Inherited Platelet Disorders: Towards DNA-based diagnosis

Claire Lentaigne\textsuperscript{1,2}, Kathleen Freson\textsuperscript{3}, Michael A Laffan\textsuperscript{1,2}, Ernest Turro\textsuperscript{4-7}, Willem H Ouwehand\textsuperscript{4,5,7,8}

\textsuperscript{1}Centre for Haematology, Hammersmith Campus, Imperial College Academic Health Sciences Centre, Imperial College London, London, United Kingdom.
\textsuperscript{2}Imperial College Healthcare NHS Trust, Du Cane Road, London, United Kingdom.
\textsuperscript{3}Department of Cardiovascular Sciences, Center for Molecular and Vascular Biology, University of Leuven, Belgium.
\textsuperscript{4}Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge, United Kingdom.
\textsuperscript{5}NHS Blood and Transplant, Cambridge Biomedical Campus, Cambridge, United Kingdom.
\textsuperscript{6}Medical Research Council Biostatistics Unit, Cambridge Institute of Public Health, Cambridge Biomedical Campus, Cambridge, United Kingdom.
\textsuperscript{7}NIHR BioResource - Rare Diseases, Cambridge University Hospitals, Cambridge Biomedical Campus, Cambridge, United Kingdom.
\textsuperscript{8}Human Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom.

~on behalf of the BRIDGE Bleeding, Thrombotic and Platelet Disorders and ThromboGenomics Consortia.

Copyright © 2016 American Society of Hematology
Abstract

Variations in platelet number, volume and function are largely genetically controlled and many loci associated with platelet traits have been identified by genome wide association studies (GWAS)\(^1\). The genome also contains a large number of rare variants, of which a tiny fraction underlie the inherited diseases of man. Research over the past three decades have led to the discovery of 51 genes harbouring variants responsible for inherited platelet disorders (IPDs). However, the majority of patients with an IPD still do not receive a molecular diagnosis. Alongside the scientific interest, molecular or genetic diagnosis is important for patients. There is increasing recognition that a number of IPDs are associated with severe pathologies, including an increased risk of malignancy and a definitive diagnosis can inform prognosis and care.

In this review we give an overview of these disorders grouped according to their effect on platelet biology and their clinical characteristics. We also discuss the challenge of identifying candidate genes and causal variants therein, how IPDs have been historically diagnosed and how this is changing with the introduction of high-throughput sequencing (HTS). Finally, we describe how integration of large genomic, epigenomic and phenotypic datasets, including whole genome sequencing (WGS) data, GWAS, epigenomic profiling, protein-protein interaction networks and standardised clinical phenotype coding, will drive the discovery of novel mechanisms of disease in the near future to improve patient diagnosis and management.
Rare Inherited Platelet Disorders

There is marked genetic heterogeneity amongst IPDs and in this section we survey the 51 genes known to harbour variants responsible for IPDs (henceforth, “IPD genes”), classified according to their principal known effect on platelet biology. They encode an array of molecules of diverse function, reflecting the complex and tightly regulated processes of megakaryopoiesis, platelet formation and platelet function (Fig. 1). For some genes, their role in platelet biology is less well defined and this is also discussed. Many IPD genes are widely transcribed across blood cell types (Fig. 2) and other tissues. Hence, patients with an IPD frequently present with pathologies reaching well beyond the blood system7 (23 genes marked with *, Fig. 1).

Megakaryopoiesis and Platelet Formation

Studies in patients with thrombocytopenia have highlighted the important roles of the Thpo/Mpl signalling pathway and transcriptional regulation in both early and late stages of megakaryopoiesis. Rare variants in THPO and MPL, the genes encoding thrombopoietin3,4 and its receptor Mpl5,6 cause congenital thrombocytopenia and seven IPD genes encode transcription factors (GATA17,8, RUNX19, FLI110, ETV611, HOXA1112, MECOM13 and GFI1B14,15) expressed in hematopoietic stem and progenitor cells (Fig. 2). Rare variants in RBM8A16 and ANKRD2617 have been shown to affect transcription factor binding (of Mecom and Runx1, respectively), altering signalling through the Thpo/Mpl pathway and resulting in thrombocytopenia, but their exact role in megakaryopoiesis is not yet clear.

