Genomics England Clinical Interpretation Partnership (GeCIP) Detailed Research Plan Form

Application Summary

GeCIP domain name	**Non-malignant haematological and haemostasis disorders (HHD)**
Project title *(max 150 characters)*	**Towards a comprehensive genetic architecture for heritable blood stem cell, myeloid and haemostasis disorders**

Objectives. *Set out the key objectives of your research. (max 200 words)*
The overarching aim of the proposed programme is to elucidate molecular mechanism in novel inherited disorders of haematopoiesis and haemostasis. This work will address a series of interrelated scientific, clinical and strategic objectives:

Scientific objectives:
1. Enrich understanding of phenotypic diversity and natural history of rare disorders of the haemopoietic and haemostasis systems associated with known genes
2. Discover causative variants in novel disease genes & their pathogenetic mechanism

Clinical objectives:
1. Enable early diagnosis and genetic counselling
2. Inform therapy for affected patients to achieve better health outcomes
3. Avoid incorrect treatments
4. Facilitate the gathering of information regarding natural history of individual disorders

Strategic objectives:
1. Extend knowledge and understanding of the genetic and molecular basis of the above disorders and implications for health and disease
2. Harmonise, coordinate and build capacity for genomic diagnostics and research in these disorders

Lay summary. *Information from this summary may be displayed on a public facing website. Provide a brief lay summary of your planned research. (max 200 words)*

The blood is one of the most complicated parts of the human body. We rely on red blood cells to carry oxygen round the body. An array of specialized white blood cells together with tiny cell fragments called platelets protect us from bleeding by clumping together with special clotting proteins when bleeding occurs and induce repair if the vessel wall is damaged. To produce this complex system, each cell relies on detailed instructions in the form of DNA. Spelling mistakes in DNA can cause disease such as bleeding, thrombus formation or anaemia, even in children. To treat patients better and advise families we need to uncover the underlying genetic spelling mistakes. This used to be an almost impossible task, but with new techniques we can read the entire DNA code of individual patients. This is called their genome; every person’s genome is unique. Researchers and doctors are still learning how to make sense of all the information it contains and especially how to identify which genetic changes cause disease. In this proposal, researchers will focus on studying the genomes of patients with blood, bleeding and clotting problems. We hope to learn more about which genes are important for blood cells and the clotting system and how they work. Together with patients we will work to bring better DNA tests to the NHS so that families can be helped quicker. Eventually new treatments may become available but this generally takes a very long time.

Technical summary. *Information from this summary may be displayed on a public facing website. Please include plans for methodology, including experimental design and expected outputs of the research. (max 500 words)*

Inherited disorders of the haemopoietic and haemostasis systems represent a rich resource for hypothesis-generating research that improves human health. By comparison with genetically engineered mouse models, a forward genetic approach offers inherent scientific advantages such as (i) increased relevance of findings to human health in the natural environment, (ii) a definite and clinically important phenotype, (iii) freedom from preconceptions as to the genetic basis and disease mechanism. Furthermore, **both the severity of associated disease and the potential for curative precision medicine mandate a molecular approach to diagnosis for individual patients**. The 100,000 Genomes Project provides the opportunity to address these imperatives by interrogating variation across the entire genome of individuals with otherwise unexplained haemopoietic and haemostasis disorders.

In our capacity as Clinical Interpretation Partners, **our overriding aim is to deliver the benefits of molecular diagnosis to participating patients, their close relatives and the NHS**. Allied to this, our research objectives are as follows:

1. Enriched understanding of phenotypic diversity and natural history of disorders associated with known disease genes. We will perform detailed clinical and laboratory assessments of patients with variants in shared disease genes, including the outcome of any therapeutic interventions. By integrating this new information with prior knowledge (incl. the UK Haemophilia Centre Directors (UKHCDO) registry of patients with bleeding, thrombotic and platelet (BPD) disorders, GOSH registry of severe congenital neutropenia, King’s College and Bart’s and the London Hospitals registries of bone marrow failure, Oxford UK registry of congenital anaemias, we will enhance understanding of genotype-phenotype correlation and discover allelic disorders.

2. Discovery of causative variants in novel disease genes & their pathogenetic mechanism. Using the novel statistical Bayesian method, named BeviMed we will interrogate variants passing pre-defined filters, we will identify novel disease loci in patients clustered by disease phenotype (HPO terms) or familial relationship. We will perform studies of intermediate phenotypes of monocytes, neutrophils, CD4+ T cells and platelets at the epigenomic, mRNA, protein, and metabolomics levels to gain insights in the underlying molecular mechanisms between causal sequence variants and pathobiology and to generate hypotheses for further research, including experimental medicine studies.

