jonathan Cairns, Amanda J. Collier, Raquel García-Vílchez, et al. 2017. “Global Reorganisation of Cis-Regulatory Units upon Lineage Commitment of Human Embryonic Stem Cells.” *eLife* 6 (March). https://doi.org/10.7554/eLife.21926.

Freson, Kathleen, Rita De Vos, Christine Wittevrongel, Chantal Thys, Johan Defoor, Luc Vanhees, Jos Vermylen, Kathelijne Peerlinck, and Chris Van Geet. 2005. “The TUBB1 Q43P Functional Polymorphism Reduces the Risk of Cardiovascular Disease in Men by Modulating Platelet Function and Structure.” *Blood* 106 (7): 2356–62.

Freson, K., and E. Turro. 2017. “High-Throughput Sequencing Approaches for Diagnosing Hereditary Bleeding and Platelet Disorders.” *Journal of Thrombosis and Haemostasis: JTH* 15 (7): 1262–72.

Fry, Anna, Thomas J. Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E. Allen. 2017. “Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population.” *American Journal of Epidemiology* 186 (9): 1026–34.

Fudenberg, Geoff, David R. Kelley, and Katherine S. Pollard. 2020. “Predicting 3D Genome Folding from DNA Sequence with Akita.” *Nature Methods* 17 (11): 1111–17.

Fulco, Charles P., Mathias Munschauer, Rockwell Anyoha, Glen Munson, Sharon R. Grossman, Elizabeth M. Perez, Michael Kane, Brian Cleary, Eric S. Lander, and Jesse

M. Engreitz. 2016. “Systematic Mapping of Functional Enhancer-Promoter Connections with CRISPR Interference.” *Science* 354 (6313): 769–73.
Furie, Bruce, and Barbara C. Furie. 2008. “Mechanisms of Thrombus Formation.” *The New England Journal of Medicine* 359 (9): 938–49.
Furlan-Magaril, Mayra, Csilla Várnai, Takashi Nagano, and Peter Fraser. 2015. “3D Genome Architecture from Populations to Single Cells.” *Current Opinion in Genetics & Development* 31 (April): 36–41.
Furlong, Eileen E. M., and Michael Levine. 2018. “Developmental Enhancers and Chromosome Topology.” *Science* 361 (6409): 1341–45.
Futuyma, Douglas, and Mark Kirkpatrick. 2017. *Evolution*. Sinauer.
Gaetano, G. de. 2001. “Historical Overview of the Role of Platelets in Hemostasis and Thrombosis.” *Haematologica* 86 (4): 349–56.
Gagniuc, Paul, and Constantin Ionescu-Tirgoviste. 2012. “Eukaryotic Genomes May Exhibit up to 10 Generic Classes of Gene Promoters.” *BMC Genomics* 13 (September): 512.
Gailani, David, and Thomas Renné. 2007. “Intrinsic Pathway of Coagulation and Arterial Thrombosis.” *Arteriosclerosis, Thrombosis, and Vascular Biology*. https://doi.org/10.1161/atvbaha.107.155952.
Gailani, D., and G. J. Broze Jr. 1991. “Factor XI Activation in a Revised Model of Blood Coagulation.” *Science* 253 (5022): 909–12.
Gale, Andrew J. 2011. “Continuing Education Course #2: Current Understanding of Hemostasis.” *Toxicologic Pathology* 39 (1): 273–80.
Gallagher, Michael D., and Alice S. Chen-Plotkin. 2018. “The Post-GWAS Era: From Association to Function.” *American Journal of Human Genetics* 102 (5): 717–30.
Gamazon, E. R., J. A. Badner, L. Cheng, C. Zhang, D. Zhang, N. J. Cox, E. S. Gershon, et al. 2013. “Enrichment of Cis-Regulatory Gene Expression SNPs and Methylation Quantitative Trait Loci among Bipolar Disorder Susceptibility Variants.” *Molecular Psychiatry* 18 (3): 340–46.
Gandrille, S., M. Alhenc-Gelas, P. Gaussem, M. F. Aillaud, E. Dupuy, I. Juhan-Vague, and M. Aiach. 1993. “Five Novel Mutations Located in Exons III and IX of the Protein C Gene in Patients Presenting with Defective Protein C Anticoagulant Activity.” *Blood*. https://doi.org/10.1182/blood.v82.1.159.bloodjournal821159.
Gardner, Anne, and Teresa Davies. 2009. *Human Genetics*. Scion Pub.
Garland, C. J., and K. A. Dora. 2017. “EDH: Endothelium-Dependent Hyperpolarization and Microvascular Signalling.” *Acta Physiologica*. https://doi.org/10.1111/apha.12649.
Garner, Chad. 2007. “Upward Bias in Odds Ratio Estimates from Genome-Wide Association Studies.” *Genetic Epidemiology* 31 (4): 288–95.
Garrod, Archibalde. 1902. “THE INCIDENCE OF ALKAPTONURIA : A STUDY IN CHEMICAL INDIVIDUALITY.” *The Lancet* 160 (4137): 1616–20.
Gaszner, Miklos, and Gary Felsenfeld. 2006. “Insulators: Exploiting Transcriptional and Epigenetic Mechanisms.” *Nature Reviews Genetics*. https://doi.org/10.1038/nrg1925.
Gaudelli, Nicole M., Alexis C. Komor, Holly A. Rees, Michael S. Packer, Ahmed H. Badran, David I. Bryson, and David R. Liu. 2017. “Programmable Base Editing of A•T to G•C in Genomic DNA without DNA Cleavage.” *Nature* 551 (7681): 464–71.
Gaziano, John Michael, John Concato, Mary Brophy, Louis Fiore, Saiju Pyarajan, James Breeling, Stacey Whitbourne, et al. 2016. “Million Veteran Program: A Mega-Biobank to Study Genetic Influences on Health and Disease.” *Journal of Clinical Epidemiology* 70 (February): 214–23.
Genetic Alliance, and District of Columbia Department of Health. 2010. *Classic Mendelian Genetics (Patterns of Inheritance)*. Genetic Alliance.
GenomeAsia100K Consortium. 2019. “The GenomeAsia 100K Project Enables Genetic Discoveries across Asia.” *Nature* 576 (7785): 106–11.

Germain, Marine, Daniel I. Chasman, Hugoline de Haan, Weihong Tang, Sara Lindström, Lu-Chen Weng, Mariza de Andrade, et al. 2015. “Meta-Analysis of 65,734 Individuals Identifies TSPAN15 and SLC44A2 as Two Susceptibility Loci for Venous Thromboembolism.” *American Journal of Human Genetics* 96 (4): 532–42.

Germeshausen, Manuela, and Matthias Ballmaier. 2021. “CAMT-MPL: Congenital Amegakaryocytic Thrombocytopenia Caused by MPL Mutations - Heterogeneity of a Monogenic Disorder - a Comprehensive Analysis of 56 Patients.” *Haematologica* 106 (9): 2439–48.

Ghahremanfard, Farahnaz, Vahid Semnani, Raheb Ghorbani, Farhad Malek, Ali Behzadfar, and Mehrdad Zahmatkesh. 2015. “Effects of Cigarette Smoking on Morphological Features of Platelets in Healthy Men.” *Saudi Medical Journal* 36 (7): 847–50.

Gialeraki, Argyri, Serena Valsami, Theodoros Pittaras, George Panayiotakopoulos, and Marianna Politou. 2018. “Oral Contraceptives and HRT Risk of Thrombosis.” *Clinical and Applied Thrombosis/hemostasis: Official Journal of the International Academy of Clinical and Applied Thrombosis/Hemostasis* 24 (2): 217–25.

Giansily-Blaizot, Muriel, Pavithra M. Rallapalli, Stephen J. Perkins, Geoffrey Kemball-Cook, Daniel J. Hampshire, Keith Gomez, Christopher A. Ludlam, and John H. McVey. 2020. “The EAHAD Blood Coagulation Factor VII Variant Database.” *Human Mutation* 41 (7): 1209–19.

Gibbins, Jonathan M. 2004. “Platelet Adhesion Signalling and the Regulation of Thrombus Formation.” *Journal of Cell Science* 117 (Pt 16): 3415–25.

Gieger, Christian, Aparna Radhakrishnan, Ana Cvejic, Weihong Tang, Eleonora Porcu, Giorgio Pistis, Jovana Serbanovic-Canic, et al. 2011. “New Gene Functions in Megakaryopoiesis and Platelet Formation.” *Nature* 480 (7376): 201–8.

Gierula, Magdalena, Isabelle I. Salles-Crawley, Salvatore Santamaria, Adrienn Teraz-Orosz, James T. B. Crawley, David A. Lane, and Josefin Ahnström. 2019. “The Roles of Factor Va and Protein S in Formation of the Activated Protein C/protein S/factor Va Inactivation Complex.” *Journal of Thrombosis and Haemostasis: JTH* 17 (12): 2056–68.

Gilbert, Luke A., Matthew H. Larson, Leonardo Morsut, Zairan Liu, Gloria A. Brar, Sandra E. Torres, Noam Stern-Ginossar, et al. 2013. “CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes.” *Cell* 154 (2): 442–51.

Gilpatrick, Timothy, Isac Lee, James E. Graham, Etienne Raimondeau, Rebecca Bowen, Andrew Heron, Bradley Downs, Saraswati Sukumar, Fritz J. Sedlazeck, and Winston Timp. 2020. “Targeted Nanopore Sequencing with Cas9-Guided Adapter Ligation.” *Nature Biotechnology* 38 (4): 433–38.

Giorgetti, Luca, Bryan R. Lajoie, Ava C. Carter, Mikael Attia, Ye Zhan, Jin Xu, Chong Jian Chen, et al. 2016. “Structural Organization of the Inactive X Chromosome in the Mouse.” *Nature* 535 (7613): 575–79.

GIS. n.d. “The (near) Complete Sequence of a Human Genome.” Accessed March 9, 2021. https://genomeinformatics.github.io/CHM13v1/.

Gleadall, Nicholas S., Barbera Veldhuisen, Jeremy Gollub, Adam S. Butterworth, John Ord, Christopher J. Penkett, Tiffany C. Timmer, et al. 2020. “Development and Validation of a Universal Blood Donor Genotyping Platform: A Multinational Prospective Study.” *Blood Advances* 4 (15): 3495–3506.

“gnomAD.” n.d. Accessed May 15, 2021. https://gnomad.broadinstitute.org/about.

Goldhaber, Samuel Z., and Ruth B. Morrison. 2002. “Cardiology Patient Pages. Pulmonary Embolism and Deep Vein Thrombosis.” *Circulation* 106 (12): 1436–38.

Goldstein, David B. 2009. “Common Genetic Variation and Human Traits.” *The New England Journal of Medicine* 360 (17): 1696–98.

Goodrich, Julia K., Moriel Singer-Berk, Rachel Son, Abigail Sveden, Jordan Wood, Eleina England, Joanne B. Cole, et al. 2021. “Determinants of Penetrance and Variable

Expressivity in Monogenic Metabolic Conditions across 77,184 Exomes." *Nature Communications* 12 (1): 3505.

Gräf, Stefan, Matthias Haimel, Marta Bleda, Charaka Hadinnapola, Laura Southgate, Wei Li, Joshua Hodgson, et al. 2018. "Identification of Rare Sequence Variation Underlying Heritable Pulmonary Arterial Hypertension." *Nature Communications* 9 (1): 1416.

Grassi, Luigi, Osagie G. Izuogu, Natasha A. N. Jorge, Denis Seyres, Mariona Bustamante, Frances Burden, Samantha Farrow, et al. 2020. "Cell Type Specific Novel lncRNAs and circRNAs in the BLUEPRINT Haematopoietic Transcriptomes Atlas." *Haematologica*, July. https://doi.org/10.3324/haematol.2019.238147.

"GRCh38.p13 - Genome - Assembly - NCBI." n.d. Accessed April 25, 2021. https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39.

Greene, Daniel, NIHR BioResource, Sylvia Richardson, and Ernest Turro. 2017. "A Fast Association Test for Identifying Pathogenic Variants Involved in Rare Diseases." *American Journal of Human Genetics* 101 (1): 104–14.

Green, Eric D., Chris Gunter, Leslie G. Biesecker, Valentina Di Francesco, Carla L. Easter, Elise A. Feingold, Adam L. Felsenfeld, et al. 2020. "Strategic Vision for Improving Human Health at The Forefront of Genomics." *Nature* 586 (7831): 683–92.

Gregor, Anne, Martin Oti, Evelyn N. Kouwenhoven, Juliane Hoyer, Heinrich Sticht, Arif B. Ekici, Susanne Kjaergaard, et al. 2013. "De Novo Mutations in the Genome Organizer CTCF Cause Intellectual Disability." *American Journal of Human Genetics* 93 (1): 124–31.

Grentzmann, G., J. A. Ingram, P. J. Kelly, R. F. Gesteland, and J. F. Atkins. 1998. "A Dual-Luciferase Reporter System for Studying Recoding Signals." *RNA* 4 (4): 479–86.

Grover, Amit, Alejandra Sanjuan-Pla, Supat Thongjuea, Joana Carrelha, Alice Giustacchini, Adriana Gambardella, Iain Macaulay, et al. 2016. "Single-Cell RNA Sequencing Reveals Molecular and Functional Platelet Bias of Aged Haematopoietic Stem Cells." *Nature Communications* 7 (March): 11075.

Grover, Steven P., and Nigel Mackman. 2018. "Tissue Factor: An Essential Mediator of Hemostasis and Trigger of Thrombosis." *Arteriosclerosis, Thrombosis, and Vascular Biology* 38 (4): 709–25.

Grozovsky, Renata, Antonija Jurak Begonja, Kaifeng Liu, Gary Visner, John H. Hartwig, Hervé Falet, and Karin M. Hoffmeister. 2015. "The Ashwell-Morell Receptor Regulates Hepatic Thrombopoietin Production via JAK2-STAT3 Signaling." *Nature Medicine* 21 (1): 47–54.

Grünewald, Julian, Ronghao Zhou, Sara P. Garcia, Sowmya Iyer, Caleb A. Lareau, Martin J. Aryee, and J. Keith Joung. 2019. "Transcriptome-Wide off-Target RNA Editing Induced by CRISPR-Guided DNA Base Editors." *Nature* 569 (7756): 433–37.

GTEx Consortium. 2013. "The Genotype-Tissue Expression (GTEx) Project." *Nature Genetics* 45 (6): 580–85.

———. 2020. "The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues." *Science* 369 (6509): 1318–30.

"GTEx Portal." n.d. Accessed May 16, 2021a. https://gtexportal.org/home/.

———. n.d. Accessed August 8, 2021b. https://gtexportal.org/home/gene/RUNX1.

Guan, Yiting, Chao Zhang, Guoliang Lyu, Xiaoke Huang, Xuebin Zhang, Tenghan Zhuang, Lumeng Jia, et al. 2020. "Senescence-Activated Enhancer Landscape Orchestrates the Senescence-Associated Secretory Phenotype in Murine Fibroblasts." *Nucleic Acids Research* 48 (19): 10909–23.

Guelen, Lars, Ludo Pagie, Emilie Brasset, Wouter Meuleman, Marius B. Faza, Wendy Talhout, Bert H. Eussen, et al. 2008. "Domain Organization of Human Chromosomes Revealed by Mapping of Nuclear Lamina Interactions." *Nature* 453 (7197): 948–51.

Guo, Michael H., Lacey Plummer, Yee-Ming Chan, Joel N. Hirschhorn, and Margaret F.

Lippincott. 2018. “Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data.” *American Journal of Human Genetics* 103 (4): 522–34.

Gupta, Rajat M., Joseph Hadaya, Aditi Trehan, Seyedeh M. Zekavat, Carolina Roselli, Derek Klarin, Connor A. Emdin, et al. 2017. “A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression.” *Cell* 170 (3): 522–33.e15.

Gusev, Alexander, S. Hong Lee, Gosia Trynka, Hilary Finucane, Bjarni J. Vilhjálmsson, Han Xu, Chongzhi Zang, et al. 2014. “Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases.” *American Journal of Human Genetics* 95 (5): 535–52.

Haddad, Yazan, Vojtech Adam, and Zbynek Heger. 2020. “Ten Quick Tips for Homology Modeling of High-Resolution Protein 3D Structures.” *PLoS Computational Biology* 16 (4): e1007449.

Halliwell, Barry, and John M. C. Gutteridge. 2015. *Free Radicals in Biology and Medicine*. Oxford University Press.

Ha, Minju. 2020. “Transcription Boosting by Nuclear Speckles.” *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/s41580-019-0203-6.

Hamosh, A. 2002. “Online Mendelian Inheritance in Man (OMIM), a Knowledgebase of Human Genes and Genetic Disorders.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/30.1.52.

Hannan, Nicholas R. F., Charis-Patricia Segeritz, Thomas Touboul, and Ludovic Vallier. 2013. “Production of Hepatocyte-like Cells from Human Pluripotent Stem Cells.” *Nature Protocols* 8 (2): 430–37.

Hansen, Anders S., Iryna Pustova, Claudia Cattoglio, Robert Tjian, and Xavier Darzacq. 2017. “CTCF and Cohesin Regulate Chromatin Loop Stability with Distinct Dynamics.” *eLife* 6 (May). https://doi.org/10.7554/eLife.25776.

Hardy, G. H. 1908. “MENDELIAN PROPORTIONS IN A MIXED POPULATION.” *Science*. https://doi.org/10.1126/science.28.706.49.

Hari Dass, Shantala A., Kathryn McCracken, Irina Pokhvisneva, Lawrence M. Chen, Elika Garg, Thao T. T. Nguyen, Zihan Wang, et al. 2019. “A Biologically-Informed Polygenic Score Identifies Endophenotypes and Clinical Conditions Associated with the Insulin Receptor Function on Specific Brain Regions.” *EBioMedicine* 42 (April): 188–202.

Harr, Jennifer C., Teresa Romeo Luperchio, Xianrong Wong, Erez Cohen, Sarah J. Wheelan, and Karen L. Reddy. 2015. “Directed Targeting of Chromatin to the Nuclear Lamina Is Mediated by Chromatin State and A-Type Lamins.” *The Journal of Cell Biology* 208 (1): 33–52.