Six IPD genes regulate the megakaryocyte cytoskeleton and the principal effect of deleterious variants is macrothrombocytopenia due to aberrant proplatelet formation (variants in MYH918, ACTN119, FLNA20, TUBB121, DIAPH122). Causal variants in the Wiskott-Aldrich Syndrome gene WAS also lead to defective proplatelet formation alongside neutropenia and eczema but, in contrast to other cytoskeletal disorders, the platelets are small23,24. So far six genes encoding proteins that primarily influence granule formation, trafficking or secretion (NBEAL225-27, NBEA28, VPS33B29, VIPAS3930, STXBP231,32, LYST33) and nine genes that cause Hermansky Pudlak Syndrome (HPS), a δ-
granule platelet disorder (HPS1, AP3B1, HPS3, HPS4, HPS5, HPS6, DTNB1, BLOC1S3, BLOC1S6) have been identified. Defects in transmembrane glycoprotein (GP) signalling pathways can lead to abnormal platelet function and thrombocytopenia with giant platelets through abnormal proplatelet formation (GP1BA, GP1BB, GP9) and thrombasthenia only (ITGA2B, ITGB3). Gain-of-function (GOF) variants in GP1BA can cause enhanced binding to Von Willebrand Factor (VWF) leading to “platelet-type” Von Willebrand’s Disease (VWD), a phenocopy of Type 2B VWD, a disorder caused by GOF variants in VWF affecting the function of its A1 domain. In both cases premature interaction between Gp1ba and the Vwf A1 domain results in bleeding characterised by thrombocytopenia and a loss of high molecular weight Vwf multimers. DNA analysis is generally required to distinguish between these two disorders. The role of some IPD genes in megakaryopoiesis and proplatelet formation is less well defined. The recent discovery of a GOF variant in SRC causing abnormal megakaryopoiesis and thrombocytopenia has cemented the central role of this universal tyrosine kinase in a variety of megakaryocyte signalling pathways and podosome formation; CYCS is expected to play a role in apoptosis; certain GOF variants in STIM1 (causal of Störmorken syndrome) result in thrombocytopenia with abnormal platelet function and bleeding, whereas loss-of-function (LOF) variants in STIM1 cause an autoimmune thrombocytopenia. LOF variants in Orai1 have been described in a mild Störmorken-like syndrome but the platelet defects are not well established. Both Orai1 and Stim1 are involved in calcium homeostasis and IPD-causing variants consequently affect platelet signalling but further research is required into their role in platelet formation. Novel missense variants were recently identified in SLFN14 in four unrelated pedigrees with moderate thrombocytopenia and platelet secretion defects. Though the mechanism is not clear, a defect in platelet formation was observed. Finally, Quebec platelet disorder, so far confined to French-Canadians, is caused by a tandem duplication in PLAU, leading to overexpression of urokinase and is characterised by thrombocytopenia, degradation of α-granule contents with normal granule structure, decreased aggregation in response to epinephrine and late onset bleeding.
**Platelet Function**

There are 14 IPD genes primarily affecting various aspects of platelet function. Six genes encode GPs which function as receptors for the hemostatically important ligands Vwf (GP1BA, GP1BB, GP9 gene defects causing Bernard Soulier syndrome (BSS))\(^ {43}\), fibrinogen (ITGA2, ITGB3 gene defects causing Glanzmann Thrombasthenia (GT))\(^ {45}\) and collagen (GP6)\(^ {56}\). The typical mode of inheritance for GT is autosomal recessive with LOF variants of ITGA2B or ITGB3. This classical thrombasthenia is reviewed extensively elsewhere\(^ {45}\), but it is worth noting that, again, rare and dominant GOF variants in ITGB3 have also been described, leading to enhanced fibrinogen binding combined with bleeding\(^ {57}\) and a mild thrombocytopenia exacerbated during pregnancy but without bleeding\(^ {58}\). Mutations in FERM1 affect integrin inside-out signalling causing a GT-like platelet phenotype with bleeding and associated with a type III leukocyte adhesion disorder\(^ {59}\). A homozygous variant in RASGRP2 was discovered as a cause of another mild GT-like platelet phenotype in a consanguineous pedigree. RASGRP2 encodes for a guanidine exchange factor that regulates Rap1b activation and consequently has a major effect on αIIbβ3 (encoded by ITGA2B, ITGB3) signalling in platelets\(^ {60}\).

G-protein coupled receptors (GPCRs) are another main type of multispan transmembrane receptors and signalling defects due to variants in P2YR12\(^ {61}\) and TBXAS1\(^ {62}\) (encoding the GPCRs for ADP and thromboxane respectively) or their downstream effectors (TBXAS1, PLA2G4A)\(^ {63}\) have been linked to IPDs. Scott syndrome does not fall into the aforementioned categories and is caused by autosomal recessive variants in ANO6, which encodes a multispan transmembrane protein involved in phospholipid scrambling\(^ {65}\). Platelets from these cases cannot properly express phosphatidylserine on their surface, which leads to defective coagulation.