In order to carry out this research, we have assembled a large team of clinicians, clinical scientists, wet lab, computational biologists and experts in statistical genomics, drawn from both academia, the EMBL European Bioinformatics Institute, the MRC Biostatistics Unit, the Wellcome Trust Sanger Institute, the Structural Genomics Consortium and academia. We will function as an extended research network and continue to secure collaborative funding to expand our already existing shared analytic and administrative infrastructure, which includes core bioinformatics, statistical method development, machine learning (with Alan Turing Institute), and a programme of regular virtual and face-to-face meetings. We will make use of the NIHR BioResource for Translational Medicine to coordinate recall for studies on mechanism of disease and for ‘first in man’ experimental medicine studies.

Expected start date	**September 2017**
Expected end date	**April 2022**

Lead Applicant 1

Name	Willem H Ouwehand
Post	Professor of Experimental Haematology Director NIHR BioResource – Rare Diseases
Department	Department of Haematology / Department of Human Genetics
Institution	University of Cambridge and Wellcome Trust Sanger Institute, Cambridge
Current commercial links	None

Lead Applicant 2

Name	Judith CW Marsh

Post	Professor of Clinical Haematology & Consultant Haematologist
Department	Dept of Haematological Medicine
Institution	King's College Hospital/King's College, London
Current commercial links	None

Administrative Support	
Name	Sofia Papadia
Email	sp605@medschl.cam.ac.uk
Telephone	01223 58 8727

Subdomain leads		
Name	**Subdomain**	**Institution**
Prof I Dokal	BMFS	Barts and The London
Prof P Ancliff	Severe congenital neutropenias	Great Ormond St Hospital
Prof M Layton	Congenital anaemias, iron disorders	Imperial College Hospital, London
Prof M Laffan	Haemostasis disorders	Hammersmith Hospital
Prof A Mumford	Platelet disorders	Bristol University Hospitals Genomics Medicine Centre Director
Prof C H Toh	Bleeding disorders	Liverpool University Hospitals

Detailed research plan

Full proposal (total max 1500 words per subdomain)

Title *(max 150 characters)*	**Towards a comprehensive genetic architecture for heritable stem & myeloid blood cell and haemostasis disorders**

Importance. *Explain the need for research in this area, and the rationale for the research planned. Give sufficient details of other past and current research to show that the aims are scientifically justified. Please refer to the 100,000 Genomes Project acceptable use(s) that apply to the proposal (page 6).*

Inherited diseases of the blood and haemostasis systems, including bone marrow failure syndromes (BMFS), stem cell and myeloid disorders (SMD), anaemia's and erythroid disorders (AED), bleeding, thrombotic and platelet disorders (BPD), cause substantial morbidity and mortality, demanding significant healthcare resources. Almost 25,000 patients with inherited BPDs are on the UKHCOD registry and the other categories cover another ~15,000 patients (incl. the haemoglobinopathies). HPO coding of the first 2,500 probands with BPD, AED and SMD showed the overlapping nature of such disorders with symptoms straddling the sub-specialities of haematology, with a predisposition to both haematological and non-haematological malignancies and pathologies of other organ systems, such as kidney, skeleton and central nervous system being present in 50% of cases.

Some of these disorders are curable by haemopoietic stem cell transplantation (HSCT) and a few can be successfully treated by gene therapy. Many other patients endure chronic ill-health despite supportive therapy such as coagulation factor replacement, blood and platelet transfusions, cytokines, antimicrobials, and nutritional support. Familial occurrence of these diseases, as well as their enrichment in consanguineous families, indicate that many are monogenic disorders, albeit a considerable fraction is caused by *de novo* variants. Disease-causing variants have already been identified in ~300 genes, leading to major advances in scientific understanding as well as manifest clinical benefits. A confirmed molecular diagnosis increases the confidence with which both natural history and treatment response can be predicted, enabling timely and tailored therapy with improved outcomes. Genetic identification of previously uncharacterised or unrecognised diseases is crucial to avoid inappropriate therapies in such patients (e.g. treatment for Immune Thrombocytopenia for patients with inherited BPDs). With improved mechanistic understanding there may also be opportunities for targeted pharmacological interventions such as small molecule inhibitors or biologic therapies (collaboration with Structural Genome Consortium). For the affected family, a molecular diagnosis also brings with it the possibilities of early (even pre-symptomatic or prenatal) diagnosis, pre-emptive therapy and genetic counselling.