Harst, Pim van der, Weihua Zhang, Irene Mateo Leach, Augusto Rendon, Niek Verweij, Joban Sehmi, Dirk S. Paul, et al. 2012. “Seventy-Five Genetic Loci Influencing the Human Red Blood Cell.” *Nature* 492 (7429): 369–75.

Hartl, Daniel L., and Andrew G. Clark. 1997. *Principles of Population Genetics*. Sinauer Associates.

Hatton, C. S., A. O. Wilkie, H. C. Drysdale, W. G. Wood, M. A. Vickers, J. Sharpe, H. Ayyub, I. M. Pretorius, V. J. Buckle, and D. R. Higgs. 1990. “Alpha-Thalassemia Caused by a Large (62 Kb) Deletion Upstream of the Human Alpha Globin Gene Cluster.” *Blood* 76 (1): 221–27.

Haudenschild, Christian C. 1984. “Morphology of Vascular Endothelial Cells in Culture.” In *Biology of Endothelial Cells*, edited by Eric A. Jaffe, 129–40. Boston, MA: Springer US.

Haver, V. M., and A. R. Gear. 1981. “Functional Fractionation of Platelets.” *The Journal of Laboratory and Clinical Medicine* 97 (2): 187–204.

Hawiger, J. 1987. “Formation and Regulation of Platelet and Fibrin Hemostatic Plug.” *Human*

Pathology 18 (2): 111–22.
Haworth, Simon, Ruth Mitchell, Laura Corbin, Kaitlin H. Wade, Tom Dudding, Ashley Budu-Aggrey, David Carslake, et al. 2019. “Apparent Latent Structure within the UK Biobank Sample Has Implications for Epidemiological Analysis.” *Nature Communications* 10 (1): 333.
Hay, Deborah, Jim R. Hughes, Christian Babbs, James O. J. Davies, Bryony J. Graham, Lars Hanssen, Mira T. Kassouf, et al. 2016. “Genetic Dissection of the α-Globin Super-Enhancer in Vivo.” *Nature Genetics* 48 (8): 895–903.
Heide-Jørgensen, Uffe, Kasper Adelborg, Johnny Kahlert, Henrik Toft Sørensen, and Lars Pedersen. 2018. “Sampling Strategies for Selecting General Population Comparison Cohorts.” *Clinical Epidemiology* 10 (September): 1325–37.
Heijnen, Harry F. G., Najet Debili, William Vainchencker, Janine Breton-Gorius, Hans J. Geuze, and Jan J. Sixma. 1998. “Multivesicular Bodies Are an Intermediate Stage in the Formation of Platelet α-Granules.” *Blood, The Journal of the American Society of Hematology* 91 (7): 2313–25.
Hernando-Herraez, Irene, Brendan Evano, Thomas Stubbs, Pierre-Henri Commere, Marc Jan Bonder, Stephen Clark, Simon Andrews, Shahragim Tajbakhsh, and Wolf Reik. 2019. “Ageing Affects DNA Methylation Drift and Transcriptional Cell-to-Cell Variability in Mouse Muscle Stem Cells.” *Nature Communications* 10 (1): 4361.
Herrera-Rivero, Marisol, Monika Stoll, Jana-Charlotte Hegenbarth, Frank Rühle, Verena Limperger, Ralf Junker, André Franke, et al. 2021. “Single- and Multimarker Genome-Wide Scans Evidence Novel Genetic Risk Modifiers for Venous Thromboembolism.” *Thrombosis and Haemostasis*, February. https://doi.org/10.1055/s-0041-1723988.
Hershey, A. D., and M. Chase. 1952. “Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage.” *The Journal of General Physiology* 36 (1): 39–56.
Herskowitz, I. 1987. “Functional Inactivation of Genes by Dominant Negative Mutations.” *Nature* 329 (6136): 219–22.
Herzel, Lydia, Diana S. M. Ottoz, Tara Alpert, and Karla M. Neugebauer. 2017. “Splicing and Transcription Touch Base: Co-Transcriptional Spliceosome Assembly and Function.” *Nature Reviews. Molecular Cell Biology* 18 (10): 637–50.
Hewitt, Robert, and Peter Watson. 2013. “Defining Biobank.” *Biopreservation and Biobanking* 11 (5): 309–15.
“HGMD® Home Page.” n.d. Accessed May 15, 2021. http://www.hgmd.cf.ac.uk/ac/index.php.
Higgs, D. R. 2013. “The Molecular Basis of -Thalassemia.” *Cold Spring Harbor Perspectives in Medicine*. https://doi.org/10.1101/cshperspect.a011718.
Hillman, Robert S., Kenneth A. Ault, Kenneth Ault, and Henry Rinder. 2005. *Hematology in Clinical Practice*. McGraw-Hill Companies,Incorporated.
Hill, William G., Michael E. Goddard, and Peter M. Visscher. 2008. “Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits.” *PLoS Genetics* 4 (2): e1000008.
Hitchcock, Ian S., Maximillian Hafer, Veena Sangkhae, and Julie A. Tucker. 2021. “The Thrombopoietin Receptor: Revisiting the Master Regulator of Platelet Production.” *Platelets*, June, 1–9.
Hitchcock, Ian S., and Kenneth Kaushansky. 2014. “Thrombopoietin from Beginning to End.” *British Journal of Haematology* 165 (2): 259–68.
Hnisz, Denes, Brian J. Abraham, Tong Ihn Lee, Ashley Lau, Violaine Saint-André, Alla A. Sigova, Heather A. Hoke, and Richard A. Young. 2013. “Super-Enhancers in the Control of Cell Identity and Disease.” *Cell* 155 (4): 934–47.
Hnisz, Denes, Daniel S. Day, and Richard A. Young. 2016. “Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control.” *Cell* 167 (5): 1188–1200.

Hoffbrand, Victor A., Daniel Catovsky, and Edward G. D. Tuddenham. 2008. *Postgraduate Haematology*. John Wiley & Sons.

Hoffman, M., and D. M. Monroe 3rd. 2001. “A Cell-Based Model of Hemostasis.” *Thrombosis and Haemostasis* 85 (6): 958–65.

Hommelsheim, Carl Maximilian, Lamprinos Frantzeskakis, Mengmeng Huang, and Bekir Ülker. 2014. “PCR Amplification of Repetitive DNA: A Limitation to Genome Editing Technologies and Many Other Applications.” *Scientific Reports* 4 (May): 5052.

Hopkin, Michael. 2008. “Biodiversity: Frozen Futures.” *Nature* 452 (7186): 404–5.

Horta, A., K. Monahan, E. Bashkirova, and S. Lomvardas. 2018. “Cell Type-Specific Interchromosomal Interactions as a Mechanism for Transcriptional Diversity.” *bioRxiv*. https://doi.org/10.1101/287532.

Hotchkiss, R. D. 1948. “The Quantitative Separation of Purines, Pyrimidines, and Nucleosides by Paper Chromatography.” *The Journal of Biological Chemistry* 175 (1): 315–32.

Ho-Tin-Noé, Benoit, Yacine Boulaftali, and Eric Camerer. 2018. “Platelets and Vascular Integrity: How Platelets Prevent Bleeding in Inflammation.” *Blood* 131 (3): 277–88.

Huang, Suming, Michael D. Litt, and Cynthia Ann Blakey. 2015. *Epigenetic Gene Expression and Regulation*. Academic Press.

Hulle, T. van der, T. van der Hulle, M. Tan, P. L. den Exter, G. C. Mol, A. Iglesias del Sol, M. A. van de Ree, M. V. Huisman, and F. A. Klok. 2013. “Selective D-Dimer Testing for the Diagnosis of Acute Deep Vein Thrombosis: A Validation Study.” *Journal of Thrombosis and Haemostasis*. https://doi.org/10.1111/jth.12419.

Huntley, Stuart, Daniel M. Baggott, Aaron T. Hamilton, Mary Tran-Gyamfi, Shan Yang, Joomyeong Kim, Laurie Gordon, Elbert Branscomb, and Lisa Stubbs. 2006. “A Comprehensive Catalog of Human KRAB-Associated Zinc Finger Genes: Insights into the Evolutionary History of a Large Family of Transcriptional Repressors.” *Genome Research* 16 (5): 669–77.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. 2020. “Pan-Cancer Analysis of Whole Genomes.” *Nature* 578 (7793): 82–93.

“Immense Discovery Power for Deeper Insights.” n.d. Accessed May 9, 2021. https://www.illumina.com/systems/sequencing-platforms/novaseq/specifications.html.

Inoue, Fumitaka, and Nadav Ahituv. 2015. “Decoding Enhancers Using Massively Parallel Reporter Assays.” *Genomics* 106 (3): 159–64.

International Human Genome Sequencing Consortium. 2004. “Finishing the Euchromatic Sequence of the Human Genome.” *Nature* 431 (7011): 931–45.

Ivanov, Ivan, Ingrid M. Verhamme, Mao-Fu Sun, Bassem Mohammed, Qiufang Cheng, Anton Matafonov, S. Kent Dickeson, Kusumam Joseph, Allen P. Kaplan, and David Gailani. 2020. “Protease Activity in Single-Chain Prekallikrein.” *Blood* 135 (8): 558–67.

Izak, Marina, and James B. Bussel. 2014. “Management of Thrombocytopenia.” *F1000prime Reports* 6 (June): 45.

Jarvik, G. P. 1998. “Complex Segregation Analyses: Uses and Limitations.” *American Journal of Human Genetics*.

Javierre, Biola M., Oliver S. Burren, Steven P. Wilder, Roman Kreuzhuber, Steven M. Hill, Sven Sewitz, Jonathan Cairns, et al. 2016. “Lineage-Specific Genome Architecture Links Enhancers and Non-Coding Disease Variants to Target Gene Promoters.” *Cell* 167 (5): 1369–84.e19.

Jeffares, Daniel C., Tobias Mourier, and David Penny. 2006. “The Biology of Intron Gain and Loss.” *Trends in Genetics: TIG* 22 (1): 16–22.

Jeggo, P. A. 1998. “DNA Breakage and Repair.” *Advances in Genetics* 38: 185–218.

Jelkmann, W. 2001. “The Role of the Liver in the Production of Thrombopoietin Compared with Erythropoietin.” *European Journal of Gastroenterology & Hepatology* 13 (7):

791–801.

Ji, Hong, Lauren I. R. Ehrlich, Jun Seita, Peter Murakami, Akiko Doi, Paul Lindau, Hwajin Lee, et al. 2010. “Comprehensive Methylome Map of Lineage Commitment from Haematopoietic Progenitors.” *Nature* 467 (7313): 338–42.

Jinek, Martin, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A. Doudna, and Emmanuelle Charpentier. 2012. “A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity.” *Science* 337 (6096): 816–21.

Jin, Fulai, Yan Li, Jesse R. Dixon, Siddarth Selvaraj, Zhen Ye, Ah Young Lee, Chia-An Yen, Anthony D. Schmitt, Celso A. Espinoza, and Bing Ren. 2013. “A High-Resolution Map of the Three-Dimensional Chromatin Interactome in Human Cells.” *Nature* 503 (7475): 290–94.

Jouanna, Jacques. 2012. “The Legacy of the Hippocratic Treatise the Nature of Man: The Theory of the Four Humours.” In *Greek Medicine from Hippocrates to Galen*, 335–59. Brill.

Joung, J. Keith, and Jeffry D. Sander. 2013. “TALENs: A Widely Applicable Technology for Targeted Genome Editing.” *Nature Reviews. Molecular Cell Biology* 14 (1): 49–55.

Juengst, Eric T., Jennifer R. Fishman, Michelle L. McGowan, and Richard A. Settersten Jr. 2014. “Serving Epigenetics before Its Time.” *Trends in Genetics: TIG* 30 (10): 427–29.

Jung, Inkyung, Anthony Schmitt, Yarui Diao, Andrew J. Lee, Tristin Liu, Dongchan Yang, Catherine Tan, et al. 2019. “A Compendium of Promoter-Centered Long-Range Chromatin Interactions in the Human Genome.” *Nature Genetics*. https://doi.org/10.1038/s41588-019-0494-8.

Junion, Guillaume, Mikhail Spivakov, Charles Girardot, Martina Braun, E. Hilary Gustafson, Ewan Birney, and Eileen E. M. Furlong. 2012. “A Transcription Factor Collective Defines Cardiac Cell Fate and Reflects Lineage History.” *Cell* 148 (3): 473–86.

Kalia, Sarah S., Kathy Adelman, Sherri J. Bale, Wendy K. Chung, Christine Eng, James P. Evans, Gail E. Herman, et al. 2017. “Recommendations for Reporting of Secondary Findings in Clinical Exome and Genome Sequencing, 2016 Update (ACMG SF v2.0): A Policy Statement of the American College of Medical Genetics and Genomics.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 19 (2): 249–55.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. “The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans.” *Nature* 581 (7809): 434–43.

Karczewski, Konrad, Matthew Solomonson, Katherine R. Chao, Julia K. Goodrich, Grace Tiao, Wenhan Lu, Bridget Riley-Gillis, et al. 2021. “Systematic Single-Variant and Gene-Based Association Testing of 3,700 Phenotypes in 281,850 UK Biobank Exomes.” medRxiv. https://doi.org/10.1101/2021.06.19.21259117.

Karpatkin, S. 1978. “Heterogeneity of Human Platelets. VI. Correlation of Platelet Function with Platelet Volume.” *Blood* 51 (2): 307–16.

Kaushansky, Kenneth, Marshall A. Lichtman, Josef Prchal, Marcel M. Levi, Oliver W. Press, Linda J. Burns, and Michael Caligiuri. 2015. *Williams Hematology, 9E*. McGraw Hill Professional.

Kearns, Nicola A., Hannah Pham, Barbara Tabak, Ryan M. Genga, Noah J. Silverstein, Manuel Garber, and René Maehr. 2015. “Functional Annotation of Native Enhancers with a Cas9–histone Demethylase Fusion.” *Nature Methods* 12 (5): 401–3.

Kelsey, Gavin, Oliver Stegle, and Wolf Reik. 2017. “Single-Cell Epigenomics: Recording the Past and Predicting the Future.” *Science* 358 (6359): 69–75.

Kentsis, A., R. Anewalt, A. Ganguly, J. B. Allen, and E. J. Neufeld. 2009. “Discordant Haemophilia A in Male Siblings due to a de Novo Mutation on a Familial Missense Mutant Allele.” *Haemophilia: The Official Journal of the World Federation of Hemophilia*

15 (4): 971–72.
Key, Nigel S., Michael Makris, and David Lillicrap. 2017. *Practical Hemostasis and Thrombosis*. John Wiley & Sons.
Khaitovich, Philipp, Wolfgang Enard, Michael Lachmann, and Svante Pääbo. 2006. “Evolution of Primate Gene Expression.” *Nature Reviews. Genetics* 7 (9): 693–702.
Khalafallah, Alhossain, Chris Jarvis, Michael Morse, Abdul-Majeed Albarzan, Phoebe Stewart, Gerald Bates, Robert Hayes, Iain Robertson, David Seaton, and Terry Brain. 2014. “Evaluation of the Innovance D-Dimer Assay for the Diagnosis of Disseminated Intravascular Coagulopathy in Different Clinical Settings.” *Clinical and Applied Thrombosis/Hemostasis*. https://doi.org/10.1177/1076029612454936.
Khan, Salwa, and Joseph D. Dickerman. 2006. “Hereditary Thrombophilia.” *Thrombosis Journal* 4 (September): 15.
Khan, Shawez, Federico Taverna, Katerina Rohlenova, Lucas Treps, Vincent Geldhof, Laura de Rooij, Liliana Sokol, et al. 2019. “EndoDB: A Database of Endothelial Cell Transcriptomics Data.” *Nucleic Acids Research* 47 (D1): D736–44.
Khera, Amit V., Mark Chaffin, Krishna G. Aragam, Mary E. Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, et al. 2018. “Genome-Wide Polygenic Scores for Common Diseases Identify Individuals with Risk Equivalent to Monogenic Mutations.” *Nature Genetics* 50 (9): 1219–24.
Kheterpal, Indu, James R. Scherer, Steven M. Clark, Arun Radhakrishnan, Jingyue Ju, Charles L. Ginther, George F. Sensabaugh, and Richard A. Mathies. 1996. “DNA Sequencing Using a Four-Color Confocal Fluorescence Capillary Array Scanner.” *Electrophoresis*. https://doi.org/10.1002/elps.1150171209.
Kilpinen, Helena, Angela Goncalves, Andreas Leha, Vackar Afzal, Sofie Ashford, Sendu Bala, Dalila Bensaddek, et al. 2017. “Common Genetic Variation Drives Molecular Heterogeneity in Human IPSCs.” https://doi.org/10.1101/055160.
Kim, Jiah, Kyu Young Han, Nimish Khanna, Taekjip Ha, and Andrew S. Belmont. 2019. “Nuclear Speckle Fusion via Long-Range Directional Motion Regulates Speckle Morphology after Transcriptional Inhibition.” *Journal of Cell Science* 132 (8). https://doi.org/10.1242/jcs.226563.
Kimura, Kouichi, Ai Wakamatsu, Yutaka Suzuki, Toshio Ota, Tetsuo Nishikawa, Riu Yamashita, Jun-Ichi Yamamoto, et al. 2006. “Diversification of Transcriptional Modulation: Large-Scale Identification and Characterization of Putative Alternative Promoters of Human Genes.” *Genome Research* 16 (1): 55–65.
King, Emily A., J. Wade Davis, and Jacob F. Degner. 2019. “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 Genetics* 15 (12): e1008489.
King, Sarah M., and Guy L. Reed. 2002. “Development of Platelet Secretory Granules.” *Seminars in Cell & Developmental Biology*. https://doi.org/10.1016/s1084952102000599.
Klann, T. S., J. B. Black, M. Chellappan, A. Safi, and L. Song. 2017. “CRISPR–Cas9 Epigenome Editing Enables High-Throughput Screening for Functional Regulatory Elements in the Human Genome.” *Nature*. https://idp.nature.com/authorize/casa?redirect_uri=https://www.nature.com/articles/nbt.3853.pdf%3Forigin%3Dppub&casa_token=F1tR2zFsQOwAAAAA:2mjVEOlVT2txEoUe9DwqKWmeFZpDNiX2SznzKJg5i4Di2qmlBKKZBNYJOhOTWHB4WQ_oWpBGSankgJC4Xs0.
Klarin, Derek, Emma Busenkell, Renae Judy, Julie Lynch, Michael Levin, Jeffery Haessler, Krishna Aragam, et al. 2019. “Genome-Wide Association Analysis of Venous Thromboembolism Identifies New Risk Loci and Genetic Overlap with Arterial Vascular