**Current Approaches to Diagnosis**

IPD diagnosis is straightforward in the major platelet function disorders such as BSS and GT, which often present with severe bleeding symptoms early in life and are easily recognised by the pattern of platelet aggregation defects\(^ {45,66}\). This is often supplemented by assessment of the storage pool either directly by nucleotide assay or...
lumi-aggregometry. In some IPDs, the platelet function defect and impaired hemostasis are part of well-defined syndromes, e.g. Chediak Higashi Syndrome (CHS) and HPS, where the platelet δ-granule defect is typically associated with immune deficiency or ocular albinism, respectively. The presence of syndromic features can help in recognition and diagnosis but diagnosis remains challenging for the majority of IPDs, which often have a mild platelet phenotype and are clinically heterogeneous. Diagnosis is further complicated by the fact that for many IPDs, the platelet count is within normal ranges and the disorder may only become apparent after a hemostatic challenge or if cases present with accompanying pathologies in other organ systems, including malignancies.

Establishing a conclusive molecular diagnosis is the bedrock of good hematological practice because it informs optimal treatment and can provide clarity about disease progression. For IPDs, this is particularly important for the severe cases and those associated with early onset clinical pathologies such as myelofibrosis, lung fibrosis, renal insufficiency and malignancy. Thrombocytopenias caused by variants in RUNX1, ETV6 and ANKRD26 are associated with increased risk of myeloid malignancy while for WAS and amegakaryocytic thrombocytopenia caused by MPL variants, treatment by allogeneic hematopoietic stem cell transplant or gene therapy may require consideration. Moreover, genetic counselling can be provided if the diagnosis is confirmed at the DNA level. Current guidelines favour a tiered approach to IPD diagnosis. DNA analysis by Sanger sequencing is at the fourth and final tier and often not applied because of its limited availability and costs. Moreover, it is used primarily to confirm an already clinically suspected genetic diagnosis and targets only a single or small group of genes. In the majority of IPDs, a single candidate gene is not readily apparent from standard laboratory tests. Consequently, a molecular diagnosis is given in only a minority of patients and even when a genetic defect is identified, the number of independent cases remains small for the majority of IPDs. Fewer than five unrelated probands have been identified for IPD genes P2Y12R, GP6, TBXA2R, PLAU, ANO6, which all were identified before the era of HTS and there is a paucity of larger case series for individual IPD genes with the exception of ACTN1 and ANKRD26.
This lack of genetic diagnosis not only hampers our ability to provide accurate information on prognosis and optimal management, but the small case numbers also impact on our ability to interpret the pathogenicity of particular variants. The advent of HTS is set to change this. In this issue of Blood, the ThromboGenomics consortium reports on a targeted HTS panel of 76 genes (63 genes in the panel reported in their publication and a further 13 genes added in the currently available version) covering the inherited bleeding, thrombotic and platelet disorders (BPDs), bringing an affordable molecular diagnosis within reach. The application of HTS will simplify the diagnostic process and reduce delay, as we discuss below.

IPD Gene Discovery to Date

Until recently, the majority of IPD genes have been discovered by candidate gene and linkage studies. These approaches resulted in the identification of 36 IPD genes by 2010, when HTS became available (Fig. 3). In 2011 three groups reported on NBEAL2 being the causal gene for Gray Platelet syndrome (GPS) and two of these discoveries were made possible by HTS. In the past five years 14 additional IPD genes have been identified, 11 by applying HTS (Fig. 3). Thrombocytopenia with absent radii (TAR) is an example of a syndrome for which the genetic roots remained elusive despite being clinically well defined for over two decades. It has a thus far unique genetic architecture typically involving a micro-deletion on one haplotype and a low-frequency regulatory variant on the other haplotype of RBM8A. The most commonly implicated regulatory variant results in reduced binding of the transcription factor Mecom leading to an insufficiency of the protein Y14 in megakaryocytes, causing thrombocytopenia. In addition, WES was also instrumental in discovering variants in GFI1B responsible for a GPS-like syndrome; ACTN1, ETV6, STIM1, DIAPH1, SRC, SLFN14 and MECOM for inherited thrombocytopenias; and RASGRP2 for GT-like disease.

The results of the rare diseases pilot phase of the 100,000 Genomes Project indicate that a large number of IPD genes remain to be discovered. Genome sequencing results of the DNA samples of hundreds of probands with uncharacterised BPDs, analysed using assigned Human Phenotype Ontology (HPO) terms have helped identify
pathogenic variants in known IPD genes in almost 20% of cases. New clustering algorithms to group cases with similar phenotypes, have been used to identify two novel IPD genes (DIAPH1 and SRC) and several putative ones. This suggests that the majority of cases either harbour pathogenic variants in unknown genes or regulatory regions or are the result of a digenic mode of inheritance.

The lack of gene discoveries on a global scale for δ-granule Storage Pool Disease (δ-SPD) is a case in point. The true prevalence of δ-SPD is unclear, as the definition is not always consistent between studies, but defects in granule release are relatively frequent amongst the mild IPDs. The greatest diagnostic success for δ-SPD has been in the multi-system syndromes such as CHS and HPS but most patients with non-syndromic δ-SPD remain undiagnosed. In on-going large-scale HTS projects such as GAPP and BRIDGE-BPD, the lack of gene discovery for δ-SPD, even in these large patient cohorts, highlights a need to analyse the non-coding regulatory regions of the genome via WGS, whilst also exploring novel methods of data analysis and integration. We discuss such methods in more detail in the final section.