GeCIP members have led research resulting in the discovery of >20 new disease genes through exome or genome sequencing studies, leading to important advances in understanding of disease mechanisms, clinical prognosis and therapy. Many discoveries have been translated into more affordable and rapid molecular diagnosis. For example, the discovery of *NBEAL2* and *RBM8A* in Gray Platelet and Thrombocytopenia and Absent Radii Syndromes, has provided the NHS with rapid diagnostic tests to manage probands and their close relatives. The majority of currently known disease-causing variants are located in the coding space but first examples of variants in the non-coding space have been identified by us (*RBM8A, PTGS1, GATA1-HDAC6)* and others. There remains significant potential to boost diagnostic yield and scientific discovery by continuing to analyse the non-coding space and the BeviMed method is scalable, fast and compute-time friendly (the only alternative method SKAT is not scalable). Furthermore, our recent completion of the production of blood cell reference epigenomes in 200 healthy controls (as part of the BLUEPINRT project) provides a firm foundation to define 'atypical epigenome profiles' in patients with non-malignant rare diseases. Joined-up analysis accros the sub-domains will therefore achieve:

- improved understanding of the phenotypic diversity of known and newly discovered inherited disorders, including response to therapy
- faster discrimination of pathogenic, likely pathogenic and variants of unknown clinical significance in known and new disease genes
- definition of the pathogenetic mechanism by which coding and non-coding variants cause disease

Research plans. *Give details of the analyses and experimental approaches, study designs and techniques that will be used and timelines for your analysis. Describe the major challenges of the research and the steps required to mitigate these.*

Phenotypes of interest:

We are interested in interrogating the full range of non-malignant haematological and the haemostasis disorders fulfilling eligibility criteria for the 100,000 Genomes Project. We have nearly completed the copying of the pheno- and geno-type data of the 2,500 probands with non-malignant haematological and haemostasis disorders who were enrolled in the NIHR BioResource. We have also reached agreement with the UK Genotyping and Phenotyping of Platelets (GAPP) study team to re-contact the 900 BPD probands enrolled in the GAPP study. These cases are extremely well phenotyped but the DNA samples of <150 have been analysed by whole exome sequencing (WES).

Aim 1: Enriched understanding of phenotypic diversity and natural history of disorders associated with known disease genes

Members of this GeCIP have extensive experience of compiling and publishing gene discovery studies, multi-centre case series and statistical method development (Albers *et al*, Nat Genetics 2011; Albers *et al*, Nat Genetics 2012; Cvejic *et al*, Nat Genetics 2013; Chen *et al*, Science 2014; Westbury *et al*, Genome Medicine 2015*;* Green *et al,* AJHG 2016; Stritt *et al*, Nature Communications 2016; Turro *et al,* Science Translational Medicine 2016; Stritt *et al*, Blood 2016; Simeoni *et al*, Blood 2016; Lentaigne *et al*, Blood 2016; Poggi *et al,* Haematologica 2016; Bariana *et al*, Brit J Haem 2017; Sivapalaratnam *et al,* Blood 2017; Pleines *et al*, J Clin Invest 2017; Greene *et al*, AJHG (accepted); Westbury *et al,* Blood (accepted). The analysis of the NIHR BioResource WGS results has already identified unanticipated defects in known genes (Sivapalaratnam *et al,* Blood 2017). This will enable improved understanding of the phenotypic spectrum and natural history of disorders. Sometimes this may amount to a demonstration that entirely different diseases can be associated with alternative effects on the same gene. We will continue to capture the clinical and laboratory information using the power of Human Phenotype Ontology (HPO) system (HPO data for 1000's of cases has already been captured during the pilot phase). We build on established capacity to perform extended refined laboratory investigations for particular areas of interest.

Aim 2: Discovery of causal coding variants in novel disease genes & their pathogenetic mechanism

Experience to date suggests that the genetic architecture underlying as yet unresolved disorders is dispersed over a large number of new loci, many of which will encode proteins that are highly connected in protein-protein interaction networks that control immunity, haematopoiesis and haemostasis. As well as classical monogenic disorders we have already encountered more complex inheritance patterns such as two rare alleles encoding proteins in the same pathway which create bi-genic pathway insufficiencies*,* combinations of extremely rare CNVs and relatively common SNVs and epigenetic mechanisms such as imprinting.

We have developed a new statistical method for Bayesian evaluation of rare variants in Mendelian disease (BeviMed) to interrogate the coding and non-coding portion of the genome. Application of the method across the BPD cases has already yielded results in uncovering causal variants localised in promoter-connected regulatory elements of known (*GATA1*) and new BPD genes (*HDAC6, PTGS1*). High resolution reference epigenome maps of 34 different types of blood and immune cells have been generated by GeCIP members as part of the BLUEPRINT epigenome project, defining the functional nature of nucleosome-depleted regulatory elements. Long range interaction between promoters and these regulatory elements have been experimentally defined in 17 of these cell types so far thereby immensely enhancing our ability to correctly link regulatory elements to the networks of genes they control. We will use the new Bayesian BeviMed method to cluster patients with composite phenotypes, exploiting the ontological relationship between HPO terms, and model the association between phenotype and genotype thereby generating estimates of the probability of variants being causal. The cut off of probability estimates will be set sensitively to select probands and pedigrees for follow-up studies. We will perform co-segregation studies and apply the extensive repertoire of cellular and biochemical methods already operational in the laboratories of the GeCIP participants to obtain corroborative evidence of gene and variant causality.