Disease.” *Nature Genetics* 51 (11): 1574–79.
Klarin, Derek, Connor A. Emdin, Pradeep Natarajan, Mark F. Conrad, INVENT Consortium, and Sekar Kathiresan. 2017. “Genetic Analysis of Venous Thromboembolism in UK Biobank Identifies the ZFPM2 Locus and Implicates Obesity as a Causal Risk Factor.” *Circulation. Cardiovascular Genetics* 10 (2). https://doi.org/10.1161/CIRCGENETICS.116.001643.
Kleinberger, Jeffrey, Kristin A. Maloney, Toni I. Pollin, and Linda Jo Bone Jeng. 2016. “An Openly Available Online Tool for Implementing the ACMG/AMP Standards and Guidelines for the Interpretation of Sequence Variants.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 18 (11): 1165.
Klemm, Sandy L., Zohar Shipony, and William J. Greenleaf. 2019. “Chromatin Accessibility and the Regulatory Epigenome.” *Nature Reviews. Genetics* 20 (4): 207–20.
Klimiankou, Maksim, Olga Klimenkova, Murat Uenalan, Alexander Zeidler, Sabine Mellor-Heineke, Siarhei Kandabarau, Julia Skokowa, Cornelia Zeidler, and Karl Welte. 2015. “GM-CSF Stimulates Granulopoiesis in a Congenital Neutropenia Patient with Loss-of-Function Biallelic Heterozygous CSF3R Mutations.” *Blood* 126 (15): 1865–67.
Klug, A., and D. Rhodes. 1987. “Zinc Fingers: A Novel Protein Fold for Nucleic Acid Recognition.” *Cold Spring Harbor Symposia on Quantitative Biology* 52: 473–82.
Köhler, Sebastian, Michael Gargano, Nicolas Matentzoglu, Leigh C. Carmody, David Lewis-Smith, Nicole A. Vasilevsky, Daniel Danis, et al. 2021. “The Human Phenotype Ontology in 2021.” *Nucleic Acids Research* 49 (D1): D1207–17.
Kondo, M., I. L. Weissman, and K. Akashi. 1997. “Identification of Clonogenic Common Lymphoid Progenitors in Mouse Bone Marrow.” *Cell* 91 (5): 661–72.
Konecny, Filip. 2009. “Inherited Trombophilic States and Pulmonary Embolism.” *Journal of Research in Medical Sciences: The Official Journal of Isfahan University of Medical Sciences* 14 (1): 43–56.
Kong, Augustine, Gudmar Thorleifsson, Michael L. Frigge, Bjarni J. Vilhjalmsson, Alexander I. Young, Thorgeir E. Thorgeirsson, Stefania Benonisdottir, et al. 2018. “The Nature of Nurture: Effects of Parental Genotypes.” *Science* 359 (6374): 424–28.
Konrad, Enrico D. H., Niels Nardini, Almuth Caliebe, Inga Nagel, Dana Young, Gabriella Horvath, Stephanie L. Santoro, et al. 2019. “CTCF Variants in 39 Individuals with a Variable Neurodevelopmental Disorder Broaden the Mutational and Clinical Spectrum.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 21 (12): 2723–33.
Kozmin, S., G. Slezak, A. Reynaud-Angelin, C. Elie, Y. de Rycke, S. Boiteux, and E. Sage. 2005. “UVA Radiation Is Highly Mutagenic in Cells That Are Unable to Repair 7,8-Dihydro-8-Oxoguanine in Saccharomyces Cerevisiae.” *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.0504497102.
Krijger, Peter H. L., Geert Geeven, Valerio Bianchi, Catharina R. E. Hilvering, and Wouter de Laat. 2020. “4C-Seq from Beginning to End: A Detailed Protocol for Sample Preparation and Data Analysis.” *Methods* 170 (January): 17–32.
Krijger, Peter Hugo Lodewijk, Bruno Di Stefano, Elzo de Wit, Francesco Limone, Chris van Oevelen, Wouter de Laat, and Thomas Graf. 2016. “Cell-of-Origin-Specific 3D Genome Structure Acquired during Somatic Cell Reprogramming.” *Cell Stem Cell* 18 (5): 597–610.
Krijger, Peter Hugo Lodewijk, and Wouter de Laat. 2016. “Regulation of Disease-Associated Gene Expression in the 3D Genome.” *Nature Reviews. Molecular Cell Biology* 17 (12): 771–82.
Krüger-Genge, Anne, Anna Blocki, Ralf-Peter Franke, and Friedrich Jung. 2019. “Vascular Endothelial Cell Biology: An Update.” *International Journal of Molecular Sciences* 20 (18). https://doi.org/10.3390/ijms20184411.

Kuchenbaecker, Karoline B., Lesley McGuffog, Daniel Barrowdale, Andrew Lee, Penny Soucy, Joe Dennis, Susan M. Domchek, et al. 2017. “Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers.” *Journal of the National Cancer Institute* 109 (7). https://doi.org/10.1093/jnci/djw302.

Kujovich, Jody Lynn. 2011. “Factor V Leiden Thrombophilia.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 13 (1): 1–16.

Kumar, Suresh, Viswanathan Chinnusamy, and Trilochan Mohapatra. 2018. “Epigenetics of Modified DNA Bases: 5-Methylcytosine and Beyond.” *Frontiers in Genetics*. https://doi.org/10.3389/fgene.2018.00640.

Kurt, Ibrahim C., Ronghao Zhou, Sowmya Iyer, Sara P. Garcia, Bret R. Miller, Lukas M. Langner, Julian Grünewald, and J. Keith Joung. 2021. “CRISPR C-to-G Base Editors for Inducing Targeted DNA Transversions in Human Cells.” *Nature Biotechnology* 39 (1): 41–46.

Kwasnieski, Jamie C., Christopher Fiore, Hemangi G. Chaudhari, and Barak A. Cohen. 2014. “High-Throughput Functional Testing of ENCODE Segmentation Predictions.” *Genome Research* 24 (10): 1595–1602.

Labant, Maryann. 2012. “Biobank Diversity Facilitates Drug & Diagnostic Development.” *Genetic Engineering & Biotechnology News* 32 (2): 42–44.

Lagerkvist, U. 1978. “‘Two out of Three’: An Alternative Method for Codon Reading.” *Proceedings of the National Academy of Sciences of the United States of America* 75 (4): 1759–62.

Lakich, D., H. H. Kazazian Jr, S. E. Antonarakis, and J. Gitschier. 1993. “Inversions Disrupting the Factor VIII Gene Are a Common Cause of Severe Haemophilia A.” *Nature Genetics* 5 (3): 236–41.

Läll, Kristi, Reedik Mägi, Andrew Morris, Andres Metspalu, and Krista Fischer. 2017. “Personalized Risk Prediction for Type 2 Diabetes: The Potential of Genetic Risk Scores.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 19 (3): 322–29.

Lambert, Samuel A., Laurent Gil, Simon Jupp, Scott C. Ritchie, Yu Xu, Annalisa Buniello, Aoife McMahon, et al. 2021. “The Polygenic Score Catalog as an Open Database for Reproducibility and Systematic Evaluation.” *Nature Genetics* 53 (4): 420–25.

Lam, Daniel D., Flavio S. J. de Souza, Sofia Nasif, Miho Yamashita, Rodrigo López-Leal, Veronica Otero-Corchon, Kana Meece, et al. 2015. “Partially Redundant Enhancers Cooperatively Maintain Mammalian Pomc Expression Above a Critical Functional Threshold.” *PLOS Genetics*. https://doi.org/10.1371/journal.pgen.1004935.

Lamond, Angus I., and David L. Spector. 2003. “Nuclear Speckles: A Model for Nuclear Organelles.” *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm1172.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature* 409 (6822): 860–921.

Lander, E. S., and N. J. Schork. 1994. “Genetic Dissection of Complex Traits.” *Science* 265 (5181): 2037–48.

Landt, Stephen G., Georgi K. Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E. Bernstein, et al. 2012. “ChIP-Seq Guidelines and Practices of the ENCODE and modENCODE Consortia.” *Genome Research* 22 (9): 1813–31.

Lange, Kenneth. 2012. *Mathematical and Statistical Methods for Genetic Analysis*. Springer Science & Business Media.

Larson, Matthew H., Luke A. Gilbert, Xiaowo Wang, Wendell A. Lim, Jonathan S. Weissman, and Lei S. Qi. 2013. “CRISPR Interference (CRISPRi) for Sequence-Specific Control of

Gene Expression.” *Nature Protocols* 8 (11): 2180–96.
Laugsch, Magdalena, Michaela Bartusel, Rizwan Rehimi, Hafiza Alirzayeva, Agathi Karaolidou, Giuliano Crispatzu, Peter Zentis, et al. 2019. “Modeling the Pathological Long-Range Regulatory Effects of Human Structural Variation with Patient-Specific hiPSCs.” *Cell Stem Cell* 24 (5): 736–52.e12.
Laurenti, Elisa, Sergei Doulatov, Sasan Zandi, Ian Plumb, Jing Chen, Craig April, Jian-Bing Fan, and John E. Dick. 2013. “The Transcriptional Architecture of Early Human Hematopoiesis Identifies Multilevel Control of Lymphoid Commitment.” *Nature Immunology* 14 (7): 756–63.
Laurenti, Elisa, and Berthold Göttgens. 2018. “From Haematopoietic Stem Cells to Complex Differentiation Landscapes.” *Nature* 553 (7689): 418–26.
Lefrançais, Emma, Guadalupe Ortiz-Muñoz, Axelle Caudrillier, Beñat Mallavia, Fengchun Liu, David M. Sayah, Emily E. Thornton, et al. 2017. “The Lung Is a Site of Platelet Biogenesis and a Reservoir for Haematopoietic Progenitors.” *Nature* 544 (7648): 105–9.
Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O’Donnell-Luria, et al. 2016. “Analysis of Protein-Coding Genetic Variation in 60,706 Humans.” *Nature* 536 (7616): 285–91.
Lelli, Katherine M., Matthew Slattery, and Richard S. Mann. 2012. “Disentangling the Many Layers of Eukaryotic Transcriptional Regulation.” *Annual Review of Genetics* 46 (August): 43–68.
Lentaigne, Claire, Kathleen Freson, Michael A. Laffan, Ernest Turro, Willem H. Ouwehand, and BRIDGE-BPD Consortium and the ThromboGenomics Consortium. 2016. “Inherited Platelet Disorders: Toward DNA-Based Diagnosis.” *Blood* 127 (23): 2814–23.
Lentaigne, Claire, Daniel Greene, Suthesh Sivapalaratnam, Remi Favier, Denis Seyres, Chantal Thys, Luigi Grassi, Sarah Mangles, Keith Sibson, Matthew Stubbs, and Others. 2019. “Germline Mutations in the Transcription Factor IKZF5 Cause Thrombocytopenia.” *Blood, The Journal of the American Society of Hematology* 134 (23): 2070–81.
Lentaigne, Claire, Daniel Greene, Suthesh Sivapalaratnam, Remi Favier, Denis Seyres, Chantal Thys, Luigi Grassi, Sarah Mangles, Keith Sibson, Matthew Stubbs, Frances Burden, et al. 2019. “Germline Mutations in the Transcription Factor IKZF5 Cause Thrombocytopenia.” *Blood* 134 (23): 2070–81.
Levine, P. H. 1973. “An Acute Effect of Cigarette Smoking on Platelet Function. A Possible Link between Smoking and Arterial Thrombosis.” *Circulation* 48 (3): 619–23.
Levin, Jack. 1977. “Blood Coagulation in the Horseshoe CRAB (Limulus Polyphemus): A Model for Mammalian Coagulation and Hemostasis.” *VIth International Congress on Thrombosis and Haemostasis*. https://doi.org/10.1055/s-0039-1682807.
Levin, Jack, and J. David Bessman. 1983. “The Inverse Relation between Platelet Volume and Platelet Number: Abnormalities in Hematologic Disease and Evidence That Platelet Size Does Not Correlate with Platelet Age.” *The Journal of Laboratory and Clinical Medicine* 101 (2): 295–307.
Levin, J., and J. D. Bessman. 1983. “The Inverse Relation between Platelet Volume and Platelet Number: Abnormalities in Hematologic Disease and Evidence That Platelet Size Does Not Correlate with ….” *The Journal of Laboratory and Clinical Medicine*. https://www.translationalres.com/article/0022-2143(83)90188-9/fulltext.
Liang, Minggao, Asim Soomro, Subia Tasneem, Luis E. Abatti, Azad Alizada, Xuefei Yuan, Liis Uusküla-Reimand, et al. 2020. “Enhancer-Gene Rewiring in the Pathogenesis of Quebec Platelet Disorder.” *Blood* 136 (23): 2679–90.
Lichou, Florence, and Gosia Trynka. 2020. “Functional Studies of GWAS Variants Are Gaining Momentum.” *Nature Communications* 11 (1): 6283.

Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. “Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.” *Science* 326 (5950): 289–93.

Li, Guoliang, Xiaoan Ruan, Raymond K. Auerbach, Kuljeet Singh Sandhu, Meizhen Zheng, Ping Wang, Huay Mei Poh, et al. 2012. “Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation.” *Cell* 148 (1-2): 84.

Li, Kailong, Yuxuan Liu, Hui Cao, Yuannyu Zhang, Zhimin Gu, Xin Liu, Andy Yu, et al. 2020. “Interrogation of Enhancer Function by Enhancer-Targeting CRISPR Epigenetic Editing.” *Nature Communications* 11 (1): 485.

Lindström, Sara, Lu Wang, Erin N. Smith, William Gordon, Astrid van Hylckama Vlieg, Mariza de Andrade, Jennifer A. Brody, et al. 2019. “Genomic and Transcriptomic Association Studies Identify 16 Novel Susceptibility Loci for Venous Thromboembolism.” *Blood* 134 (19): 1645–57.

Liu, Xin, Yuannyu Zhang, Yong Chen, Mushan Li, Feng Zhou, Kailong Li, Hui Cao, et al. 2017. “In Situ Capture of Chromatin Interactions by Biotinylated dCas9.” *Cell* 170 (5): 1028–43.e19.

Liu, Xuanyao, Hilary K. Finucane, Alexander Gusev, Gaurav Bhatia, Steven Gazal, Luke O’Connor, Brendan Bulik-Sullivan, et al. 2017. “Functional Architectures of Local and Distal Regulation of Gene Expression in Multiple Human Tissues.” *American Journal of Human Genetics* 100 (4): 605–16.

Li, Wei, Qing Zhang, Naoki Oiso, Edward K. Novak, Rashi Gautam, Edward P. O’Brien, Caroline L. Tinsley, et al. 2003. “Hermansky-Pudlak Syndrome Type 7 (HPS-7) Results from Mutant Dysbindin, a Member of the Biogenesis of Lysosome-Related Organelles Complex 1 (BLOC-1).” *Nature Genetics* 35 (1): 84–89.

Li, Yun, Cristen Willer, Serena Sanna, and Gonçalo Abecasis. 2009. “Genotype Imputation.” *Annual Review of Genomics and Human Genetics* 10: 387–406.

Long, Tao, Michael Hicks, Hung-Chun Yu, William H. Biggs, Ewen F. Kirkness, Cristina Menni, Jonas Zierer, et al. 2017. “Whole-Genome Sequencing Identifies Common-to-Rare Variants Associated with Human Blood Metabolites.” *Nature Genetics* 49 (4): 568–78.

López-Cortegano, Eugenio, and Armando Caballero. 2019. “GWEHS: A Genome-Wide Effect Sizes and Heritability Screener.” *Genes* 10 (8). https://doi.org/10.3390/genes10080558.

López, J. A. 1994. “The Platelet Glycoprotein Ib-IX Complex.” *Blood Coagulation & Fibrinolysis: An International Journal in Haemostasis and Thrombosis* 5 (1): 97–119.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550.

Luan, Jing, Guanjue Xiang, Pablo Aurelio Gómez-García, Jacob M. Tome, Zhe Zhang, Marit W. Vermunt, Haoyue Zhang, et al. 2021. “Distinct Properties and Functions of CTCF Revealed by a Rapidly Inducible Degron System.” *Cell Reports* 34 (8): 108783.

Lupiáñez, Darío G., Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, et al. 2015. “Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions.” *Cell* 161 (5): 1012–25.

Lykke-Andersen, Søren, and Torben Heick Jensen. 2015. “Nonsense-Mediated mRNA Decay: An Intricate Machinery That Shapes Transcriptomes.” *Nature Reviews. Molecular Cell Biology* 16 (11): 665–77.

Lynch, Michael, Bruce Walsh, and Others. 1998. “Genetics and Analysis of Quantitative Traits.”

http://www.invemar.org.co/redcostera1/invemar/docs/RinconLiterario/2011/febrero/AG_8.pdf.

Lyons, Paul A., Tim F. Rayner, Sapna Trivedi, Julia U. Holle, Richard A. Watts, David R. W. Jayne, Bo Baslund, et al. 2012. “Genetically Distinct Subsets within ANCA-Associated Vasculitis.” *The New England Journal of Medicine* 367 (3): 214–23.

MacArthur, Daniel G., and Chris Tyler-Smith. 2010. “Loss-of-Function Variants in the Genomes of Healthy Humans.” *Human Molecular Genetics* 19 (R2): R125–30.

MacDonald, Richard A. 1961. “Lifespan of Liver Cells: Autoradiographic Study Using Tritiated Thymidine in Normal, Cirrhotic, and Partially Hepatectomized Rats.” *Archives of Internal Medicine* 107 (3): 335–43.