Assigning Pathogenicity to Novel Variants: Use of Public Databases

With the advent of HTS it has become possible to survey the exonic fraction of large numbers of genomes for variants by WES. Exomes comprise ~2% of the genome (~64 Mb) while current WGS captures 98.3% of the 3.2 billion bases of the genome at a minimum of 15x coverage. Information about pathogenic and likely pathogenic variants is maintained in databases but until recently it was not possible to verify their allele frequencies in the general population. Several initiatives such as the 1000 Genomes and UK10K projects and initiatives to aggregate results from many smaller WES projects, as achieved by the Exome Aggregation Consortium (ExAC), have provided information on exonic variants observed in more than 71,000 individuals of mainly Caucasoid ancestry. This catalogue has made an immense contribution to improving the accuracy of assignment of pathogenicity to DNA variants observed in IPD cases. There is still a relative paucity of sequencing data from individuals of other ethnicities. Consequently, a larger number of variants absent from control samples is
observed in their DNA, making it harder to distinguish variants which are relatively common in particular populations from variants responsible for disease.

HTS has also made possible large-scale WGS projects such as the 100,000 Genomes Project\textsuperscript{83} which are complemented by GWAS in large population cohorts such as the UK Biobank of 500,000 healthy individuals\textsuperscript{92}. Both projects link genotypes with health and social care records and will lead to a better understanding of the relationship between variants and diseases. In particular, they will allow more accurate assignment of pathogenicity to rare variants underlying the thousands of inherited diseases. Single nucleotide variants (SNV) can be called accurately by WES but calling of short insertions and deletions and especially of copy number and structural variants is more challenging. Here WGS can achieve far greater sensitivity and specificity than WES. WGS will therefore further improve the accuracy of frequencies for all classes of variants in databases, make it easier to discover the genetic determinants of rare inherited diseases and also open up opportunities to explore the non-coding regulatory part of the genome.

Germline variants are inevitable consequences of meiosis and DNA repair, and accumulate over generations. Given the existence of many non-pathogenic variants in any individual’s genome, the main challenge faced by researchers when interpreting HTS data of an IPD case is determining which variants are causing the disorder. The correct identification of novel causal variants critically depends on their allele frequency in relevant control samples\textsuperscript{93}. Additionally, studies are required to uncover the function of novel genes and the consequences of candidate rare variants. It has been shown that specific GOF variants in \textit{DIAPH1} and \textit{SRC} can cause a defect in megakaryopoiesis while this is not expected of LOF variants\textsuperscript{22,48}. Similar observations can be made for \textit{SFLN14} were all variants are located in the ATPase-AAA-4 domain while rare variants outside this domain seem not to result in an IPD\textsuperscript{54}.

There are publicly accessible databases like ClinVar\textsuperscript{94} and DECIPHER\textsuperscript{95}, Exome variant server\textsuperscript{96} and access-for-a-fee databases like Human Gene Mutation Database (HGMD)\textsuperscript{97} that record disease-associated variants. HGMD maintains a catalogue of high-penetrance variants derived from the literature\textsuperscript{97}. None of these databases are 100% accurate: For example, 539 rare variants denoted disease-causing in HGMD were
observed in the 1000 Genomes Project at a frequency above 1\%\textsuperscript{98} and 140 variants labelled as causing BPDs showed a frequency in ExAC above 0.1\%, yet the evidence supporting a claim to pathogenicity was deemed insufficient for all but four variants\textsuperscript{82}.

In conclusion, large-scale population sequencing projects have greatly enhanced our ability to interpret the pathogenicity of potential candidate IPD-causing variants, but there are pitfalls: any rare variant must be interpreted in the context of the ethnicity of the individual and great care must be taken not to over-interpret the pathogenicity of novel variants absent from control datasets, even in established IPD genes, without back-up from functional studies or observation of the same variant in several unrelated cases with similar phenotypes.

The Future of IPD Gene Discovery

HTS will undoubtedly assist the discovery of variants causing IPDs in many new genes over the next decade but genomic sequencing alone cannot explain the mechanisms underlying the relationship between genotype and phenotype in cases with an apparent inherited disorder\textsuperscript{99}. In order to ascertain the functional consequences of rare variants, it is essential that knowledge of phenotypes and pathways is integrated systematically within a frame of reference similar to that of the human genome\textsuperscript{100}.