Subsequent assays will be designed to explore the effects of the variant(s) on the epigenome, mRNA and protein landscapes (collaboration with Prof Lamond, University of Dundee for proteomics), complemented by immunoblotting +/- immunofluorescence microscopy/cytometry and high resolution microscopy, respectively. We have, in partnership with the HipSci consortium already generated induced pluripotent stem cells (iPSC) cells from 30 patients and also successfully applied CRISPR genome editing to knock-in causal variants in newly discovered genes in reference IPSC lines and/or the variant(s) will be modelled in cultured or primary cells by standard techniques. The structural effect of amino acid substitutions will be modelled into known molecular structures. Assays for forward programming of IPSCs into megakaryocytes, erythroblasts and neutrophils are operational in the laboratories of several of the GeCIP members.

Alternative models systems (mouse knock-ins/knock-outs, zebrafish) are also available within the GeCIP as well as by collaboration and may have an important part to play, particularly for those genes not previously linked to haemopoiesis or haemostasis.

Collaborations including with other GeCIPs. *Outline your major planned academic, healthcare, patient and industrial collaborations. This should include collaborations and data sharing with other GeCIPs. Please attach letters of support.*

Members of our domain are active in a variety of funded research programmes, the aims of which mirror the 100,000 Genomes Project, as well as several overlapping GeCIP domains including myeloid malignancies (Schuh, Oxford), immunology (Hambleton [Newcastle], Smith [Cambridge], Thrasher [GOSH/UCL]). These individuals embody the collaborative relationship between projects and include representation from national initiatives on rare disease research as they relate to blood and haemostasis disorders. External collaborators who have specifically expressed interest in working cooperatively with the GeCIP domain include Kuijpers (University of Amsterdam), Freson (University of Leuven), Furie (Boston) and Poncz (Philadelphia). Through our overlapping membership we reach out to the wider clinical community of haemostasis experts (UKHCDO) and non-malignant haematologists, including paediatricians. Ri Liesner at GOSH is Chair of UKHCDO and has been extremely supportive in enhancing enrolment through the Haemophilia centres.

Training. *Describe the planned involvement of trainees in the research and any specific training that will form part of your plan.*

Members of this GeCIP are already actively engaged in delivering the Masters of Genomic Medicine at their local Genomic Medicine Centres (GMCs) and a Clinical Sciences MPhil in Rare Diseases at Cambridge University. Furthermore we are raising awareness of genomic medicine and the 100,000 Genomes Project within our clinical communities, through involvement with local GMCs and by participation in clinical and scientific meetings of national and international ‘learned’ societies.

There is widespread recognition across the GeCIP of the opportunities provided by the 100,000 Genomes Project to engender an interdisciplinary approach to translational research involving clinicians, bench researchers, bioinformaticians, statisticians and experts in machine learning. We already have trained a new cadre of researchers and NHS employees with skills and understanding across traditional subject boundaries. Activities associated with this GeCIP will provide ample opportunity to host research projects of varying complexity and length. We will already deliver a training and research environment for five MRC Clinical Research fellows and eight PhD fellows in statistics, bioinformatics and machine learning, supported by funding schemes of the main UK research funders such as MRC, NIHR, the Wellcome Trust and the Medical Charities. Movement of clinical training fellows and trainee clinical scientists between centres is being encouraged and shown to be popular with trainees (whether research or NHS).

People and track record. *Explain why the group is well qualified to do this research, how the investigators would work together.*

This GeCIP includes a large number of members with established track records in clinical and bench-based research relevant for the subdomains. These individuals are drawn from coherent and collaborative research communities centred on highly specialised clinical practice. They include international leaders in their field, who run well-funded research groups linked to and supported by NIHR Biomedical Research Centres (BRCs). The quality of their research is outstanding with publications in high-impact journals, like Cell, JCI, Nature, Nature Genetics, Nature Medicine, New England Journal of Medicine and Science amongst others. Members within subdomains have often worked and published with each other before and will function as a collaborative unit. BMFS, congenital anaemias and neutropenias are similarly represented by Marsh, Ancliff, Dokal, Layton and Roberts and bleeding, thrombotic and platelet disorders by Laffan, Mumford, Freson, Ouwehand and Toh.