Macfarlane, R. G. 1964. “AN ENZYME CASCADE IN THE BLOOD CLOTTING MECHANISM, AND ITS FUNCTION AS A BIOCHEMICAL AMPLIFIER.” *Nature* 202 (May): 498–99.

Mackay, Trudy F. C., and Jason H. Moore. 2015. “Erratum to: Why Epistasis Is Important for Tackling Complex Human Disease Genetics.” *Genome Medicine*. https://doi.org/10.1186/s13073-015-0205-8.

Mackman, Nigel. 2009. “The Role of Tissue Factor and Factor VIIa in Hemostasis.” *Anesthesia and Analgesia* 108 (5): 1447–52.

Mahmoud, Medhat, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J. Sedlazeck. 2019. “Structural Variant Calling: The Long and the Short of It.” *Genome Biology* 20 (1): 246.

Malara, Alessandro, Vittorio Abbonante, Christian A. Di Buduo, Lorenzo Tozzi, Manuela Currao, and Alessandra Balduini. 2015. “The Secret Life of a Megakaryocyte: Emerging Roles in Bone Marrow Homeostasis Control.” *Cellular and Molecular Life Sciences: CMLS* 72 (8): 1517–36.

Malone, James, Ele Holloway, Tomasz Adamusiak, Misha Kapushesky, Jie Zheng, Nikolay Kolesnikov, Anna Zhukova, Alvis Brazma, and Helen Parkinson. 2010. “Modeling Sample Variables with an Experimental Factor Ontology.” *Bioinformatics* 26 (8): 1112–18.

Mangalpally, Kiran Kumar R., Alan Siqueiros-Garcia, Muthiah Vaduganathan, Jing-Fei Dong, Neal S. Kleiman, and Sasidhar Guthikonda. 2010. “Platelet Activation Patterns in Platelet Size Sub-Populations: Differential Responses to Aspirin in Vitro.” *Journal of Thrombosis and Thrombolysis* 30 (3): 251–62.

Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, et al. 2009. “Finding the Missing Heritability of Complex Diseases.” *Nature* 461 (7265): 747–53.

Mao, G. F., L. E. Goldfinger, D. C. Fan, M. P. Lambert, G. Jalagadugula, R. Freishtat, and A. K. Rao. 2017. “Dysregulation ofPLDN(pallidin) Is a Mechanism for Platelet Dense Granule Deficiency inRUNX1haplodeficiency.” *Journal of Thrombosis and Haemostasis*. https://doi.org/10.1111/jth.13619.

Marcdante, Karen, and Robert M. Kliegman. 2018. *Nelson Essentials of Pediatrics E-Book*. Elsevier Health Sciences.

Marchini, Jonathan, Peter Donnelly, and Lon R. Cardon. 2005. “Genome-Wide Strategies for Detecting Multiple Loci That Influence Complex Diseases.” *Nature Genetics* 37 (4): 413–17.

Marchini, Jonathan, and Bryan Howie. 2010. “Genotype Imputation for Genome-Wide Association Studies.” *Nature Reviews. Genetics* 11 (7): 499–511.

Marcus, Aaron J., M. Johan Broekman, Joan H. F. Drosopoulos, David J. Pinsky, Naziba Islam, Richard B. Gayle, and Charles R. Maliszewski. 2001. “Thromboregulation by Endothelial Cells.” *Arteriosclerosis, Thrombosis, and Vascular Biology*. https://doi.org/10.1161/01.atv.21.2.178.

Margolin, J. F., J. R. Friedman, W. K. Meyer, H. Vissing, H. J. Thiesen, and F. J. Rauscher 3rd. 1994. “Krüppel-Associated Boxes Are Potent Transcriptional Repression Domains.” *Proceedings of the National Academy of Sciences of the United States of America* 91 (10): 4509–13.

Marraffini, Luciano A. 2015. “CRISPR-Cas Immunity in Prokaryotes.” *Nature* 526 (7571): 55–61.

Martin, Alicia R., Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. 2019. “Clinical Use of Current Polygenic Risk Scores May Exacerbate Health Disparities.” *Nature Genetics*. https://doi.org/10.1038/s41588-019-0379-x.

Martin, J. F., T. Shaw, J. Heggie, and D. G. Penington. 1983. “Measurement of the Density of Human Platelets and Its Relationship to Volume.” *British Journal of Haematology* 54 (3): 337–52.

Martin-Ramirez, Javier, Menno Hofman, Maartje van den Biggelaar, Robert P. Hebbel, and Jan Voorberg. 2012. “Establishment of Outgrowth Endothelial Cells from Peripheral Blood.” *Nature Protocols* 7 (9): 1709–15.

Massai, Diana, Giulia Soloperto, Diego Gallo, Xiao Yun Xu, and Umberto Morbiducci. 2012. “Shear-Induced Platelet Activation and Its Relationship with Blood Flow Topology in a Numerical Model of Stenosed Carotid Bifurcation.” *European Journal of Mechanics - B/Fluids*. https://doi.org/10.1016/j.euromechflu.2012.03.011.

Mast, Alan E. 2016. “Tissue Factor Pathway Inhibitor: Multiple Anticoagulant Activities for a Single Protein.” *Arteriosclerosis, Thrombosis, and Vascular Biology* 36 (1): 9–14.

“Matrix Profile: CTCF - MA0139.1 - from JASPAR 2018.” n.d. Accessed August 31, 2021. http://jaspar.genereg.net/matrix/MA0139.1/?revcomp=1.

Matsuda, Shigeru, Takehiro Yasukawa, Yuriko Sakaguchi, Kenji Ichiyanagi, Motoko Unoki, Kazuhito Gotoh, Kei Fukuda, Hiroyuki Sasaki, Tsutomu Suzuki, and Dongchon Kang. 2018. “Accurate Estimation of 5-Methylcytosine in Mammalian Mitochondrial DNA.” *Scientific Reports*. https://doi.org/10.1038/s41598-018-24251-z.

Matthews, J. C., K. Hori, and M. J. Cormier. 1977. “Purification and Properties of Renilla Reniformis Luciferase.” *Biochemistry* 16 (1): 85–91.

Maura, Francesco, Andrea Degasperi, Ferran Nadeu, Daniel Leongamornlert, Helen Davies, Luiza Moore, Romina Royo, et al. 2019. “A Practical Guide for Mutational Signature Analysis in Hematological Malignancies.” *Nature Communications* 10 (1): 2969.

Maurano, Matthew T., Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, et al. 2012. “Systematic Localization of Common Disease-Associated Variation in Regulatory DNA.” *Science* 337 (6099): 1190–95.

Mayer, Louisa, Maria Jasztal, Mercedes Pardo, Salvadora Aguera de Haro, Janine Collins, Tadbir K. Bariana, Peter A. Smethurst, et al. 2018. “Nbeal2 Interacts with Dock7, Sec16a, and Vac14.” *Blood* 131 (9): 1000–1011.

McCarthy, Mark I., Gonçalo R. Abecasis, Lon R. Cardon, David B. Goldstein, Julian Little, John P. A. Ioannidis, and Joel N. Hirschhorn. 2008. “Genome-Wide Association Studies for Complex Traits: Consensus, Uncertainty and Challenges.” *Nature Reviews. Genetics* 9 (5): 356–69.

McCarty, Nicholas S., Alicia E. Graham, Lucie Studená, and Rodrigo Ledesma-Amaro. 2020. “Multiplexed CRISPR Technologies for Gene Editing and Transcriptional Regulation.” *Nature Communications* 11 (1): 1281.

McCord, Rachel Patton, Noam Kaplan, and Luca Giorgetti. 2020. “Chromosome Conformation Capture and Beyond: Toward an Integrative View of Chromosome Structure and Function.” *Molecular Cell* 77 (4): 688–708.

McGovern, Amanda, Stefan Schoenfelder, Paul Martin, Jonathan Massey, Kate Duffus, Darren Plant, Annie Yarwood, et al. 2016. “Capture Hi-C Identifies a Novel Causal Gene, IL20RA, in the Pan-Autoimmune Genetic Susceptibility Region 6q23.” *Genome*

Biology 17 (1): 212.
McGrath, J., and D. Solter. 1984. “Completion of Mouse Embryogenesis Requires Both the Maternal and Paternal Genomes.” *Cell* 37 (1): 179–83.
McKinnon, Peter J., and Keith W. Caldecott. 2007. “DNA Strand Break Repair and Human Genetic Disease.” *Annual Review of Genomics and Human Genetics*. https://doi.org/10.1146/annurev.genom.7.080505.115648.
McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. “The Ensembl Variant Effect Predictor.” *Genome Biology* 17 (1): 122.
McNabb, David S., Robin Reed, and Robert A. Marciniak. 2005. “Dual Luciferase Assay System for Rapid Assessment of Gene Expression in Saccharomyces Cerevisiae.” *Eukaryotic Cell* 4 (9): 1539–49.
McNicol, Archibald, and Sara J. Israels. 1999. “Platelet Dense Granules.” *Thrombosis Research*. https://doi.org/10.1016/s0049-3848(99)00015-8.
McPherson, J. D., M. Marra, L. Hillier, R. H. Waterston, A. Chinwalla, J. Wallis, M. Sekhon, et al. 2001. “A Physical Map of the Human Genome.” *Nature* 409 (6822): 934–41.
Mega, J. L., N. O. Stitziel, J. G. Smith, D. I. Chasman, M. Caulfield, J. J. Devlin, F. Nordio, et al. 2015. “Genetic Risk, Coronary Heart Disease Events, and the Clinical Benefit of Statin Therapy: An Analysis of Primary and Secondary Prevention Trials.” *The Lancet* 385 (9984): 2264–71.
Megy, Karyn, Kate Downes, Ilenia Simeoni, Loredana Bury, Joannella Morales, Rutendo Mapeta, Daniel B. Bellissimo, et al. 2019. “Curated Disease-Causing Genes for Bleeding, Thrombotic, and Platelet Disorders: Communication from the SSC of the ISTH.” *Journal of Thrombosis and Haemostasis: JTH* 17 (8): 1253–60.
Meisinger, Christa, Holger Prokisch, Christian Gieger, Nicole Soranzo, Divya Mehta, Dieter Rosskopf, Peter Lichtner, et al. 2009. “A Genome-Wide Association Study Identifies Three Loci Associated with Mean Platelet Volume.” *American Journal of Human Genetics* 84 (1): 66–71.
Melchinger, Hannah, Kanika Jain, Tarun Tyagi, and John Hwa. 2019. “Role of Platelet Mitochondria: Life in a Nucleus-Free Zone.” *Frontiers in Cardiovascular Medicine* 6 (October): 153.
Meletis, John, and V. Goratsa. 2002. “The Derivatives of the Hellenic Word ‘Haema’(hema, Blood) in the English Language.” *Haema* 5 (2): 140–63.
Meletis, John, and Kostas Konstantopoulos. 2010. “The Beliefs, Myths, and Reality Surrounding the Word Hema (Blood) from Homer to the Present.” *Anemia*. https://doi.org/10.1155/2010/857657.
Melhem, Motasem, Mohamed Abu-Farha, Dinu Antony, Ashraf Al Madhoun, Chiara Bacchelli, Fadi Alkayal, Irina AlKhairi, et al. 2017. “Novel G6B Gene Variant Causes Familial Autosomal Recessive Thrombocytopenia and Anemia.” *European Journal of Haematology* 98 (3): 218–27.
Mendel, Gregor. 1866. “Versuche über Plflanzenhybriden. Verhandlungen Des Naturforschenden Vereines in Brünn, Bd. IV Für Das Jahr 1865.” *Abhandlungen*, 3–47.
“Mendelian Inheritance.” n.d. Accessed May 12, 2021. https://www.genome.gov/genetics-glossary/Mendelian-Inheritance.
Meuleman, W., D. Peric-Hupkes, J. Kind, J-B Beaudry, L. Pagie, M. Kellis, M. Reinders, L. Wessels, and B. van Steensel. 2013. “Constitutive Nuclear Lamina-Genome Interactions Are Highly Conserved and Associated with A/T-Rich Sequence.” *Genome Research*. https://doi.org/10.1101/gr.141028.112.
Michalopoulos, George K., and Bharat Bhushan. 2021. “Liver Regeneration: Biological and Pathological Mechanisms and Implications.” *Nature Reviews. Gastroenterology & Hepatology* 18 (1): 40–55.

Michelson, A. D., J. Loscalzo, B. Melnick, B. S. Coller, and R. I. Handin. 1986. “Partial Characterization of a Binding Site for von Willebrand Factor on Glycocalicin.” *Blood* 67 (1): 19–26.

Michelson, Alan D., Marco Cattaneo, Andrew Frelinger, and Peter Newman. 2019. *Platelets*. Academic Press.

Michiels, Carine. 2003. “Endothelial Cell Functions.” *Journal of Cellular Physiology* 196 (3): 430–43.

Mifsud, Borbala, Filipe Tavares-Cadete, Alice N. Young, Robert Sugar, Stefan Schoenfelder, Lauren Ferreira, Steven W. Wingett, et al. 2015. “Mapping Long-Range Promoter Contacts in Human Cells with High-Resolution Capture Hi-C.” *Nature Genetics* 47 (6): 598–606.

Miga, Karen H., Sergey Koren, Arang Rhie, Mitchell R. Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, et al. 2020. “Telomere-to-Telomere Assembly of a Complete Human X Chromosome.” *Nature* 585 (7823): 79–84.

Minikel, Eric Vallabh, Konrad J. Karczewski, Hilary C. Martin, Beryl B. Cummings, Nicola Whiffin, Daniel Rhodes, Jessica Alföldi, et al. 2020. “Evaluating Drug Targets through Human Loss-of-Function Genetic Variation.” *Nature* 581 (7809): 459–64.

Misteli, Tom, Javier F. Cáceres, and David L. Spector. 1997. “The Dynamics of a Pre-mRNA Splicing Factor in Living Cells.” *Nature*. https://doi.org/10.1038/387523a0.

Moncada, S., E. A. Higgs, and J. R. Vane. 1977. “Human Arterial and Venous Tissues Generate Prostacyclin (prostaglandin X), a Potent Inhibitor of Platelet Aggregation.” *The Lancet* 1 (8001): 18–20.

Monroe, Dougald M., and Maureane Hoffman. 2006. “What Does It Take to Make the Perfect Clot?” *Arteriosclerosis, Thrombosis, and Vascular Biology* 26 (1): 41–48.

Morange, Pierre-Emmanuel, Franck Peiretti, Lenaick Gourhant, Carole Proust, Omar Soukarieh, Anne-Sophie Pulcrano-Nicolas, Ganapathi-Varma Saripella, et al. 2021. “A Rare Coding Mutation in the MAST2 Gene Causes Venous Thrombosis in a French Family with Unexplained Thrombophilia: The Breizh MAST2 Arg89Gln Variant.” *PLoS Genetics* 17 (1): e1009284.

Moreau, Thomas, Amanda L. Evans, Louella Vasquez, Marloes R. Tijssen, Ying Yan, Matthew W. Trotter, Daniel Howard, et al. 2016. “Large-Scale Production of Megakaryocytes from Human Pluripotent Stem Cells by Chemically Defined Forward Programming.” *Nature Communications* 7 (April): 11208.

Morena-Barrio, Belén de la, Jonathan Stephens, María Eugenia de la Morena-Barrio, Luca Stefanucci, José Padilla, Antonia Miñano, Nicholas Gleadall, et al. 2020. “Long-Read Sequencing Resolves Structural Variants in SERPINC1 Causing Antithrombin Deficiency and Identifies a Complex Rearrangement and a Retrotransposon Insertion Not Characterized by Routine Diagnostic Methods.” *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/2020.08.28.271932.

Morgens, David W., Michael Wainberg, Evan A. Boyle, Oana Ursu, Carlos L. Araya, C. Kimberly Tsui, Michael S. Haney, et al. 2017. “Genome-Scale Measurement of off-Target Activity Using Cas9 Toxicity in High-Throughput Screens.” *Nature Communications* 8 (May): 15178.

Morrell, Craig N., Angela A. Aggrey, Lesley M. Chapman, and Kristina L. Modjeski. 2014. “Emerging Roles for Platelets as Immune and Inflammatory Cells.” *Blood* 123 (18): 2759–67.

Morris-Rosendahl, Deborah J., and Marc-Antoine Crocq. 2020. “Neurodevelopmental Disorders-the History and Future of a Diagnostic Concept.” *Dialogues in Clinical Neuroscience* 22 (1): 65–72.

Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. “Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq.” *Nature*

Methods 5 (7): 621–28.
Moss, M. L. 1981. “Genetics, Epigenetics, and Causation.” *American Journal of Orthodontics* 80 (4): 366–75.
Mullaney, Julienne M., Ryan E. Mills, W. Stephen Pittard, and Scott E. Devine. 2010. “Small Insertions and Deletions (INDELs) in Human Genomes.” *Human Molecular Genetics* 19 (R2): R131–36.
Mulvey, B., T. Lagunas, and J. D. Dougherty. 2020. “The Oft-Overlooked Massively Parallel Reporter Assay: Where, When, and Which Psychiatric Genetic Variants Are Functional?” *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.02.02.931337v2.abstract.
Munafò, Marcus R., Kate Tilling, Amy E. Taylor, David M. Evans, and George Davey Smith. 2018. “Collider Scope: When Selection Bias Can Substantially Influence Observed Associations.” *International Journal of Epidemiology* 47 (1): 226–35.
Mundal, H. H., P. Hjemdahl, and K. Gjesdal. 1998. “Acute Effects of Cigarette Smoking on Platelet Function and Plasma Catecholamines in Hypertensive and Normotensive Men.” *American Journal of Hypertension* 11 (6 Pt 1): 677–81.
Mus, El, Patrick R. Hof, and Henri Tiedge. 2007. “Dendritic BC200 RNA in Aging and in Alzheimer’s Disease.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (25): 10679–84.
Myint, Leslie, Dimitrios G. Avramopoulos, Loyal A. Goff, and Kasper D. Hansen. 2019. “Linear Models Enable Powerful Differential Activity Analysis in Massively Parallel Reporter Assays.” *BMC Genomics* 20 (1): 209.
Naeem, Muhammad, Saman Majeed, Mubasher Zahir Hoque, and Irshad Ahmad. 2020. “Latest Developed Strategies to Minimize the Off-Target Effects in CRISPR-Cas-Mediated Genome Editing.” *Cells* 9 (7). https://doi.org/10.3390/cells9071608.
Nagano, Takashi, Yaniv Lubling, Eitan Yaffe, Steven W. Wingett, Wendy Dean, Amos Tanay, and Peter Fraser. 2015. “Single-Cell Hi-C for Genome-Wide Detection of Chromatin Interactions That Occur Simultaneously in a Single Cell.” *Nature Protocols* 10 (12): 1986–2003.
Nakamura, Muneaki, Yuchen Gao, Antonia A. Dominguez, and Lei S. Qi. 2021. “CRISPR Technologies for Precise Epigenome Editing.” *Nature Cell Biology* 23 (1): 11–22.
Nakamura, Sou, Naoya Takayama, Shinji Hirata, Hideya Seo, Hiroshi Endo, Kiyosumi Ochi, Ken-Ichi Fujita, et al. 2014. “Expandable Megakaryocyte Cell Lines Enable Clinically Applicable Generation of Platelets from Human Induced Pluripotent Stem Cells.” *Cell Stem Cell* 14 (4): 535–48.
Nakato, Ryuichiro, Youichiro Wada, Ryo Nakaki, Genta Nagae, Yuki Katou, Shuichi Tsutsumi, Natsu Nakajima, et al. 2019. “Comprehensive Epigenome Characterization Reveals Diverse Transcriptional Regulation across Human Vascular Endothelial Cells.” *Epigenetics & Chromatin* 12 (1): 77.
Narendra, Varun, Pedro P. Rocha, Disi An, Ramya Raviram, Jane A. Skok, Esteban O. Mazzoni, and Danny Reinberg. 2015. “CTCF Establishes Discrete Functional Chromatin Domains at the Hox Clusters during Differentiation.” *Science* 347 (6225): 1017–21.
Nasser, Joseph, Drew T. Bergman, Charles P. Fulco, Philine Guckelberger, Benjamin R. Doughty, Tejal A. Patwardhan, Thouis R. Jones, et al. 2021. “Genome-Wide Enhancer Maps Link Risk Variants to Disease Genes.” *Nature*. https://doi.org/10.1038/s41586-021-03446-x.
“National Center for Biotechnology Information.” n.d. Accessed May 15, 2021. https://www.ncbi.nlm.nih.gov/.
Neerman, Nir, Gregory Faust, Naomi Meeks, Shira Modai, Limor Kalfon, Tzipora

Falik-Zaccai, and Alexander Kaplun. 2019. “A Clinically Validated Whole Genome Pipeline for Structural Variant Detection and Analysis.” *BMC Genomics* 20 (Suppl 8): 545.

Nelson, Matthew R., Hannah Tipney, Jeffery L. Painter, Judong Shen, Paola Nicoletti, Yufeng Shen, Aris Floratos, et al. 2015. “The Support of Human Genetic Evidence for Approved Drug Indications.” *Nature Genetics* 47 (8): 856–60.

Nesta, Alex V., Denisse Tafur, and Christine R. Beck. 2020. “Hotspots of Human Mutation.” *Trends in Genetics: TIG*, November. https://doi.org/10.1016/j.tig.2020.10.003.

Nestorowa, Sonia, Fiona K. Hamey, Blanca Pijuan Sala, Evangelia Diamanti, Mairi Shepherd, Elisa Laurenti, Nicola K. Wilson, David G. Kent, and Berthold Göttgens. 2016. “A Single-Cell Resolution Map of Mouse Hematopoietic Stem and Progenitor Cell Differentiation.” *Blood* 128 (8): e20–31.

Neuenschwander, P. F. 2006. “COAGULATION CASCADE | Overview.” In *Encyclopedia of Respiratory Medicine*, edited by Geoffrey J. Laurent and Steven D. Shapiro, 478–86. Oxford: Academic Press.

Newland, Stephen A., Iain C. Macaulay, Andres R. Floto, Edwin C. de Vet, Willem H. Ouwehand, Nicholas A. Watkins, Paul A. Lyons, and Duncan R. Campbell. 2007. “The Novel Inhibitory Receptor G6B Is Expressed on the Surface of Platelets and Attenuates Platelet Function in Vitro.” *Blood* 109 (11): 4806–9.

Newton, Richard, and Lorenz Wernisch. 2015. “Investigating Inter-Chromosomal Regulatory Relationships through a Comprehensive Meta-Analysis of Matched Copy Number and Transcriptomics Data Sets.” *BMC Genomics* 16 (November): 967.

Ngo, Jacky Chi Ki, Mingdong Huang, David A. Roth, Barbara C. Furie, and Bruce Furie. 2008. “Crystal Structure of Human Factor VIII: Implications for the Formation of the Factor IXa-Factor VIIIa Complex.” *Structure*. https://doi.org/10.1016/j.str.2008.03.001.

Nguyen, Thomas A., Richard D. Jones, Andrew R. Snavely, Andreas R. Pfenning, Rory Kirchner, Martin Hemberg, and Jesse M. Gray. 2016. “High-Throughput Functional Comparison of Promoter and Enhancer Activities.” *Genome Research* 26 (8): 1023–33.

Nicchia, Elena, P. Giordano, C. Greco, Daniela De Rocco, and Anna Savoia. 2016. “Molecular Diagnosis of Thrombocytopenia-Absent Radius Syndrome Using next-Generation Sequencing.” *International Journal of Laboratory Hematology* 38 (4): 412–18.

Niedringhaus, Thomas P., Denitsa Milanova, Matthew B. Kerby, Michael P. Snyder, and Annelise E. Barron. 2011. “Landscape of next-Generation Sequencing Technologies.” *Analytical Chemistry* 83 (12): 4327–41.

Nieswandt, Bernhard, and Steve P. Watson. 2003. “Platelet-Collagen Interaction: Is GPVI the Central Receptor?” *Blood* 102 (2): 449–61.

“NIHR BioResource Home Page.” n.d. Accessed May 13, 2021. https://bioresource.nihr.ac.uk/.

Nilsen, Timothy W., and Brenton R. Graveley. 2010. “Expansion of the Eukaryotic Proteome by Alternative Splicing.” *Nature* 463 (7280): 457–63.

Nimmo, Rachael A., Gillian E. May, and Tariq Enver. 2015. “Primed and Ready: Understanding Lineage Commitment through Single Cell Analysis.” *Trends in Cell Biology* 25 (8): 459–67.

Noordermeer, Daan, Elzo de Wit, Petra Klous, Harmen van de Werken, Marieke Simonis, Melissa Lopez-Jones, Bert Eussen, Annelies de Klein, Robert H. Singer, and Wouter de Laat. 2011. “Variegated Gene Expression Caused by Cell-Specific Long-Range DNA Interactions.” *Nature Cell Biology* 13 (8): 944–51.

Nora, Elphège P., Anton Goloborodko, Anne-Laure Valton, Johan H. Gibcus, Alec Uebersohn, Nezar Abdennur, Job Dekker, Leonid A. Mirny, and Benoit G. Bruneau. 2017. “Targeted Degradation of CTCF Decouples Local Insulation of Chromosome

Domains from Genomic Compartmentalization." *Cell* 169 (5): 930–44.e22.
Nora, Elphège P., Bryan R. Lajoie, Edda G. Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, et al. 2012. "Spatial Partitioning of the Regulatory Landscape of the X-Inactivation Centre." *Nature* 485 (7398): 381–85.
Notta, Faiyaz, Sasan Zandi, Naoya Takayama, Stephanie Dobson, Olga I. Gan, Gavin Wilson, Kerstin B. Kaufmann, et al. 2016. "Distinct Routes of Lineage Development Reshape the Human Blood Hierarchy across Ontogeny." *Science* 351 (6269): aab2116.
Nurden, Alan T. 2011. "Platelets, Inflammation and Tissue Regeneration." *Thrombosis and Haemostasis* 105 Suppl 1 (May): S13–33.
Oberlin, Estelle, Manuela Tavian, Istvàn Blazsek, and Bruno Péault. 2002. "Blood-Forming Potential of Vascular Endothelium in the Human Embryo." *Development* 129 (17): 4147–57.
O'Connor, Luke J. 2021. "The Distribution of Common-Variant Effect Sizes." *Nature Genetics*, July, 1–7.
Oetjens, M. T., M. A. Kelly, A. C. Sturm, C. L. Martin, and D. H. Ledbetter. 2019. "Quantifying the Polygenic Contribution to Variable Expressivity in Eleven Rare Genetic Disorders." *Nature Communications* 10 (1): 4897.
Ohno, Susumu. 1972. "So much'junk'DNA in Our Genome." In *Evolution of Genetic Systems, Brookhaven Symp. Biol.*, 366–70.
"OMIM Gene Map Statistics." n.d. Accessed May 15, 2021. https://www.omim.org/statistics/geneMap.
Ong, Chin-Tong, and Victor G. Corces. 2011. "Enhancer Function: New Insights into the Regulation of Tissue-Specific Gene Expression." *Nature Reviews. Genetics* 12 (4): 283–93.
"On the Pathology of the Blood." 1831. *The Medico-Chirurgical Review* 15 (30): 337–54.
Orkin, Stuart H., and Leonard I. Zon. 2008. "Hematopoiesis: An Evolving Paradigm for Stem Cell Biology." *Cell* 132 (4): 631–44.
Osawa, M., K. Hanada, H. Hamada, and H. Nakauchi. 1996. "Long-Term Lymphohematopoietic Reconstitution by a Single CD34-Low/negative Hematopoietic Stem Cell." *Science* 273 (5272): 242–45.
Oudelaar, A. Marieke, Caroline L. Harrold, Lars L. P. Hanssen, Jelena M. Telenius, Douglas R. Higgs, and Jim R. Hughes. 2019. "A Revised Model for Promoter Competition Based on Multi-Way Chromatin Interactions at the α-Globin Locus." *Nature Communications* 10 (1): 5412.
Oudenrijn, S. van den, M. Bruin, C. C. Folman, M. Peters, L. B. Faulkner, M. de Haas, and A. E. von dem Borne. 2000. "Mutations in the Thrombopoietin Receptor, Mpl, in Children with Congenital Amegakaryocytic Thrombocytopenia." *British Journal of Haematology* 110 (2): 441–48.
"Our Future Health." n.d. Accessed August 6, 2021. https://ourfuturehealth.org.uk/.
Overturf, K., M. al-Dhalimy, C. N. Ou, M. Finegold, and M. Grompe. 1997. "Serial Transplantation Reveals the Stem-Cell-like Regenerative Potential of Adult Mouse Hepatocytes." *The American Journal of Pathology* 151 (5): 1273–80.
Ozaki, Kouichi, Yozo Ohnishi, Aritoshi Iida, Akihiko Sekine, Ryo Yamada, Tatsuhiko Tsunoda, Hiroshi Sato, et al. 2002. "Functional SNPs in the Lymphotoxin-α Gene That Are Associated with Susceptibility to Myocardial Infarction." *Nature Genetics* 32 (4): 650–54.
Padmakumar, Manisha, Eveline Van Raes, Chris Van Geet, and Kathleen Freson. 2019. "Blood Platelet Research in Autism Spectrum Disorders: In Search of Biomarkers." *Research and Practice in Thrombosis and Haemostasis*. https://doi.org/10.1002/rth2.12239.
Palazzo, Alexander F., and T. Ryan Gregory. 2014. "The Case for Junk DNA." *PLoS*

Genetics 10 (5): e1004351.
Palazzo, Alexander F., and Eliza S. Lee. 2015. “Non-Coding RNA: What Is Functional and What Is Junk?” *Frontiers in Genetics* 6 (January): 2.
Palmer, R. M., A. G. Ferrige, and S. Moncada. 1987. “Nitric Oxide Release Accounts for the Biological Activity of Endothelium-Derived Relaxing Factor.” *Nature* 327 (6122): 524–26.
Pan, Qun, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. 2008. “Deep Surveying of Alternative Splicing Complexity in the Human Transcriptome by High-Throughput Sequencing.” *Nature Genetics*. https://doi.org/10.1038/ng.259.
Pansky, Ben. 1982. *Review of Medical Embryology*. Macmillan Publishing Company.
Papatheodorou, Irene, Pablo Moreno, Jonathan Manning, Alfonso Muñoz-Pomer Fuentes, Nancy George, Silvie Fexova, Nuno A. Fonseca, et al. 2020. “Expression Atlas Update: From Tissues to Single Cells.” *Nucleic Acids Research* 48 (D1): D77–83.
Patnaik, M. M., and S. Moll. 2008. “Inherited Antithrombin Deficiency: A Review.” *Haemophilia: The Official Journal of the World Federation of Hemophilia* 14 (6): 1229–39.
Patterson, Nick, Alkes L. Price, and David Reich. 2006. “Population Structure and Eigenanalysis.” *PLoS Genetics* 2 (12): e190.
Paul, Franziska, Ya ’ara Arkin, Amir Giladi, Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Deborah Winter, et al. 2016. “Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors.” *Cell* 164 (1-2): 325.
Payer, Lindsay M., and Kathleen H. Burns. 2019. “Transposable Elements in Human Genetic Disease.” *Nature Reviews. Genetics* 20 (12): 760–72.
Peelman, L. J., P. Chardon, M. Nunes, C. Renard, C. Geffrotin, M. Vaiman, A. Van Zeveren, W. Coppieters, A. van de Weghe, and Y. Bouquet. 1995. “The BAT1 Gene in the MHC Encodes an Evolutionarily Conserved Putative Nuclear RNA Helicase of the DEAD Family.” *Genomics* 26 (2): 210–18.
Peleg, Shahaf, Christian Feller, Andreas G. Ladurner, and Axel Imhof. 2016. “The Metabolic Impact on Histone Acetylation and Transcription in Ageing.” *Trends in Biochemical Sciences* 41 (8): 700–711.
Pennacchio, Len A., Wendy Bickmore, Ann Dean, Marcelo A. Nobrega, and Gill Bejerano. 2013. “Enhancers: Five Essential Questions.” *Nature Reviews. Genetics* 14 (4): 288–95.
Perdomo, Jose, Feng Yan, Halina H. L. Leung, and Beng H. Chong. 2017. “Megakaryocyte Differentiation and Platelet Formation from Human Cord Blood-Derived CD34+ Cells.” *Journal of Visualized Experiments: JoVE*, no. 130 (December). https://doi.org/10.3791/56420.
Pers, Tune H., Juha M. Karjalainen, Yingleong Chan, Harm-Jan Westra, Andrew R. Wood, Jian Yang, Julian C. Lui, et al. 2015. “Biological Interpretation of Genome-Wide Association Studies Using Predicted Gene Functions.” *Nature Communications* 6 (January): 5890.
Pertea, Mihaela, Alaina Shumate, Geo Pertea, Ales Varabyou, Yu-Chi Chang, Anil K. Madugundu, Akhilesh Pandey, and Steven L. Salzberg. 2018. “Thousands of Large-Scale RNA Sequencing Experiments Yield a Comprehensive New Human Gene List and Reveal Extensive Transcriptional Noise.” *bioRxiv*. https://doi.org/10.1101/332825.
Petersen, Romina, John J. Lambourne, Biola M. Javierre, Luigi Grassi, Roman Kreuzhuber, Dace Ruklisa, Isabel M. Rosa, et al. 2017. “Platelet Function Is Modified by Common Sequence Variation in Megakaryocyte Super Enhancers.” *Nature Communications* 8 (July): 16058.
Petrova, Natalia V., and Cathy H. Wu. 2006. “Prediction of Catalytic Residues Using Support

Vector Machine with Selected Protein Sequence and Structural Properties." *BMC Bioinformatics* 7 (June): 312.

Peyvandi, Flora, Roberta Palla, Marzia Menegatti, and Pier Mannuccio Mannucci. 2009. "Introduction. Rare Bleeding Disorders: General Aspects of Clinical Features, Diagnosis, and Management." *Seminars in Thrombosis and Hemostasis* 35 (4): 349–55.

Peyvandi, F., R. Palla, M. Menegatti, S. M. Siboni, S. Halimeh, B. Faeser, H. Pergantou, et al. 2012. "Coagulation Factor Activity and Clinical Bleeding Severity in Rare Bleeding Disorders: Results from the European Network of Rare Bleeding Disorders." *Journal of Thrombosis and Haemostasis: JTH* 10 (4): 615–21.

Phylipsen, Marion, John F. Prior, Erna Lim, Neela Lingam, Ingrid P. Vogelaar, Piero C. Giordano, Jill Finlayson, and Cornelis L. Harteveld. 2010. "Thalassemia in Western Australia: 11 Novel Deletions Characterized by Multiplex Ligation-Dependent Probe Amplification." *Blood Cells, Molecules & Diseases* 44 (3): 146–51.

Pillay, Janesh, Ineke den Braber, Nienke Vrisekoop, Lydia M. Kwast, Rob J. de Boer, José A. M. Borghans, Kiki Tesselaar, and Leo Koenderman. 2010. "In Vivo Labeling with 2H2O Reveals a Human Neutrophil Lifespan of 5.4 Days." *Blood* 116 (4): 625–27.

Plenge, Robert M., Edward M. Scolnick, and David Altshuler. 2013. "Validating Therapeutic Targets through Human Genetics." *Nature Reviews. Drug Discovery* 12 (8): 581–94.

Plo, Isabelle, Christine Bellanné-Chantelot, Matthieu Mosca, Stefania Mazzi, Caroline Marty, and William Vainchenker. 2017. "Genetic Alterations of the Thrombopoietin/MPL/JAK2 Axis Impacting Megakaryopoiesis." *Frontiers in Endocrinology* 8 (September): 234.

Podolak-Dawidziak, M., V. Hancock, R. Lelchuk, S. Kotlarek-Haus, and J. F. Martin. 1995. "The Expression of mRNA for Fibrinogen in Megakaryocytes Isolated from Patients with T-Cell Lymphoma." *British Journal of Haematology*. https://doi.org/10.1111/j.1365-2141.1995.tb05304.x.