Data from a wide variety of sources, including GWAS, Online Medelian Inheritance in Man (OMIM) and mouse genome databases can be used to annotate candidate regions, better understand individual proteins and their roles in pathways relevant to megakaryopoiesis and platelet formation. Identification of the regulatory DNA elements has become feasible thanks to the results from projects like ENCODE\textsuperscript{101}, Roadmap\textsuperscript{102} and Blueprint\textsuperscript{103}, coordinated by the International Human Epigenome Consortium (IHEC). Integration of different layers of information, including methylation and histone modification states, expression quantitative trait loci data, transcriptomics and proteomics, requires the development of new statistical methods. Some insights on how this richness of information can aid the discovery of novel causes of disease will be discussed in the following sections.

*Human Phenotype Ontology*
To maintain power of gene discovery, cases sharing similar clinical and laboratory phenotypes need to be identified and new methods have been sought to cluster similar cases in the ever-growing cohorts undergoing genome sequencing. One of the widely used phenotype annotation standard for rare diseases is the HPO terms system\(^{104}\). HPO coding is used by the Deciphering of Developmental Disorders\(^ {105}\) and 100,000 Genomes Projects and so far 1,247 probands with BPDs have been HPO coded\(^ {2}\). This revealed the presence of non-hematological pathologies in 60% of cases, particularly in the central nervous (e.g. autism spectrum disorder), skeletal (e.g. osteoporosis) and immune systems\(^ {2}\). This insight into the more complex spectrum of pathologies in IPD cases is important for the provision of care, which often warrants a multi-disciplinary approach. Additionally, standardised phenotyping by means of HPO terms is critical for IPD gene discovery across large collections of cases. Indeed, genome sequencing combined with HPO coding supported the identification of the \(\text{DIAPH1}\) variant in two unrelated pedigrees with similar phenotype terms ‘Thrombocytopenia’ and ‘Deafness’\(^ {22}\). It also allowed integration with existing phenotype databases, such as the one for mouse phenotype ontology (MPO) terms, which aided the discovery of the \(\text{SRC}\) variant because cases and knockout mice shared HPO terms\(^ {48}\). These discoveries are critically dependent on new statistical methods which exploit the power of HPO-based patient coding together with genotypes obtained by sequencing\(^ {84,106,107}\).

In order to discover which genes are pertinent to the remaining IPDs, screening of large case collections will be essential. The small number of reported independent cases in the majority of IPDs and the lack of discovery in SPD to date indicates that extremely large collections are needed to bring together adequate numbers of unrelated index cases with a shared genetic basis. International collaboration is therefore warranted and will also bring together expertise about the clinical evolution of disease for each of the IPD genes. This collaborative approach will provide the platform to evaluate existing and new interventions to gather evidence for the best approach to treatment. That such international approaches can be successful has been demonstrated by the BRIDGE-BPD and ThromboGenomics consortium efforts reported in this issue\(^ {22,82}\) and in other journals\(^ {48,84}\).
The Non-coding Regulatory Space

The genome comprises 3.2 billion base pairs, 98% of which do not form part of a recognised gene. The non-exonic portion of the genome is largely regulatory in nature but the landscape of regulatory elements differs between cell types. The aforementioned IHEC initiatives have generated accurate maps of regulatory elements. The discoveries that specific heterozygous GOF variants in the 5’UTR of **ANKRD26** cause thrombocytopenia\(^\text{108}\) and that the vast majority of TAR syndrome cases of Northern European ancestry are due to compound inheritance of one **RBMT8A** null allele and a low-frequency SNV in its 5’UTR\(^\text{16}\) indicate that cell-specific knowledge of the regulatory space may help to identify novel mechanisms of disease in future. It is likely that similar non-coding regulatory variants in known or novel IPD genes remain to be discovered and integrating IHEC reference epigenome maps with the catalogue of blood cell-type-specific transcript isoform usage can help interpret the consequence of non-coding DNA variants observed in IPD cases\(^\text{108}\). The exploration of the non-coding portion of the genome for variants causal of IPDs requires many parallel approaches recognising the complexity of the regulatory networks at play. For example the cell-specific role of non-coding RNAs has been recently highlighted\(^\text{109}\). Finally the rapidly accumulating ChIP-seq data defining the binding sites for transcription factors in different cell types, including data generated in megakaryocytes will also be of value\(^\text{110}\).

GWAS for Platelet Traits

Meta-analysis of GWAS for blood cell traits has led to the discovery of nearly 200 common SNVs exerting small effects on blood cell indices\(^\text{111}\). These variants mark known and new genomic regions important for hematopoiesis and blood cell survival. With GWAS being performed on ever-larger population samples such as the UK Biobank and Million Veteran cohorts, the number of associated SNVs is predicted to rise substantially, with increasing numbers of rare variants with larger effect sizes being revealed. A picture is therefore emerging of hundreds of genes identified by GWAS controlling the life cycle of platelets and the maximum effect sizes of variants in these genes on platelet traits are inversely correlated with their MAFs due to selection. This assumption is illustrated by recent observations in **TUBB1**. The common non-
coding SNV rs4812048 was linked with an effect on platelet volume by GWAS\textsuperscript{1} and rare non-synonymous SNVs have since been reported as a cause of macrothrombocytopenia in man\textsuperscript{21,112-114}. This suggests that integrating knowledge about GWAS loci affecting platelet indices alongside genome sequencing data of collections of patients with IPDs of unknown molecular etiology may reveal novel candidate genes.