We consider it imperative for the success of the 100,000 Genomes Project that we cement the different subdomains together so we can achieve the highest impact on improving patient care and deliver impactful research. Aggregating phenotype and genotype data across the subdomain makes sound biological and statistical sense because of the shared nature of many of the pathways and disease phenotypes and the immense richness of the annotation of the non-coding space of blood, immune and liver cells thereby providing an exemplar of how to capitalise on the investment in WGS. To maximise this synergy we will adopt a shared approach to data analysis involving the iterative pooling of genomic resources, generated through diverse research programmes including the 100,000 Genomes Project itself. This will ensure that sequencing data can be queried against the richest possible functional annotation across the different types of blood and liver cells.

In order to deliver this platform, the GeCIP has a credible and talented membership representing the quantitative sciences with experts in bioinformatics (both clinical and genomics), computational biology and statistical genomics. A group of ~30 GeCIP members (2/3 quantitative and 1/3 clinical) with leading teams in Statistical Genomics and method development have already met to develop effective joined up analytic approaches. Several GeCIP members have shared appointments with the MRC Biostatistics Unit in Cambridge (director: Richardson) and with the Wellcome Trust Centre for Human Genetics in Oxford (director: Donnelly). Our GeCIP will foster the working relationship with these two centres as the development and application of new analytical methods will be critical for the successful delivery of the proposed programme of research and its translation into improved clinical care.

We will function as an extended research network with nodes and hubs. To benefit maximally from the potential research synergy, our intention will be to secure collaborative funding to support a shared analytic and administrative infrastructure, including dedicated bioinformatic support, communication hubs and a programme of regular virtual and face-to-face meetings.

Clinical interpretation. *(Where relevant to your GeCIP) Describe your plans to ensure patient benefit through clinical interpretation relevant to your domain. This should specifically address variant interpretation and feedback and your interaction with the cross-cutting Validation and Feedback domain.*

The validation and reporting of variants in **known disease genes** is a core activity of existing clinical diagnostic services, which are strongly represented within the GeCIP. We have established a collaboration with the ISTH Standardisation and Science Committee for Genomics in Thrombosis and Haemostasis and the American Society of Haematology Task Force for Precision Medicine. We are active opinion leaders in these two groups which aim to bring order in the ClinVar database.

For the time being (until cost of WGS is ~£250/sample) diagnostic next generation sequencing (NGS) platforms for panels of genes relevant for all the subdomains are available at CUH (~80 genes for BPD and 70 genes for BMF genes; (MDT for BPD chaired by Gomez, UCL and for BMF by Dokal, Barts) and OUH for RBC and SMD; MDT chaired by Schuh). The NGS panels are available as an accredited clinical diagnostic test with defined validation and reporting processes. Establishing these tests has yielded useful skills and insights into the bioinformatic and interpretational challenges of NGS in a clinical diagnostic setting, with relevance to reporting services provided by the GMCs. In particular this experience highlights the added value of (i) integrating corroboratory evidence at protein and/or functional level and (ii) strong representation from a range of expert clinicians, statistical experts and clinical geneticists, providing the correct forum for patient-centric interpretation of genetic data. We are committed to offering support to GMCs in reporting Tier 1 variants in known genes and in turn will seek their support in the confirmatory testing of pathogenic and likely pathogenic variants by Sanger sequencing.

Judging when novel variants (**coding or non-coding**) reach the status of actionable clinical findings is an ongoing challenge within the rare disease community. The GeCIP will adhere to international and national guidelines and follow an extremely conservative approach before genes and variants therein are moved to Tier 1 status. A recent analysis of the pathogenic variants for the BPD genes in the HGMD database has revealed the presence of large numbers of variants erroneously labelled as pathogenic. The GeCIP leads are aware of the importance of balancing the desire to deliver patient benefit by rapid reporting with the potential risks of incorrect interpretation of variants. The GeCIP and its subdomains provide a forum in which to consider evidence for disease causation as it is generated. In practice we expect to assemble a case for the designation of novel variants as disease-causing and present it to the *Validation and Feedback* domain for their approval prior to publication and reporting.

Beneficiaries. *How will the research benefit patients and healthcare institutions including the NHS, other researchers in the field? Are there other likely beneficiaries?*