Poort, Swibertus, Hans Vos, Rogier Bertina, Chularatana Mahasandana, Voravarn Tanphaichitr, Gavivann Veerakul, Suthida Kankirawatana, Vinai Suvatte, and Parichat Pung-amritt. 1999. "Compound Heterozygosity for One Novel and One Recurrent Mutation in a Thai Patient with Severe Protein S Deficiency." *Thrombosis and Haemostasis*. https://doi.org/10.1055/s-0037-1614440.

Popescu, Dorin-Mirel, Rachel A. Botting, Emily Stephenson, Kile Green, Simone Webb, Laura Jardine, Emily F. Calderbank, et al. 2019. "Decoding Human Fetal Liver Haematopoiesis." *Nature* 574 (7778): 365–71.

Qin, Yufeng, Sara A. Grimm, John D. Roberts, Kaliopi Chrysovergis, and Paul A. Wade. 2020. "Alterations in Promoter Interaction Landscape and Transcriptional Network Underlying Metabolic Adaptation to Diet." *Nature Communications* 11 (1): 1–16.

Quinsey, Noelene S., Ainslie L. Greedy, Stephen P. Bottomley, James C. Whisstock, and Robert N. Pike. 2004. "Antithrombin: In Control of Coagulation." *The International Journal of Biochemistry & Cell Biology*. https://doi.org/10.1016/s1357-2725(03)00244-9.

Rafii, Shahin, and David Lyden. 2003. "Therapeutic Stem and Progenitor Cell Transplantation for Organ Vascularization and Regeneration." *Nature Medicine* 9 (6): 702–12.

Rao, Suhas S. P., Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, et al. 2014. "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping." *Cell* 159 (7): 1665–80.

Raskob, Gary, and ISTH Steering Committee for World Thrombosis Day. 2014. "Thrombosis: A Major Contributor to Global Disease Burden." *Thrombosis and Haemostasis*. https://doi.org/10.1160/th14-08-0671.

Raslova, Hana, Audrey Kauffmann, Dalila Sekkaï, Hugues Ripoche, Fréderic Larbret,

Thomas Robert, Diana Tronik Le Roux, et al. 2007. “Interrelation between Polyploidization and Megakaryocyte Differentiation: A Gene Profiling Approach.” *Blood* 109 (8): 3225–34.

Rath, Ana, Annie Olry, Ferdinand Dhombres, Maja Miličić Brandt, Bruno Urbero, and Segolene Ayme. 2012. “Representation of Rare Diseases in Health Information Systems: The Orphanet Approach to Serve a Wide Range of End Users.” *Human Mutation* 33 (5): 803–8.

Razin, A., and A. D. Riggs. 1980. “DNA Methylation and Gene Function.” *Science* 210 (4470): 604–10.

Reese, Jessica A., Jennifer D. Peck, David R. Deschamps, Jennifer J. McIntosh, Eric J. Knudtson, Deirdra R. Terrell, Sara K. Vesely, and James N. George. 2018. “Platelet Counts during Pregnancy.” *The New England Journal of Medicine* 379 (1): 32–43.

Rein, Chantelle M., Umesh R. Desai, and Frank C. Church. 2011. “Chapter Seven - Serpin–Glycosaminoglycan Interactions.” In *Methods in Enzymology*, edited by James C. Whisstock and Phillip I. Bird, 501:105–37. Academic Press.

Relethford, J. H. 2008. “Genetic Evidence and the Modern Human Origins Debate.” *Heredity* 100 (6): 555–63.

Repke, D., C. H. Gemmell, A. Guha, V. T. Turitto, G. J. Broze Jr, and Y. Nemerson. 1990. “Hemophilia as a Defect of the Tissue Factor Pathway of Blood Coagulation: Effect of Factors VIII and IX on Factor X Activation in a Continuous-Flow Reactor.” *Proceedings of the National Academy of Sciences of the United States of America* 87 (19): 7623–27.

Reuter, Jason A., Damek V. Spacek, and Michael P. Snyder. 2015. “High-Throughput Sequencing Technologies.” *Molecular Cell* 58 (4): 586–97.

Rhodes, Christopher J., Ken Batai, Marta Bleda, Matthias Haimel, Laura Southgate, Marine Germain, Michael W. Pauciulo, et al. 2019. “Genetic Determinants of Risk in Pulmonary Arterial Hypertension: International Genome-Wide Association Studies and Meta-Analysis.” *The Lancet. Respiratory Medicine* 7 (3): 227–38.

Ribas, G., M. Neville, J. L. Wixon, and J. Cheng. 1999. “Genes Encoding Three New Members of the Leukocyte Antigen 6 Superfamily and a Novel Member of Ig Superfamily, Together with Genes Encoding the Regulatory ….” *The Journal of*. https://www.jimmunol.org/content/163/1/278.short.

Richards, Sue, ; on behalf of the ACMG Laboratory Quality Assurance Committee, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, et al. 2015. “Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.” *Genetics in Medicine*. https://doi.org/10.1038/gim.2015.30.

Richmond, Jillian M., and John E. Harris. 2014. “Immunology and Skin in Health and Disease.” *Cold Spring Harbor Perspectives in Medicine* 4 (12): a015339.

Riggs, A. D., and P. A. Jones. 1983. “5-Methylcytosine, Gene Regulation, and Cancer.” *Advances in Cancer Research* 40: 1–30.

Rios, Raquel, Bruno Sangro, Ignacio Herrero, Jorge Quiroga, and Jesus Prieto. 2005. “The Role of Thrombopoietin in the Thrombocytopenia of Patients with Liver Cirrhosis.” *The American Journal of Gastroenterology* 100 (6): 1311–16.

Ripatti, Samuli, Emmi Tikkanen, Marju Orho-Melander, Aki S. Havulinna, Kaisa Silander, Amitabh Sharma, Candace Guiducci, et al. 2010. “A Multilocus Genetic Risk Score for Coronary Heart Disease: Case-Control and Prospective Cohort Analyses.” *The Lancet* 376 (9750): 1393–1400.

Roberts, Lara N., Raj K. Patel, and Roopen Arya. 2010. “Haemostasis and Thrombosis in Liver Disease.” *British Journal of Haematology* 148 (4): 507–21.

Robertson, A., and I. M. Lerner. 1949. “The Heritability of All-or-None Traits: Viability of

Poultry." *Genetics* 34 (4): 395–411.
Robinson, Peter N., Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. 2008. "The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease." *American Journal of Human Genetics* 83 (5): 610–15.
Rodriquez-Merchan, E. C., and Christine A. Lee. 2008. Inhibitors in Patients with Haemophilia. John Wiley & Sons.
Rood, Jennifer E., and Aviv Regev. 2021. "The Legacy of the Human Genome Project." Science 373 (6562): 1442–43.
Rosenbluh, Joseph, Han Xu, William Harrington, Stanley Gill, Xiaoxing Wang, Francisca Vazquez, David E. Root, Aviad Tsherniak, and William C. Hahn. 2017. "Complementary Information Derived from CRISPR Cas9 Mediated Gene Deletion and Suppression." *Nature Communications* 8 (May): 15403.
Rosenblum, B. B., L. G. Lee, S. L. Spurgeon, S. H. Khan, S. M. Menchen, C. R. Heiner, and S. M. Chen. 1997. "New Dye-Labeled Terminators for Improved DNA Sequencing Patterns." *Nucleic Acids Research* 25 (22): 4500–4504.
Rosenfeld, Jeffrey A., Zhibin Wang, Dustin E. Schones, Keji Zhao, Rob DeSalle, and Michael Q. Zhang. 2009. "Determination of Enriched Histone Modifications in Non-Genic Portions of the Human Genome." *BMC Genomics* 10 (March): 143.
Rothmayr, Christine. 2009. "Biobanks: Governance in Comparative Perspective - Edited by Herbert Gottweis and Alan Petersen." *Governance*. https://doi.org/10.1111/j.1468-0491.2009.01461_6.x.
Ruggeri, Zaverio M. 2009. "Platelet Adhesion under Flow." *Microcirculation* 16 (1): 58–83.
Ruiz-Velasco, Mariana, Manjeet Kumar, Mang Ching Lai, Pooja Bhat, Ana Belen Solis-Pinson, Alejandro Reyes, Stefan Kleinsorg, Kyung-Min Noh, Toby J. Gibson, and Judith B. Zaugg. 2017. "CTCF-Mediated Chromatin Loops between Promoter and Gene Body Regulate Alternative Splicing across Individuals." *Cell Systems* 5 (6): 628–37.e6.
Sabater-Lleal, Maria, Jennifer E. Huffman, Paul S. de Vries, Jonathan Marten, Michael A. Mastrangelo, Ci Song, Nathan Pankratz, et al. 2019. "Genome-Wide Association Transethnic Meta-Analyses Identifies Novel Associations Regulating Coagulation Factor VIII and von Willebrand Factor Plasma Levels." *Circulation* 139 (5): 620–35.
Sabatine, Marc S., Robert P. Giugliano, Anthony C. Keech, Narimon Honarpour, Stephen D. Wiviott, Sabina A. Murphy, Julia F. Kuder, et al. 2017. "Evolocumab and Clinical Outcomes in Patients with Cardiovascular Disease." *The New England Journal of Medicine* 376 (18): 1713–22.
Sadler, J. E. 1997. "Thrombomodulin Structure and Function." *Thrombosis and Haemostasis* 78 (1): 392–95.
———. 1998. "Biochemistry and Genetics of von Willebrand Factor." *Annual Review of Biochemistry* 67: 395–424.
Sakata, T., K. Kario, Y. Katayama, T. Matsuyama, H. Kato, and T. Miyata. 2000. "Studies on Congenital Protein C Deficiency in Japanese: Prevalence, Genetic Analysis, and Relevance to the Onset of Arterial Occlusive Diseases." *Seminars in Thrombosis and Hemostasis* 26 (1): 11–16.
Sallusto, F., D. Lenig, R. Förster, M. Lipp, and A. Lanzavecchia. 1999. "Two Subsets of Memory T Lymphocytes with Distinct Homing Potentials and Effector Functions." *Nature* 401 (6754): 708–12.
Sanger, F. 1988. "Sequences, Sequences, and Sequences." *Annual Review of Biochemistry* 57: 1–28.
Sanger, F., S. Nicklen, and A. R. Coulson. 1977. "DNA Sequencing with Chain-Terminating Inhibitors." *Proceedings of the National Academy of Sciences of the United States of America* 74 (12): 5463–67.

Sanjana, Neville E., Ophir Shalem, and Feng Zhang. 2014. “Improved Vectors and Genome-Wide Libraries for CRISPR Screening.” *Nature Methods* 11 (8): 783–84.

Sanjuan-Pla, Alejandra, Iain C. Macaulay, Christina T. Jensen, Petter S. Woll, Tiago C. Luis, Adam Mead, Susan Moore, et al. 2013. “Platelet-Biased Stem Cells Reside at the Apex of the Haematopoietic Stem-Cell Hierarchy.” *Nature* 502 (7470): 232–36.

Sanna, Serena, Anne U. Jackson, Ramaiah Nagaraja, Cristen J. Willer, Wei-Min Chen, Lori L. Bonnycastle, Haiqing Shen, et al. 2008. “Common Variants in the GDF5-UQCC Region Are Associated with Variation in Human Height.” *Nature Genetics* 40 (2): 198–203.

Sanyal, Amartya, Bryan R. Lajoie, Gaurav Jain, and Job Dekker. 2012. “The Long-Range Interaction Landscape of Gene Promoters.” *Nature* 489 (7414): 109–13.

Sati, Satish, and Giacomo Cavalli. 2017. “Chromosome Conformation Capture Technologies and Their Impact in Understanding Genome Function.” *Chromosoma* 126 (1): 33–44.

Sawada, N., and T. Ishikawa. 1988. “Reduction of Potential for Replicative but Not Unscheduled DNA Synthesis in Hepatocytes Isolated from Aged as Compared to Young Rats.” *Cancer Research* 48 (6): 1618–22.

Scandurra, Valeria, Leonardo Emberti Gialloreti, Francesca Barbanera, Marirosa Rosaria Scordo, Angelo Pierini, and Roberto Canitano. 2019. “Neurodevelopmental Disorders and Adaptive Functions: A Study of Children With Autism Spectrum Disorders (ASD) And/or Attention Deficit and Hyperactivity Disorder (ADHD).” *Frontiers in Psychiatry / Frontiers Research Foundation* 10 (September): 673.

Schmidt, Florian, Fabian Kern, and Marcel H. Schulz. 2020. “Integrative Prediction of Gene Expression with Chromatin Accessibility and Conformation Data.” *Epigenetics & Chromatin* 13 (1): 4.

Schoenfelder, Stefan, and Peter Fraser. 2019. “Long-Range Enhancer–promoter Contacts in Gene Expression Control.” *Nature Reviews. Genetics* 20 (8): 437–55.

Schoenfelder, Stefan, Mayra Furlan-Magaril, Borbala Mifsud, Filipe Tavares-Cadete, Robert Sugar, Biola-Maria Javierre, Takashi Nagano, et al. 2015. “The Pluripotent Regulatory Circuitry Connecting Promoters to Their Long-Range Interacting Elements.” *Genome Research* 25 (4): 582–97.

Schoenfelder, Stefan, Biola-Maria Javierre, Mayra Furlan-Magaril, Steven W. Wingett, and Peter Fraser. 2018. “Promoter Capture Hi-C: High-Resolution, Genome-Wide Profiling of Promoter Interactions.” *Journal of Visualized Experiments: JoVE*, no. 136 (June). https://doi.org/10.3791/57320.

Schoenfelder, Stefan, Robert Sugar, Andrew Dimond, Biola-Maria Javierre, Harry Armstrong, Borbala Mifsud, Emilia Dimitrova, et al. 2015. “Polycomb Repressive Complex PRC1 Spatially Constrains the Mouse Embryonic Stem Cell Genome.” *Nature Genetics* 47 (10): 1179–86.

Schwartzentruber, Jeremy, Sarah Cooper, Jimmy Z. Liu, Inigo Barrio-Hernandez, Erica Bello, Natsuhiko Kumasaka, Adam M. H. Young, et al. 2021. “Genome-Wide Meta-Analysis, Fine-Mapping and Integrative Prioritization Implicate New Alzheimer’s Disease Risk Genes.” *Nature Genetics* 53 (3): 392–402.

Schwarzer, Wibke, Nezar Abdennur, Anton Goloborodko, Aleksandra Pekowska, Geoffrey Fudenberg, Yann Loe-Mie, Nuno A. Fonseca, et al. 2017. “Two Independent Modes of Chromatin Organization Revealed by Cohesin Removal.” *Nature* 551 (7678): 51–56.

Schwessinger, Ron, Matthew Gosden, Damien Downes, Richard C. Brown, A. Marieke Oudelaar, Jelena Telenius, Yee Whye Teh, Gerton Lunter, and Jim R. Hughes. 2020. “DeepC: Predicting 3D Genome Folding Using Megabase-Scale Transfer Learning.” *Nature Methods* 17 (11): 1118–24.

Sedlazeck, Fritz J., Hayan Lee, Charlotte A. Darby, and Michael C. Schatz. 2018. “Piercing the Dark Matter: Bioinformatics of Long-Range Sequencing and Mapping.” *Nature*

Reviews. Genetics 19 (6): 329–46.
Sedlazeck, Fritz J., Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C. Schatz. 2018. “Accurate Detection of Complex Structural Variations Using Single-Molecule Sequencing.” *Nature Methods* 15 (6): 461–68.
Sellaro, T. L. 2007. “Ravindra, aK, Stolz, DB & Badylak, SF Maintenance of Hepatic Sinusoidal Endothelial Cell Phenotype in Vitrousing Organ-Specific Extracellular Matrix Scaffolds.” *Tissue Engineering* 13: 2301–10.
Seltsam, Axel, Michael Hallensleben, Anke Kollmann, and Rainer Blasczyk. 2003. “The Nature of Diversity and Diversification at the ABO Locus.” *Blood* 102 (8): 3035–42.
Sexton, Travis, and Susan S. Smyth. 2014. “Novel Mediators and Biomarkers of Thrombosis.” *Journal of Thrombosis and Thrombolysis* 37 (1): 1–3.
Shah, Naisha, Ying-Chen Claire Hou, Hung-Chun Yu, Rachana Sainger, C. Thomas Caskey, J. Craig Venter, and Amalio Telenti. 2018. “Identification of Misclassified ClinVar Variants via Disease Population Prevalence.” *The American Journal of Human Genetics*. https://doi.org/10.1016/j.ajhg.2018.02.019.
Shattil, S. J., H. Kashiwagi, and N. Pampori. 1998. “Integrin Signaling: The Platelet Paradigm.” *Blood* 91 (8): 2645–57.
Sheppard, Brooke, Nadav Rappoport, Po-Ru Loh, Stephan J. Sanders, Noah Zaitlen, and Andy Dahl. 2021. “A Model and Test for Coordinated Polygenic Epistasis in Complex Traits.” *Proceedings of the National Academy of Sciences of the United States of America* 118 (15). https://doi.org/10.1073/pnas.1922305118.
Sherf, Bruce A., Shauna L. Navarro, Rita R. Hannah, Keith V. Wood, and Others. 1996. “Dual-Luciferase Reporter Assay: An Advanced Co-Reporter Technology Integrating Firefly and Renilla Luciferase Assays.” *Promega Notes Magazine* 57 (2): 2–8.
Shin, Ha Youn, Michaela Willi, Kyung HyunYoo, Xianke Zeng, Chaochen Wang, Gil Metser, and Lothar Hennighausen. 2016. “Hierarchy within the Mammary STAT5-Driven Wap Super-Enhancer.” *Nature Genetics* 48 (8): 904–11.
Shin, So-Youn, Eric B. Fauman, Ann-Kristin Petersen, Jan Krumsiek, Rita Santos, Jie Huang, Matthias Arnold, et al. 2014. “An Atlas of Genetic Influences on Human Blood Metabolites.” *Nature Genetics* 46 (6): 543–50.
Shi, Wenqiang, Oriol Fornes, and Wyeth W. Wasserman. 2019. “Gene Expression Models Based on Transcription Factor Binding Events Confer Insight into Functional Cis-Regulatory Variants.” *Bioinformatics* 35 (15): 2610–17.
Simeoni, Ilenia, Jonathan C. Stephens, Fengyuan Hu, Sri V. V. Deevi, Karyn Megy, Tadbir K. Bariana, Claire Lentaigne, et al. 2016. “A High-Throughput Sequencing Test for Diagnosing Inherited Bleeding, Thrombotic, and Platelet Disorders.” *Blood* 127 (23): 2791–2803.
Simeoni, I., O. Shamardina, S. V. V. Deevi, and M. Thomas. 2019. “GRID–Genomics of Rare Immune Disorders: A Highly Sensitive and Specific Diagnostic Gene Panel for Patients with Primary Immunodeficiencies.” *bioRxiv*. https://www.biorxiv.org/content/10.1101/431544v3.abstract.
Sims, Matthew C., Louisa Mayer, Janine H. Collins, Tadbir K. Bariana, Karyn Megy, Cecile Lavenu-Bombled, Denis Seyres, et al. 2020. “Novel Manifestations of Immune Dysregulation and Granule Defects in Gray Platelet Syndrome.” *Blood* 136 (17): 1956–67.
Sivapalaratnam, Suthesh, Janine Collins, and Keith Gomez. 2017. “Diagnosis of Inherited Bleeding Disorders in the Genomic Era.” *British Journal of Haematology* 179 (3): 363–76.
Sivapalaratnam, Suthesh, A. Koneti Rao, Willem Ouwehand, and Kathleen Freson. 2021. “Chapter 119: Inherited Platelet Disorders.” In *Williams Hematology 10th Edition*, edited by Kenneth Kaushansky, Josef T. Prchal, Linda J. Burns, Marshall A. Lichtman, Marcel