\textit{Protein-protein Networks}

Knowledge about the molecules and pathways underlying megakaryopoiesis and the formation of platelets is rapidly expanding. Bringing knowledge gathered from GWAS\textsuperscript{111} and inherited BPDs (the 76 BPD genes reported by Simeoni \textit{et al}, this issue\textsuperscript{82}) together with information about genes identified by ChIP-seq\textsuperscript{110} we defined 200 unique genes deemed important for these processes (Fig. 4B). The proteins encoded by these “seed genes” were used as baits to retrieve their first-order interactors from the Reactome and IntAct databases using previously reported informatics approaches\textsuperscript{1,115}. This information was displayed as a protein-protein interaction network (PPIN) consisting of 1,684 nodes (proteins) connected by 5,360 biochemical interactions (edges) using the Cytoscape application (Fig. 4A). To make this PPIN, which encompasses the knowledge from thousands of publications available to the scientific and medical communities in a format allowing its unrestricted use, we have made a Cytoscape file available for download (Supplemental data). For example, it can be used to resolve IPDs with digenic roots as illustrated by observations in \textit{TBXA2R}: four reports describe LOF variants of \textit{TBXA2R} in patients with platelet defects as a dominant finding. Although heterozygosity for the \textit{TBXA2R} variants correlated with the platelet defect in these carriers, there was no association with bleeding problems. It was noted that a potential second genetic factor would be required to cause bleeding. Two sub-networks (Fig. 4C, D) of proteins important for the signalling events downstream of the thromboxane receptor (Fig. 4C) and for the synthesis of thromboxane (Fig. 4D) may highlight potential candidates for this second gene\textsuperscript{78,116-118}. Indeed we observed a case with severe bleeding and a LOF variant in \textit{TBXA2R} (Fig. 4C) and a second putative causal variant was identified in \textit{PTGS1}, which encodes for Cyclooxygenase-1\textsuperscript{119}. 
Conclusion

HTS has enabled the discovery of many novel IPD genes in the past five years and this new knowledge can be rapidly integrated in diagnostic platforms like the ThromboGenomics HTS test, which will simplify and hasten diagnosis of IPDs. There is however still an immense challenge to resolve the genetic basis of the remaining IPDs and to gather better evidence for the best treatments. This can only happen through international collaboration and knowledge sharing. We must seek permission from patients and their families to share their genotype and phenotype data and invest in standardized phenotyping using internationally agreed terms like those of the HPO system. There is also an obligation for the research and clinical communities to pursue the development of informatics environments for the safe sharing of anonymized and linked-anonymized data. WGS combined with emerging data from GWAS, ChIP-seq, proteomics and mouse knock-out studies amongst others will also help explore the non-coding regulatory space and identify novel candidate IPD genes and variants.

Finally, as there are many potential pitfalls when interpreting the role of novel rare variants, it is important to apply rigorous standards when assigning pathogenicity. Providing a molecular diagnosis to patients is highly desirable but making incorrect assumptions about variants could be harmful.
Acknowledgments

We would like to thank the members of the BRIDGE-BPD and ThromboGenomics Consortia for their contributions. The BRIDGE-BPD and ThromboGenomics studies, including the enrolment of cases, the sequencing and analysis received support from the NIHR BioResource – Rare Diseases. The NIHR BioResource is funded by the National Institute for Health Research (NIHR). We gratefully acknowledge the patients and their relatives who participated in the NIHR BioResource – Rare Diseases studies without whom much of this research would not be possible. The contributions by Jo Westmoreland (Medical Research Council {MRC} Laboratory of Molecular Biology) and by Dr Myrto Kostadima and Stuart Meacham (University of Cambridge, Ouwehand Group) in creating Fig. 1, Fig. 2 and Fig. 4 respectively are acknowledged.

C.L. is the recipient of a Clinical Research Training Fellowship award from the MRC and M.A.L. and C.L. are also supported by the Imperial College London NIHR Biomedical Research Centre. E.T. is supported by the NIHR BioResource and research in the Ouwehand laboratory receives support from the British Heart Foundation, European Commission, MRC, NHS Blood and Transplant, NIHR and Wellcome Trust.

Authorship

Contribution: C.L. wrote the paper; E.T. performed data analysis and edited the manuscript; M.A.L. and W.H.O. are Co-chairs of the BRIDGE-BPD consortium; K.F. and W.H.O. are Co-chair and Chair of the Scientific and Standardization Committee (SSC) on “Genomics in Thrombosis and Haemostasis” of the International Society on Thrombosis and Haemostasis. This SSC oversaw the development of the ThromboGenomics HTS test referred to in this review and described by Simeoni et al in this issue of Blood. K.F., M.A.L. and W.H.O. edited the manuscript.