We know from experience that accurate molecular diagnosis empowers patients, families and their healthcare teams to deliver best care. These benefits will attach not only to those enrolled in the project but also to patients with the same disorder globally, who will be more likely to achieve a molecular diagnosis as a result of our discoveries. As clinicians and clinical scientists already working with patients in the NHS, and fully engaged with international learned societies, members of this GeCIP are well-placed to ensure the incorporation of new knowledge into routine diagnostics. At a minimum, a molecular diagnosis enables screening of family members and genetic counselling, including the possibilities of prenatal and preimplantation genetic diagnosis where appropriate. Over time, clinicians learn more about the behaviour of individual disorders, which informs understanding of prognosis and hence the appropriateness of risky but potentially curative therapies such as stem cell transplantation – including knowing when not to transplant. Early ascertainment of the molecular basis of disease improves the accuracy of therapeutic intervention including the potential for precision medicine. Although often stereotyped as highly costly therapies, recent examples all represent repurposing of drugs already licensed for other indications, yet offering transformative improvement in patients' wellbeing. Furthermore, discoveries within our subdomain have the potential to inform scientific understanding of fundamental processes relevant to common disorders such as cardiovascular conditions (heart attack and stroke), neurological (15% of BPD cases have neurological conditions) and malignant conditions (a large portion of the non-malignant haematological conditions substantially increase the risk of malignancies).

Commercial exploitation. *(Where relevant to your GeCIP) Genomics England has a very explicit intellectual property policy. We and other funders need to know if the proposed research likely to generate commercially exploitable results. Do you have commercial partners in place?*

GeCIP members will be guided by GEL and their host institutions to optimise the 'return of investment' in the 100,000 Genomics Project for UK PLC, including through commercial partnership.

References. *Provide key references related to the research you set out.*

1. Sowerby, JM *et al.* Nbeal2 is required for neutrophil and NK cell function, and pathogen defence. JCI. 2017 (accepted for publication)
2. Greene, D *et al.* A fast integrative genetic association test for rare diseases. Am J Hum Genet. 2017 (accepted for publication)
3. Petersen, R *et al*. Platelet function is modified by common sequence variation in megakaryocyte super enhancers. Nat. Commun. 8, 16058 doi: 10.1038/ncomms16058. 2017 (Epub ahead of print)
4. Westbury, S *et al*. Expanded repertoire of RASGRP2 variants responsible for platelet dysfunction and severe bleeding. Blood. 2017 (accepted for publication)

5. Carss KJ, Arno G, Erwood M, Stephens J, Sanchis-Juan A, Hull S, Megy K, Grozeva D, Dewhurst E, Malka S, Plagnol V, Penkett C, Stirrups K, Rizzo R, Wright G, Josifova D, Bitner-Glindzicz M, Scott RH, Clement E, Allen L, Armstrong R, Brady AF, Carmichael J, Chitre M, Henderson RH, Hurst J, MacLaren RE, Murphy E, Paterson J, Rosser E, Thompson DA, Wakeling E, Ouwehand WH, Michaelides M, Moore AT, Webster AR, Raymond FL. Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease. Am J Hum Genet. 2017;100(1):75-90.
6. Guo BB, Allcock RJ, Mirzai B, Malherbe JA, Choudry FA, Frontini M, Chuah H, Liang J, Kavanagh SE, Howman R, Ouwehand WH, Fuller KA, Erber WN. Megakaryocytes in Myeloproliferative Neoplasms Have Unique Somatic Mutations. The American journal of pathology. 2017. Epub ahead of print.
7. Kilpinen H, Goncalves A, Leha A, Afzal V, Alasoo K, Ashford S, Bala S, Bensaddek D, Casale FP, Culley OJ, Danecek P, Faulconbridge A, Harrison PW, Kathuria A, McCarthy D, McCarthy SA, Meleckyte R, Memari Y, Moens N, Soares F, Mann A, Streeter I, Agu CA, Alderton A, Nelson R, Harper S, Patel M, White A, Patel SR, Clarke L, Halai R, Kirton CM, Kolb-Kokocinski A, Beales P, Birney E, Danovi D, Lamond AI, Ouwehand WH, Vallier L, Watt FM, Durbin R, Stegle O, Gaffney DJ. Common genetic variation drives molecular heterogeneity in human iPSCs. Nature. 2017. Epub ahead of print.
8. Kohler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Ayme S, Baynam G, Bello SM, Boerkoel CF, Boycott KM, Brudno M, Buske OJ, Chinnery PF, Cipriani V, Connell LE, Dawkins HJ, DeMare LE, Devereau AD, de Vries BB, Firth HV, Freson K, Greene D, Hamosh A, Helbig I, Hum C, Jahn JA, James R, Krause R, SJ FL, Lochmuller H, Lyon GJ, Ogishima S, Olry A, Ouwehand WH, Pontikos N, Rath A, Schaefer F, Scott RH, Segal M, Sergouniotis PI, Sever R, Smith CL, Straub V, Thompson R, Turner C, Turro E, Veltman MW, Vulliamy T, Yu J, von Ziegenweidt J, Zankl A, Zuchner S, Zemojtel T, Jacobsen JO, Groza T, Smedley D, Mungall CJ, Haendel M, Robinson PN. The Human Phenotype Ontology in 2017. Nucleic Acids Res. 2017;45(D1):D865-d76.
9. Pleines I, Woods J, Chappaz S, Kew V, Foad N, Ballester-Beltran J, Aurbach K, Lincetto C, Lane RM, Schevzov G, Alexander WS, Hilton DJ, Astle WJ, Downes K, Nurden P, Westbury SK, Mumford AD, Obaji SG, Collins PW, Delerue F, Ittner LM, Bryce NS, Holliday M, Lucas CA, Hardeman EC, Ouwehand WH, Gunning PW, Turro E, Tijssen MR, Kile BT. Mutations in tropomyosin 4 underlie a rare form of human macrothrombocytopenia. J Clin Invest. 2017;127(3):814-29.
10. Poggi M, Canault M, Favier M, Turro E, Saultier P, Ghalloussi D, Baccini V, Vidal L, Mezzapesa A, Chelghoum N, Mohand-Oumoussa B, Falaise C, Favier R, Ouwehand WH, Fiore M, Peiretti F, Morange PE, Saut N, Bernot D, Greinacher A, BioResource N, Nurden AT, Nurden P, Freson K, Tregouet DA, Raslova H, Alessi MC. Germline variants in ETV6 underlie reduced platelet