Levi, David C. Linch. McGraw-Hill Education.
Sivapalaratnam, Suthesh, Sarah K. Westbury, Jonathan C. Stephens, Daniel Greene, Kate Downes, Anne M. Kelly, Claire Lentaigne, et al. 2017. “Rare Variants in GP1BB Are Responsible for Autosomal Dominant Macrothrombocytopenia.” *Blood* 129 (4): 520–24.
Skipper, Magdalena, Alex Eccleston, Noah Gray, Therese Heemels, Nathalie Le Bot, Barbara Marte, and Ursula Weiss. 2015. “Presenting the Epigenome Roadmap.” *Nature* 518 (7539): 313.
Slatko, Barton E., Andrew F. Gardner, and Frederick M. Ausubel. 2018. “Overview of Next-Generation Sequencing Technologies.” *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]* 122 (1): e59.
Smemo, Scott, Juan J. Tena, Kyoung-Han Kim, Eric R. Gamazon, Noboru J. Sakabe, Carlos Gómez-Marín, Ivy Aneas, et al. 2014. “Obesity-Associated Variants within FTO Form Long-Range Functional Connections with IRX3.” *Nature* 507 (7492): 371–75.
Smith, John Maynard, and John M. Smith. 1989. *Evolutionary Genetics*. Oxford University Press.
Smolkin, M. B., and P. L. Perrotta. 2017. “Chapter 18 - Molecular Diagnostics for Coagulopathies.” In *Diagnostic Molecular Pathology*, edited by William B. Coleman and Gregory J. Tsongalis, 221–33. Academic Press.
Song, W. J., M. G. Sullivan, R. D. Legare, S. Hutchings, X. Tan, D. Kufrin, J. Ratajczak, et al. 1999. “Haploinsufficiency of CBFA2 Causes Familial Thrombocytopenia with Propensity to Develop Acute Myelogenous Leukaemia.” *Nature Genetics* 23 (2): 166–75.
Sood, Raman, Yasuhiko Kamikubo, and Paul Liu. 2017. “Role of RUNX1 in Hematological Malignancies.” *Blood* 129 (15): 2070–82.
Soranzo, Nicole, Augusto Rendon, Christian Gieger, Chris I. Jones, Nicholas A. Watkins, Stephan Menzel, Angela Döring, et al. 2009. “A Novel Variant on Chromosome 7q22.3 Associated with Mean Platelet Volume, Counts, and Function.” *Blood*. https://doi.org/10.1182/blood-2008-10-184234.
Soranzo, Nicole, Tim D. Spector, Massimo Mangino, Brigitte Kühnel, Augusto Rendon, Alexander Teumer, Christina Willenborg, et al. 2009. “A Genome-Wide Meta-Analysis Identifies 22 Loci Associated with Eight Hematological Parameters in the HaemGen Consortium.” *Nature Genetics* 41 (11): 1182–90.
Soskic, Blagoje, Eddie Cano-Gamez, Deborah J. Smyth, Wendy C. Rowan, Nikolina Nakic, Jorge Esparza-Gordillo, Lara Bossini-Castillo, et al. 2019. “Chromatin Activity at GWAS Loci Identifies T Cell States Driving Complex Immune Diseases.” *Nature Genetics* 51 (10): 1486–93.
Spagnolo, Antonio G., Viviana Daloiso, and Paola Parente. 2011. “Biobanks: Ethical and Legal Aspects of the Collection and Storage of Human Biological Material in Italy.” *Human Tissue Research*. https://doi.org/10.1093/acprof:oso/9780199587551.003.0012.
Spector, D. L., and A. I. Lamond. 2011. “Nuclear Speckles.” *Cold Spring Harbor*. https://cshperspectives.cshlp.org/content/3/2/a000646.short.
Spitz, François, and Eileen E. M. Furlong. 2012. “Transcription Factors: From Enhancer Binding to Developmental Control.” *Nature Reviews. Genetics* 13 (9): 613–26.
Spivakov, Mikhail. 2014. “Spurious Transcription Factor Binding: Non-Functional or Genetically Redundant?” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 36 (8): 798–806.
Stadhouders, Ralph, Guillaume J. Filion, and Thomas Graf. 2019. “Transcription Factors and 3D Genome Conformation in Cell-Fate Decisions.” *Nature* 569 (7756): 345–54.
Stanger, Ben Z. 2015. “Cellular Homeostasis and Repair in the Mammalian Liver.” *Annual Review of Physiology* 77: 179–200.
Steensel, Bas van, and Andrew S. Belmont. 2017. “Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression.” *Cell* 169 (5):

780–91.

Steensel, Bas van, and Eileen E. M. Furlong. 2019. “The Role of Transcription in Shaping the Spatial Organization of the Genome.” *Nature Reviews. Molecular Cell Biology* 20 (6): 327–37.

Steimle, V., B. Durand, E. Barras, M. Zufferey, M. R. Hadam, B. Mach, and W. Reith. 1995. “A Novel DNA-Binding Regulatory Factor Is Mutated in Primary MHC Class II Deficiency (bare Lymphocyte Syndrome).” *Genes & Development* 9 (9): 1021–32.

Stelzer, Yonatan, Ido Sagi, Ofra Yanuka, Rachel Eiges, and Nissim Benvenisty. 2014. “The Noncoding RNA IPW Regulates the Imprinted DLK1-DIO3 Locus in an Induced Pluripotent Stem Cell Model of Prader-Willi Syndrome.” *Nature Genetics* 46 (6): 551–57.

Stephenson, Emily, Gary Reynolds, Rachel A. Botting, Fernando J. Calero-Nieto, Michael D. Morgan, Zewen Kelvin Tuong, Karsten Bach, et al. 2021. “Single-Cell Multi-Omics Analysis of the Immune Response in COVID-19.” *Nature Medicine*, April. https://doi.org/10.1038/s41591-021-01329-2.

Stevens, Eric L., Greg Heckenberg, Elisha D. O. Roberson, Joseph D. Baugher, Thomas J. Downey, and Jonathan Pevsner. 2011. “Inference of Relationships in Population Data Using Identity-by-Descent and Identity-by-State.” *PLoS Genetics* 7 (9): e1002287.

Stockklausner, Clemens, Anne-Christine Klotter, Nicole Dickemann, Isabelle N. Kuhlee, Christin M. Duffert, Carolin Kerber, Niels H. Gehring, and Andreas E. Kulozik. 2015. “The Thrombopoietin Receptor P106L Mutation Functionally Separates Receptor Signaling Activity from Thrombopoietin Homeostasis.” *Blood* 125 (7): 1159–69.

Stritt, Simon, Paquita Nurden, Remi Favier, Marie Favier, Silvia Ferioli, Sanjeev K. Gotru, Judith M. M. van Eeuwijk, et al. 2016. “Defects in TRPM7 Channel Function Deregulate Thrombopoiesis through Altered Cellular Mg 2+ Homeostasis and Cytoskeletal Architecture.” *Nature Communications* 7 (1): 1–13.

Stritt, Simon, Paquita Nurden, Ernest Turro, Daniel Greene, Sjoert B. Jansen, Sarah K. Westbury, Romina Petersen, et al. 2016. “A Gain-of-Function Variant in DIAPH1 Causes Dominant Macrothrombocytopenia and Hearing Loss.” *Blood* 127 (23): 2903–14.

Stunnenberg, Hendrik G., International Human Epigenome Consortium, and Martin Hirst. 2016a. “The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery.” *Cell* 167 (5): 1145–49.

———. 2016b. “The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery.” *Cell* 167 (7): 1897.

Sun, Jianlong, Azucena Ramos, Brad Chapman, Jonathan B. Johnnidis, Linda Le, Yu-Jui Ho, Allon Klein, Oliver Hofmann, and Fernando D. Camargo. 2014. “Clonal Dynamics of Native Haematopoiesis.” *Nature* 514 (7522): 322–27.

Su, Yu-Kai, Jia Wei Lin, Jing-Wen Shih, Hao-Yu Chuang, Iat-Hang Fong, Chi-Tai Yeh, and Chien-Min Lin. 2020. “Targeting BC200/miR218-5p Signaling Axis for Overcoming Temozolomide Resistance and Suppressing Glioma Stemness.” *Cells* 9 (8). https://doi.org/10.3390/cells9081859.

Swanson, James M. 2012. “The UK Biobank and Selection Bias.” *The Lancet*.

Symmons, Orsolya, Leslie Pan, Silvia Remeseiro, Tugce Aktas, Felix Klein, Wolfgang Huber, and François Spitz. 2016. “The Shh Topological Domain Facilitates the Action of Remote Enhancers by Reducing the Effects of Genomic Distances.” *Developmental Cell* 39 (5): 529–43.

Taliun, Daniel, Daniel N. Harris, Michael D. Kessler, Jedidiah Carlson, Zachary A. Szpiech, Raul Torres, Sarah A. Gagliano Taliun, et al. 2021. “Sequencing of 53,831 Diverse Genomes from the NHLBI TOPMed Program.” *Nature* 590 (7845): 290–99.

Tang, Fuchou, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan

Xu, Xiaohui Wang, et al. 2009. “mRNA-Seq Whole-Transcriptome Analysis of a Single Cell.” *Nature Methods* 6 (5): 377–82.

Tefferi, Ayalew, Juergen Thiele, Attilio Orazi, Hans Michael Kvasnicka, Tiziano Barbui, Curtis A. Hanson, Giovanni Barosi, et al. 2007. “Proposals and Rationale for Revision of the World Health Organization Diagnostic Criteria for Polycythemia Vera, Essential Thrombocythemia, and Primary Myelofibrosis: Recommendations from an Ad Hoc International Expert Panel.” *Blood, The Journal of the American Society of Hematology* 110 (4): 1092–97.

Thakore, Pratiksha I., Anthony M. D’Ippolito, Lingyun Song, Alexias Safi, Nishkala K. Shivakumar, Ami M. Kabadi, Timothy E. Reddy, Gregory E. Crawford, and Charles A. Gersbach. 2015. “Highly Specific Epigenome Editing by CRISPR-Cas9 Repressors for Silencing of Distal Regulatory Elements.” *Nature Methods* 12 (12): 1143–49.

Thaventhiran, James E. D., Hana Lango Allen, Oliver S. Burren, William Rae, Daniel Greene, Emily Staples, Zinan Zhang, et al. 2020. “Whole-Genome Sequencing of a Sporadic Primary Immunodeficiency Cohort.” *Nature* 583 (7814): 90–95.

“The 100,000 Genomes Project.” 2014. July 21, 2014. https://www.genomicsengland.co.uk/about-genomics-england/the-100000-genomes-project/.

“The Cost of Sequencing a Human Genome.” n.d. Accessed May 9, 2021. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.

The GTEx Consortium. 2015. “The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans.” *Science* 348 (6235): 648–60.

The Physical Foundation of Biology. 1958. Elsevier.

Therizols, Pierre, Robert S. Illingworth, Celine Courilleau, Shelagh Boyle, Andrew J. Wood, and Wendy A. Bickmore. 2014. “Chromatin Decondensation Is Sufficient to Alter Nuclear Organization in Embryonic Stem Cells.” *Science* 346 (6214): 1238–42.

Thiele, Wilko, Jaya Krishnan, Melanie Rothley, Debra Weih, Diana Plaumann, Vanessa Kuch, Luca Quagliata, Herbert A. Weich, and Jonathan P. Sleeman. 2012. “VEGFR-3 Is Expressed on Megakaryocyte Precursors in the Murine Bone Marrow and Plays a Regulatory Role in Megakaryopoiesis.” *Blood* 120 (9): 1899–1907.

Thomas, K. R., and M. R. Capecchi. 1986. “Targeting of Genes to Specific Sites in the Mammalian Genome.” *Cold Spring Harbor Symposia on Quantitative Biology* 51 Pt 2: 1101–13.

Thrombosis, U. K. n.d. “Thrombosis UK.” Accessed August 1, 2021. https://thrombosisuk.org/.

Tijssen, Marloes R., Ana Cvejic, Anagha Joshi, Rebecca L. Hannah, Rita Ferreira, Ariel Forrai, Dana C. Bellissimo, et al. 2011. “Genome-Wide Analysis of Simultaneous GATA1/2, RUNX1, FLI1, and SCL Binding in Megakaryocytes Identifies Hematopoietic Regulators.” *Developmental Cell* 20 (5): 597–609.

Tijssen, Marloes R., Franca di Summa, Sonja van den Oudenrijn, Jaap Jan Zwaginga, C. Ellen van der Schoot, Carlijn Voermans, and Masja de Haas. 2008. “Functional Analysis of Single Amino-Acid Mutations in the Thrombopoietin-Receptor Mpl Underlying Congenital Amegakaryocytic Thrombocytopenia.” *British Journal of Haematology*. https://doi.org/10.1111/j.1365-2141.2008.07139.x.

Torkamani, Ali, Nathan E. Wineinger, and Eric J. Topol. 2018. “The Personal and Clinical Utility of Polygenic Risk Scores.” *Nature Reviews. Genetics* 19 (9): 581–90.

Tormoen, Garth W., Ayesha Khader, András Gruber, and Owen J. T. McCarty. 2013. “Physiological Levels of Blood Coagulation Factors IX and X Control Coagulation Kinetics in Anin Vitromodel of Circulating Tissue Factor.” *Physical Biology*. https://doi.org/10.1088/1478-3975/10/3/036003.

Touboul, Thomas, Nicholas R. F. Hannan, Sébastien Corbineau, Amélie Martinez, Clémence Martinet, Sophie Branchereau, Sylvie Mainot, et al. 2010. “Generation of Functional Hepatocytes from Human Embryonic Stem Cells under Chemically Defined Conditions That Recapitulate Liver Development.” *Hepatology*. https://doi.org/10.1002/hep.23506.

Treangen, Todd J., and Steven L. Salzberg. 2011. “Repetitive DNA and next-Generation Sequencing: Computational Challenges and Solutions.” *Nature Reviews. Genetics* 13 (1): 36–46.

Tripodi, Armando, and Pier Mannuccio Mannucci. 2011. “The Coagulopathy of Chronic Liver Disease.” *The New England Journal of Medicine* 365 (2): 147–56.

Turro, Ernest, William J. Astle, Karyn Megy, Stefan Gräf, Daniel Greene, Olga Shamardina, Hana Lango Allen, et al. 2020a. “Whole-Genome Sequencing of Patients with Rare Diseases in a National Health System.” *Nature* 583 (7814): 96–102.

———. 2020b. “Whole-Genome Sequencing of Patients with Rare Diseases in a National Health System.” *Nature* 583 (7814): 96–102.

Turro, Ernest, William J. Astle, and Simon Tavaré. 2014. “Flexible Analysis of RNA-Seq Data Using Mixed Effects Models.” *Bioinformatics* 30 (2): 180–88.

Turro, Ernest, Daniel Greene, Anouck Wijgaerts, Chantal Thys, Claire Lentaigne, Tadbir K. Bariana, Sarah K. Westbury, et al. 2016. “A Dominant Gain-of-Function Mutation in Universal Tyrosine Kinase SRC Causes Thrombocytopenia, Myelofibrosis, Bleeding, and Bone Pathologies.” *Science Translational Medicine* 8 (328): 328ra30.

Tweedie, Susan, Bryony Braschi, Kristian Gray, Tamsin E. M. Jones, Ruth L. Seal, Bethan Yates, and Elspeth A. Bruford. 2021. “Genenames.org: The HGNC and VGNC Resources in 2021.” *Nucleic Acids Research* 49 (D1): D939–46.

Tycko, Josh, Michael Wainberg, Georgi K. Marinov, Oana Ursu, Gaelen T. Hess, Braeden K. Ego, Aradhana, et al. 2019. “Mitigation of off-Target Toxicity in CRISPR-Cas9 Screens for Essential Non-Coding Elements.” *Nature Communications* 10 (1): 4063.

UKHCDO-Annual-Report-2019.pdf. n.d.

Undas, Anetta, Kathleen E. Brummel-Ziedins, and Kenneth G. Mann. 2005. “Statins and Blood Coagulation.” *Arteriosclerosis, Thrombosis, and Vascular Biology* 25 (2): 287–94.

Urrutia, Raul. 2003. “KRAB-Containing Zinc-Finger Repressor Proteins.” *Genome Biology* 4 (10): 231.