Conflict-of-interest disclosure

The authors declare no competing financial interests.
References


83. The 100,000 Genomes Project. http://www.genomicsengland.co.uk/.
97. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and


Figure Legends

Figure 1: The 51 genes underlying Inherited Platelet Disorders.

The cartoon depicts the process of megakaryopoiesis and platelet formation. Each of the 51 known inherited platelet disorder (IPD) genes are indicated and categorised according to their effect on megakaryocyte and platelet biology. IPDs typically associated with phenotypes outside of the blood system are indicated by *; HSC, hematopoietic stem cell.

Figure 2: Expression levels of 51 genes underlying Inherited Platelet Disorders across hematopoietic stem and progenitor cells.

High relative expression is shown in red and low relative expression in blue. The expression of each gene is normalized compared to a mean expression of zero across all the samples. Genes are ordered and colour-coded according to their predicted effect on platelet biology (as in Fig. 1). Information about the levels of transcripts for the 51 genes determined by RNA-seq was retrieved from Chen et al. HSC, hematopoietic stem cell; MPP, multipotent progenitor; CLP, common lymphoid progenitor; CMP, common myeloid progenitor; GMP, granulocyte-monocyte progenitor; MEP, megakaryocyte-erythrocyte precursor; EB, erythroblast; MK, megakaryocyte.

Figure 3: Genomic location of the 51 genes underlying Inherited Platelet Disorders.

Circos diagram illustrating the location of known IPD genes across human chromosomes. Track 1: Cytoband with chromosome name with centromeres in blue. Track 2: Genomic location of 51 established inherited platelet disorder (IPD) genes and the year in which variants in the gene were first identified as a cause of IPD in humans in brackets. Gene names in red represent genes identified by high throughput sequencing. Track 3: Log_{10} of the number of amino acids encoded by the reference CCDS transcript. Log_{10} scale is indicated at 12 o’clock. Track 4: Log_{10} of the number of rare variants predicted to affect amino acid sequence observed in 6,390 individuals enrolled to the NIHR BioResource – Rare Diseases; Log_{10} scale is indicated at 12 o’clock.

Figure 4: Protein-protein interaction network reflecting the molecules and pathways implicated in megakaryopoiesis, the formation of platelets, thrombosis and hemostasis.

A. Protein-protein interaction network (PPIN) of 1,684 nodes (proteins) connected by 5,360 edges (biochemical reactions). The 1,517 first-order interacting nodes and all but 24 of the 5,360 edges were obtained from the Reactome (n=3,625) and IntAct (n=1,711) databases. The 24 edges were added on basis of manual literature curation. The 200 baits are colored as per the Venn diagram (B) except for the eight baits present in more than one category, which are pink and the eight prototype proteins involved in the synthesis of thromboxone and signalling via the thromboxane-thromboxane receptor (Tbxa2r) pathway. The Venn diagram shows the three gene sets in ochre, blue and purple for the ThromboGenomics HTS test platform gene set, the platelet volume and count GWAS gene set or the gene set identified by ChiP-seq in human megakaryocytes and showing binding of all five
transcription factors (Fli1, Gata1, Gata2, Runx1, Tal1) at their promoter\textsuperscript{110}, respectively. \textbf{C and D}. Sub-networks retrieved from the PPIN in \textbf{(A)}. \textbf{C}. A sub-network of 156 nodes and 874 edges obtained by retrieving the first order interactors of Tbxa2r, the receptor for thromboxane. \textbf{D}. A sub-network of 26 nodes and 42 edges involved in the synthesis of thromboxane and obtained by selecting the first order interactors of Tbxas1 (Thromboxane synthase 1) and Pla2g4a (Phospholipase A2). The red nodes in \textbf{C} and \textbf{D} are a set of prototype proteins related with thromboxane synthesis and signalling and the other colored nodes are baits. The surface area of the red colored nodes in \textbf{A} and all colored nodes in \textbf{C} and \textbf{D} equate with their transcript level determined by sequencing of RNA from human megakaryocytes (data retrieved from Chen \textit{et al})\textsuperscript{103}. An interactive version of the network, containing gene expression levels and other annotation features is available for download in Cytoscape format from Supplemental data.
Fig. 1

Early Megakaryopoiesis

<table>
<thead>
<tr>
<th>HSC</th>
<th>Early MK</th>
</tr>
</thead>
<tbody>
<tr>
<td>TPO/MPL signalling</td>
<td>Transcription regulation</td>
</tr>
<tr>
<td>THPO/MPL</td>
<td>GATA1</td>
</tr>
<tr>
<td></td>
<td>RUNX1</td>
</tr>
<tr>
<td></td>
<td>FLt1*</td>
</tr>
<tr>
<td></td>
<td>ETV6</td>
</tr>
<tr>
<td></td>
<td>GF11B</td>
</tr>
<tr>
<td></td>
<td>HoxA11*</td>
</tr>
<tr>
<td></td>
<td>MECom*</td>
</tr>
<tr>
<td></td>
<td>ANKR26</td>
</tr>
<tr>
<td></td>
<td>RBM8A*</td>
</tr>
</tbody>
</table>