formation, platelet dysfunction and increased levels of circulating CD34+ progenitors. Haematologica. 2017;102(2):282-94.

11. Sivapalaratnam S, Westbury SK, Stephens JC, Greene D, Downes K, Kelly AM, Lentaigne C, Astle WJ, Huizinga EG, Nurden P, Papadia S, Peerlinck K, Penkett CJ, Perry DJ, Roughley C, Simeoni I, Stirrups K, Hart DP, Tait RC, Mumford AD, BioResource N, Laffan MA, Freson K, Ouwehand WH, Kunishima S, Turro E. Rare variants in GP1BB are responsible for autosomal dominant macrothrombocytopenia. Blood. 2017;129(4):520-4.
12. Zou S, Teixeira AM, Kostadima M, Astle WJ, Radhakrishnan A, Simon LM, Truman L, Fang JS, Hwa J, Zhang PX, van der Harst P, Bray PF, Ouwehand WH, Frontini M, Krause DS. SNP in human ARHGEF3 promoter is associated with DNase hypersensitivity, transcript level and platelet function, and Arhgef3 KO mice have increased mean platelet volume. PLoS One. 2017;12(5):e0178095.
13. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, Mead D, Bouman H, Riveros-Mckay F, Kostadima MA, Lambourne JJ, Sivapalaratnam S, Downes K, Kundu K, Bomba L, Berentsen K, Bradley JR, Daugherty LC, Delaneau O, Freson K, Garner SF, Grassi L, Guerrero J, Haimel M, Janssen-Megens EM, Kaan A, Kamat M, Kim B, Mandoli A, Marchini J, Martens JH, Meacham S, Megy K, O'Connell J, Petersen R, Sharifi N, Sheard SM, Staley JR, Tuna S, van der Ent M, Walter K, Wang SY, Wheeler E, Wilder SP, Iotchkova V, Moore C, Sambrook J, Stunnenberg HG, Di Angelantonio E, Kaptoge S, Kuijpers TW, Carrillo-de-Santa-Pau E, Juan D, Rico D, Valencia A, Chen L, Ge B, Vasquez L, Kwan T, Garrido-Martin D, Watt S, Yang Y, Guigo R, Beck S, Paul DS, Pastinen T, Bujold D, Bourque G, Frontini M, Danesh* J, Roberts* DJ, Ouwehand* WH, Butterworth* AS, Soranzo* N. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell. 2016;167(5):1415-29.e19.
14. Bermejo E, Alberto MF, Luceros AS, Nurden AT, Simeoni I, Ouwehand WH, Turro E, Nurden P. Functional and molecular characterization of an inherited abnormal platelet function related to a new CalDAG-GEFI protein variant in an argentinean family. Journal of Thrombosis and Haemostasis. 2016;14:116-7.
15. Breeze CE, Paul DS, van Dongen J, Butcher LM, Ambrose JC, Barrett JE, Lowe R, Rakyan VK, Iotchkova V, Frontini M, Downes K, Ouwehand WH, Laperle J, Jacques PE, Bourque G, Bergmann AK, Siebert R, Vellenga E, Saeed S, Matarese F, Martens JH, Stunnenberg HG, Teschendorff AE, Herrero J, Birney E, Dunham I, Beck S. eFORGE: A Tool for Identifying Cell Type-Specific Signal in Epigenomic Data. Cell Rep. 2016;17(8):2137-50.
16. Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martin D, Watt S, Yan Y, Kundu K, Ecker S, Datta A, Richardson D, Burden F, Mead D, Mann AL, Fernandez JM, Rowlston S, Wilder SP, Farrow S, Shao X, Lambourne JJ, Redensek A, Albers CA, Amstislavskiy V, Ashford S, Berentsen K, Bomba L, Bourque G, Bujold D, Busche S, Caron M, Chen SH, Cheung W, Delaneau O,