Vanhoorelbeke, K., H. Ulrichts, A. Schoolmeester, and H. Deckmyn. 2003. “Inhibition of Platelet Adhesion to Collagen as a New Target for Antithrombotic Drugs.” *Current Drug Targets. Cardiovascular & Haematological Disorders* 3 (2): 125–40.

Van Hout, Cristopher V., Ioanna Tachmazidou, Joshua D. Backman, Joshua D. Hoffman, Daren Liu, Ashutosh K. Pandey, Claudia Gonzaga-Jauregui, et al. 2020. “Exome Sequencing and Characterization of 49,960 Individuals in the UK Biobank.” *Nature* 586 (7831): 749–56.

Veitia, Reiner A. 2002. “Exploring the Etiology of Haploinsufficiency.” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 24 (2): 175–84.

Vekemans, Katrien, and Filip Braet. 2005. “Structural and Functional Aspects of the Liver and Liver Sinusoidal Cells in Relation to Colon Carcinoma Metastasis.” *World Journal of Gastroenterology: WJG* 11 (33): 5095–5102.

Velten, Lars, Simon F. Haas, Simon Raffel, Sandra Blaszkiewicz, Saiful Islam, Bianca P. Hennig, Christoph Hirche, et al. 2017. “Human Haematopoietic Stem Cell Lineage Commitment Is a Continuous Process.” *Nature Cell Biology* 19 (4): 271–81.

Versteeg, Henri H., Johan W. M. Heemskerk, Marcel Levi, and Pieter H. Reitsma. 2013. “New Fundamentals in Hemostasis.” *Physiological Reviews* 93 (1): 327–58.

Vet, Edwin C. J. M. de, Begoña Aguado, and R. Duncan Campbell. 2001. “G6b, a Novel Immunoglobulin Superfamily Member Encoded in the Human Major Histocompatibility Complex, Interacts with SHP-1 and SHP-2 *.” *The Journal of Biological Chemistry* 276

(45): 42070–76.
Victor Hoffbrand, A., Douglas R. Higgs, David M. Keeling, and Atul B. Mehta. 2016. *Postgraduate Haematology*. John Wiley & Sons.
Vieux-Rochas, Maxence, Pierre J. Fabre, Marion Leleu, Denis Duboule, and Daan Noordermeer. 2015. “Clustering of Mammalian Hox Genes with Other H3K27me3 Targets within an Active Nuclear Domain.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (15): 4672–77.
Villard, J., M. Peretti, K. Masternak, E. Barras, G. Caretti, R. Mantovani, and W. Reith. 2000. “A Functionally Essential Domain of RFX5 Mediates Activation of Major Histocompatibility Complex Class II Promoters by Promoting Cooperative Binding between RFX and NF-Y.” *Molecular and Cellular Biology* 20 (10): 3364–76.
Vincent, Lisa M., Sinh Tran, Ruzica Livaja, Tracy A. Bensend, Dianna M. Milewicz, and Björn Dahlbäck. 2013. “Coagulation Factor V(A2440G) Causes East Texas Bleeding Disorder via TFPIα.” *The Journal of Clinical Investigation* 123 (9): 3777–87.
Viñuela, Ana, Andrew A. Brown, Juan Fernandez, Mun-Gwan Hong, Caroline A. Brorsson, Robert W. Koivula, David Davtian, et al. 2021. “Genetic Analysis of Blood Molecular Phenotypes Reveals Regulatory Networks Affecting Complex Traits: A DIRECT Study.” *bioRxiv*. medRxiv. https://doi.org/10.1101/2021.03.26.21254347.
Visel, Axel, Shyam Prabhakar, Jennifer A. Akiyama, Malak Shoukry, Keith D. Lewis, Amy Holt, Ingrid Plajzer-Frick, Veena Afzal, Edward M. Rubin, and Len A. Pennacchio. 2008. “Ultraconservation Identifies a Small Subset of Extremely Constrained Developmental Enhancers.” *Nature Genetics* 40 (2): 158–60.
Visel, Axel, Leila Taher, Hani Girgis, Dalit May, Olga Golonzhka, Renee V. Hoch, Gabriel L. McKinsey, et al. 2013. “A High-Resolution Enhancer Atlas of the Developing Telencephalon.” *Cell* 152 (4): 895–908.
Visscher, Peter M., William G. Hill, and Naomi R. Wray. 2008. “Heritability in the Genomics Era — Concepts and Misconceptions.” *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2322.
Vuckovic, Dragana, Erik L. Bao, Parsa Akbari, Caleb A. Lareau, Abdou Mousas, Tao Jiang, Ming-Huei Chen, et al. 2020. “The Polygenic and Monogenic Basis of Blood Traits and Diseases.” *Cell* 182 (5): 1214–31.e11.
Waldmann, Robert, Johnw Rebuck, Hidehiko Saito, Josephp Abraham, June Caldwell, and Oscard Ratnoff. 1975. “FITZGERALD FACTOR: A HITHERTO UNRECOGNISED COAGULATION FACTOR.” *The Lancet*. https://doi.org/10.1016/s0140-6736(75)92008-5.
Wallace, Chris. 2020. “Eliciting Priors and Relaxing the Single Causal Variant Assumption in Colocalisation Analyses.” *PLoS Genetics* 16 (4): e1008720.
Wang, Zhibin, Chongzhi Zang, Jeffrey A. Rosenfeld, Dustin E. Schones, Artem Barski, Suresh Cuddapah, Kairong Cui, et al. 2008. “Combinatorial Patterns of Histone Acetylations and Methylations in the Human Genome.” *Nature Genetics* 40 (7): 897–903.
Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. “RNA-Seq: A Revolutionary Tool for Transcriptomics.” *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2484.
Watson, J. D., and F. H. Crick. 1953. “Molecular Structure of Nucleic Acids; a Structure for Deoxyribose Nucleic Acid.” *Nature* 171 (4356): 737–38.
Watson, S. P., J. M. J. Herbert, and A. Y. Pollitt. 2010. “GPVI and CLEC-2 in Hemostasis and Vascular Integrity.” *Journal of Thrombosis and Haemostasis: JTH* 8 (7): 1456–67.
Wei, Wei, Salih Tuna, Michael J. Keogh, Katherine R. Smith, Timothy J. Aitman, Phil L. Beales, David L. Bennett, et al. 2019. “Germline Selection Shapes Human Mitochondrial DNA Diversity.” *Science* 364 (6442). https://doi.org/10.1126/science.aau6520.

Wendling, F., E. Maraskovsky, N. Debili, C. Florindo, M. Teepe, M. Titeux, N. Methia, J. Breton-Gorius, D. Cosman, and W. Vainchenker. 1994. “cMpl Ligand Is a Humoral Regulator of Megakaryocytopoiesis.” *Nature* 369 (6481): 571–74.

West, Adam G., Miklos Gaszner, and Gary Felsenfeld. 2002. “Insulators: Many Functions, Many Mechanisms.” *Genes & Development* 16 (3): 271–88.

Westbury, Sarah K., Ernest Turro, Daniel Greene, Claire Lentaigne, Anne M. Kelly, Tadbir K. Bariana, Ilenia Simeoni, et al. 2015. “Human Phenotype Ontology Annotation and Cluster Analysis to Unravel Genetic Defects in 707 Cases with Unexplained Bleeding and Platelet Disorders.” *Genome Medicine* 7 (1): 36.

White, Michael A. 2015. “Understanding How Cis-Regulatory Function Is Encoded in DNA Sequence Using Massively Parallel Reporter Assays and Designed Sequences.” *Genomics* 106 (3): 165–70.

Whyte, Warren A., David A. Orlando, Denes Hnisz, Brian J. Abraham, Charles Y. Lin, Michael H. Kagey, Peter B. Rahl, Tong Ihn Lee, and Richard A. Young. 2013. “Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes.” *Cell*. https://doi.org/10.1016/j.cell.2013.03.035.

Wienberg, J. 2005. “Fluorescence in Situ Hybridization to Chromosomes as a Tool to Understand Human and Primate Genome Evolution.” *Cytogenetic and Genome Research* 108 (1-3): 139–60.

Wijchers, Patrick J., Geert Geeven, Michael Eyres, Atze J. Bergsma, Mark Janssen, Marjon Verstegen, Yun Zhu, et al. 2015. “Characterization and Dynamics of Pericentromere-Associated Domains in Mice.” *Genome Research* 25 (7): 958–69.

Wijchers, Patrick J., Peter H. L. Krijger, Geert Geeven, Yun Zhu, Annette Denker, Marjon J. A. M. Verstegen, Christian Valdes-Quezada, et al. 2016. “Cause and Consequence of Tethering a SubTAD to Different Nuclear Compartments.” *Molecular Cell* 61 (3): 461–73.

Wilkie, Andrew O. M. 2003. “Why Study Human Limb Malformations?” *Journal of Anatomy* 202 (1): 27–35.

Wilkins, Bridget S., Wendy N. Erber, David Bareford, Georgina Buck, Keith Wheatley, Clare L. East, Beverley Paul, Claire N. Harrison, Anthony R. Green, and Peter J. Campbell. 2008. “Bone Marrow Pathology in Essential Thrombocythemia: Interobserver Reliability and Utility for Identifying Disease Subtypes.” *Blood* 111 (1): 60–70.

Willer, Cristen J., Serena Sanna, Anne U. Jackson, Angelo Scuteri, Lori L. Bonnycastle, Robert Clarke, Simon C. Heath, et al. 2008. “Newly Identified Loci That Influence Lipid Concentrations and Risk of Coronary Artery Disease.” *Nature Genetics* 40 (2): 161–69.

Wilson, Anne, Elisa Laurenti, Gabriela Oser, Richard C. van der Wath, William Blanco-Bose, Maike Jaworski, Sandra Offner, et al. 2008. “Hematopoietic Stem Cells Reversibly Switch from Dormancy to Self-Renewal during Homeostasis and Repair.” *Cell* 135 (6): 1118–29.

Wilson, Nicola K., David G. Kent, Florian Buettner, Mona Shehata, Iain C. Macaulay, Fernando J. Calero-Nieto, Manuel Sánchez Castillo, et al. 2015. “Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations.” *Cell Stem Cell* 16 (6): 712–24.

Wingett, Steven, Philip Ewels, Mayra Furlan-Magaril, Takashi Nagano, Stefan Schoenfelder, Peter Fraser, and Simon Andrews. 2015. “HiCUP: Pipeline for Mapping and Processing Hi-C Data.” *F1000Research* 4 (November): 1310.

Wit, Elzo de, Erica S. M. Vos, Sjoerd J. B. Holwerda, Christian Valdes-Quezada, Marjon J. A. M. Verstegen, Hans Teunissen, Erik Splinter, Patrick J. Wijchers, Peter H. L. Krijger, and Wouter de Laat. 2015. “CTCF Binding Polarity Determines Chromatin Looping.” *Molecular Cell* 60 (4): 676–84.

Wonkam, Ambroise. 2021. “Sequence Three Million Genomes across Africa.” *Nature* 590

(7845): 209–11.
Wood, K. V., J. R. de Wet, N. Dewji, and M. DeLuca. 1984. “Synthesis of Active Firefly Luciferase by in Vitro Translation of RNA Obtained from Adult Lanterns.” *Biochemical and Biophysical Research Communications* 124 (2): 592–96.
Wray, Naomi R., Cisca Wijmenga, Patrick F. Sullivan, Jian Yang, and Peter M. Visscher. 2018. “Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model.” *Cell* 173 (7): 1573–80.
Wright, Caroline F., Tomas W. Fitzgerald, Wendy D. Jones, Stephen Clayton, Jeremy F. McRae, Margriet van Kogelenberg, Daniel A. King, et al. 2015. “Genetic Diagnosis of Developmental Disorders in the DDD Study: A Scalable Analysis of Genome-Wide Research Data.” *The Lancet* 385 (9975): 1305–14.
Wright, S. 1920. “The Relative Importance of Heredity and Environment in Determining the Piebald Pattern of Guinea-Pigs.” *Proceedings of the National Academy of Sciences of the United States of America* 6 (6): 320–32.
WTCCC. 2007. “Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls.” *Nature* 447 (7145): 661–78.
Xiang, Jiale, Jiyun Yang, Lisha Chen, Qiang Chen, Haiyan Yang, Chengcheng Sun, Qing Zhou, and Zhiyu Peng. 2020. “Reinterpretation of Common Pathogenic Variants in ClinVar Revealed a High Proportion of Downgrades.” *Scientific Reports*. https://doi.org/10.1038/s41598-019-57335-5.
Xin, Dedong, Landian Hu, and Xiangyin Kong. 2008. “Alternative Promoters Influence Alternative Splicing at the Genomic Level.” *PloS One* 3 (6): e2377.
Xu, Duo, Omer Gokcumen, and Ekta Khurana. 2020. “Loss-of-Function Tolerance of Enhancers in the Human Genome.” *PLoS Genetics* 16 (4): e1008663.
Yang, Jian, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. 2011. “GCTA: A Tool for Genome-Wide Complex Trait Analysis.” *American Journal of Human Genetics* 88 (1): 76–82.
Yardımcı, Galip Gürkan, Hakan Ozadam, Michael E. G. Sauria, Oana Ursu, Koon-Kiu Yan, Tao Yang, Abhijit Chakraborty, et al. 2019. “Measuring the Reproducibility and Quality of Hi-C Data.” *Genome Biology* 20 (1): 57.
Yeo, Nan Cher, Alejandro Chavez, Alissa Lance-Byrne, Yingleong Chan, David Menn, Denitsa Milanova, Chih-Chung Kuo, et al. 2018. “An Enhanced CRISPR Repressor for Targeted Mammalian Gene Regulation.” *Nature Methods* 15 (8): 611–16.
Ying, Yue, Xingyu Yang, Kai Zhao, Jifang Mao, Ying Kuang, Zhugang Wang, Ruilin Sun, and Jian Fei. 2015. “The Krüppel-Associated Box Repressor Domain Induces Reversible and Irreversible Regulation of Endogenous Mouse Genes by Mediating Different Chromatin States.” *Nucleic Acids Research* 43 (3): 1549–61.
Young, Alexander I. 2019. “Solving the Missing Heritability Problem.” *PLoS Genetics* 15 (6): e1008222.
Young, Richard A. 2011. “Control of the Embryonic Stem Cell State.” *Cell* 144 (6): 940–54.
Youssefian, T., and E. M. Cramer. 2000. “Megakaryocyte Dense Granule Components Are Sorted in Multivesicular Bodies.” *Blood* 95 (12): 4004–7.
Zeigler, F. C., F. de Sauvage, H. R. Widmer, G. A. Keller, C. Donahue, R. D. Schreiber, B. Malloy, P. Hass, D. Eaton, and W. Matthews. 1994. “In Vitro Megakaryocytopoietic and Thrombopoietic Activity of c-Mpl Ligand (TPO) on Purified Murine Hematopoietic Stem Cells.” *Blood* 84 (12): 4045–52.
Zerbino, Daniel R., Nathan Johnson, Thomas Juetteman, Dan Sheppard, Steven P. Wilder, Ilias Lavidas, Michael Nuhn, et al. 2016. “Ensembl Regulation Resources.” *Database: The Journal of Biological Databases and Curation* 2016 (February). https://doi.org/10.1093/database/bav119.
Zerbino, Daniel R., Steven P. Wilder, Nathan Johnson, Thomas Juettemann, and Paul R.

Flicek. 2015. “The Ensembl Regulatory Build.” *Genome Biology* 16 (March): 56.
Zhang, Bin, Marta Spreafico, Chunlei Zheng, Angela Yang, Petra Platzer, Michael U. Callaghan, Zekai Avci, et al. 2008. “Genotype-Phenotype Correlation in Combined Deficiency of Factor V and Factor VIII.” *Blood* 111 (12): 5592–5600.
Zhao, Hao, Zhifu Sun, Jing Wang, Haojie Huang, Jean-Pierre Kocher, and Liguo Wang. 2014. “CrossMap: A Versatile Tool for Coordinate Conversion between Genome Assemblies.” *Bioinformatics* 30 (7): 1006–7.
Zheng, Hui, and Wei Xie. 2019. “The Role of 3D Genome Organization in Development and Cell Differentiation.” *Nature Reviews. Molecular Cell Biology* 20 (9): 535–50.
Zheng, Jian-Tao, Cui-Xiang Lin, Zhao-Yu Fang, and Hong-Dong Li. 2020. “Intron Retention as a Mode for RNA-Seq Data Analysis.” *Frontiers in Genetics* 11 (July): 586.
Zhou, Zhou Zhou, Ming-Jiang Xu, and Bin Gao. 2016. “Hepatocytes: A Key Cell Type for Innate Immunity.” *Cellular & Molecular Immunology*. https://doi.org/10.1038/cmi.2015.97.
Zollner, Sebastian, and Jonathan K. Pritchard. 2007. “Overcoming the Winner’s Curse: Estimating Penetrance Parameters from Case-Control Data.” *American Journal of Human Genetics* 80 (4): 605–15.
Zou, Xueqing, Genomics England Research Consortium, Gene Ching Chiek Koh, Arjun Scott Nanda, Andrea Degasperi, Katie Urgo, Theodoros I. Roumeliotis, et al. 2021. “A Systematic CRISPR Screen Defines Mutational Mechanisms Underpinning Signatures Caused by Replication Errors and Endogenous DNA Damage.” *Nature Cancer*. https://doi.org/10.1038/s43018-021-00200-0.
Zrimec, Jan, Christoph S. Börlin, Filip Buric, Azam Sheikh Muhammad, Rhongzen Chen, Verena Siewers, Vilhelm Verendel, Jens Nielsen, Mats Töpel, and Aleksej Zelezniak. 2020. “Deep Learning Suggests That Gene Expression Is Encoded in All Parts of a Co-Evolving Interacting Gene Regulatory Structure.” *Nature Communications* 11 (1): 6141.
Zuin, Jessica, Jesse R. Dixon, Michael I. J. A. van der Reijden, Zhen Ye, Petros Kolovos, Rutger W. W. Brouwer, Mariëtte P. C. van de Corput, et al. 2014. “Cohesin and CTCF Differentially Affect Chromatin Architecture and Gene Expression in Human Cells.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (3): 996–1001.