Late Megakaryopoiesis & Pro-platelet formation

<table>
<thead>
<tr>
<th>Late MK</th>
<th>Pro-platelet formation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Granule biogenesis &amp; trafficking</td>
<td>Cytoskeleton regulation</td>
</tr>
<tr>
<td></td>
<td>MYH9*</td>
</tr>
<tr>
<td></td>
<td>WAS*</td>
</tr>
<tr>
<td></td>
<td>ACTN1</td>
</tr>
<tr>
<td></td>
<td>F-LNA*</td>
</tr>
<tr>
<td></td>
<td>TUBB1</td>
</tr>
<tr>
<td></td>
<td>DIAPH1*</td>
</tr>
</tbody>
</table>

Pathway Incompletely Defined

| CYCS |
| SRC2 |
| SLFN14 |
| PLAU |
| STIM1* |

Platelet function

<table>
<thead>
<tr>
<th>GPCR signalling</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2RY13</td>
</tr>
<tr>
<td>TBXA2R</td>
</tr>
<tr>
<td>TBXAS1*</td>
</tr>
<tr>
<td>PLAKGAP</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GP receptor signalling</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITGA2B</td>
</tr>
<tr>
<td>RASGRF2</td>
</tr>
<tr>
<td>ITGB3</td>
</tr>
<tr>
<td>VWF</td>
</tr>
<tr>
<td>GP1BA</td>
</tr>
<tr>
<td>FERMT3</td>
</tr>
<tr>
<td>GP1IR</td>
</tr>
<tr>
<td>GPA</td>
</tr>
<tr>
<td>GP9</td>
</tr>
</tbody>
</table>

* Presence of phenotypes outside the blood system
Glossary of Terms

**Inherited Platelet Disorder (IPD):** Any genetic abnormality of platelet number, morphology or function in humans

**IPD Gene:** A gene in which a DNA variant has been shown to cause an IPD in humans

**Bleeding, Thrombotic and Platelet Disorders (BPD):** Term used by the NIHR BioResource and the ThromboGenomics consortium to describe cases with genetic variants leading to a tendency to bleed abnormally, or a tendency to abnormal thrombosis or an abnormality of the count, volume, morphology or function of platelets.

**High-Throughput Sequencing (HTS):** Term encompassing a variety of modern techniques, which allow the sequencing of DNA and RNA more quickly and cheaply than traditional Sanger sequencing *(also referred to as next-generation sequencing)*.

**Minor Allele Frequency (MAF):** The frequency at which the least common allele at a particular variant position occurs in a population.

**Whole Exome Sequencing (WES):** The sequencing by HTS of the coding portion of the genome comprising ~65 million bases.

**Whole Genome Sequencing (WGS):** The sequencing by HTS of the entire genome comprising ~3.2 billion bases.

**Single Nucleotide Variant (SNV):** A single base pair alteration in a DNA sequence compared to the reference genome for a particular species. A SNV does not necessarily alter the function of a gene. SNVs can be common (MAF>5%), low frequency (MAF between 1% and 5%) or rare (MAF<1%).

**Loss of Function (LOF):** A variant allele causing the complete or partial absence of normal gene function compared to the wild-type allele.

**Gain of Function (GOF):** A variant allele causing a gain of normal gene function compared to the wild-type allele.

**5-prime untranslated region (5’UTR):** The section of the message RNA (mRNA) upstream from the start codon (where translation of mRNA starts). The 5’UTR is important in regulating translation.

**ChIP-seq (Chromatin immunoprecipitation combined with HTS):** A powerful method for identifying modification of DNA-associated proteins like histones or used for mapping the binding sites of proteins like transcription factors; the method is used extensively to enhance the functional annotation of the non-coding portion of the genome.
**Genome Wide Association Study (GWAS):** A study comparing the frequencies of millions of SNVs across thousands of people with and without a particular disease or assessing the effect of these SNVs on a quantitative trait. This type of study explores the genome in a hypothesis-free approach for variants that confer a risk for a particular disease or are exerting an effect on the value of a certain trait.

**Protein-protein interaction network (PPIN):** A representation of physical contacts established between proteins. PPINs can encompass knowledge on biochemical pathways derived from thousands of experiments and may be represented graphically in a network diagram.
Inherited platelet disorders: towards DNA-based diagnosis
Claire Lentaigne, Kathleen Freson, Michael A. Laffan, Ernest Turro and Willem H. Ouwehand