Dermitzakis ET, Elding H, Colgiu I, Bagger FO, Flicek P, Habibi E, Iotchkova V, Janssen-Megens E, Kim B, Lehrach H, Lowy E, Mandoli A, Matarese F, Maurano MT, Morris JA, Pancaldi V, Pourfarzad F, Rehnstrom K, Rendon A, Risch T, Sharifi N, Simon MM, Sultan M, Valencia A, Walter K, Wang SY, Frontini M, Antonarakis SE, Clarke L, Yaspo ML, Beck S, Guigo R, Rico D, Martens JH, Ouwehand WH, Kuijpers TW, Paul DS, Stunnenberg HG, Stegle O, Downes K, Pastinen T, Soranzo N. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell. 2016;167(5):1398-414.e24.

17. Farlik M, Halbritter F, Muller F, Choudry FA, Ebert P, Klughammer J, Farrow S, Santoro A, Ciaurro V, Mathur A, Uppal R, Stunnenberg HG, Ouwehand WH, Laurenti E, Lengauer T, Frontini M, Bock C. DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. Cell Stem Cell. 2016;19(6):808-22.
18. Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, Cairns J, Wingett SW, Varnai C, Thiecke MJ, Burden F, Farrow S, Cutler AJ, Rehnstrom K, Downes K, Grassi L, Kostadima M, Freire-Pritchett P, Wang F, Stunnenberg HG, Todd JA, Zerbino DR, Stegle O, Ouwehand WH, Frontini M, Wallace C, Spivakov M, Fraser P. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell. 2016;167(5):1369-84.e19.
19. Libertini E, Heath SC, Hamoudi RA, Gut M, Ziller MJ, Czyz A, Ruotti V, Stunnenberg HG, Frontini M, Ouwehand WH, Meissner A, Gut IG, Beck S. Information recovery from low coverage whole-genome bisulfite sequencing. Nat Commun. 2016;7:11306.
20. Libertini E, Heath SC, Hamoudi RA, Gut M, Ziller MJ, Herrero J, Czyz A, Ruotti V, Stunnenberg HG, Frontini M, Ouwehand WH, Meissner A, Gut IG, Beck S. Saturation analysis for whole-genome bisulfite sequencing data. Nat Biotechnol. 2016.
21. Moreau T, Evans AL, Vasquez L, Tijssen MR, Yan Y, Trotter MW, Howard D, Colzani M, Arumugam M, Wu WH, Dalby A, Lampela R, Bouet G, Hobbs CM, Pask DC, Payne H, Ponomaryov T, Brill A, Soranzo N, Ouwehand WH, Pedersen RA, Ghevaert C. Large-scale production of megakaryocytes from human pluripotent stem cells by chemically defined forward programming. Nat Commun. 2016;7:11208.
22. Paul DS, Teschendorff AE, Dang MA, Lowe R, Hawa MI, Ecker S, Beyan H, Cunningham S, Fouts AR, Ramelius A, Burden F, Farrow S, Rowlston S, Rehnstrom K, Frontini M, Downes K, Busche S, Cheung WA, Ge B, Simon MM, Bujold D, Kwan T, Bourque G, Datta A, Lowy E, Clarke L, Flicek P, Libertini E, Heath S, Gut M, Gut IG, Ouwehand WH, Pastinen T, Soranzo N, Hofer SE, Karges B, Meissner T, Boehm BO, Cilio C, Elding Larsson H, Lernmark A, Steck AK, Rakyan VK, Beck S, Leslie RD. Increased DNA methylation variability in

type 1 diabetes across three immune effector cell types. Nat Commun. 2016;7:13555.

23. Schutte J, Wang H, Antoniou S, Jarratt A, Wilson NK, Riepsaame J, Calero-Nieto FJ, Moignard V, Basilico S, Kinston SJ, Hannah RL, Chan MC, Nurnberg ST, Ouwehand WH, Bonzanni N, de Bruijn MF, Gottgens B. An experimentally validated network of nine haematopoietic transcription factors reveals mechanisms of cell state stability. Elife. 2016;5:e11469.
24. Schuyler RP, Merkel A, Raineri E, Altucci L, Vellenga E, Martens JH, Pourfarzad F, Kuijpers TW, Burden F, Farrow S, Downes K, Ouwehand WH, Clarke L, Datta A, Lowy E, Flicek P, Frontini M, Stunnenberg HG, Martin-Subero JI, Gut I, Heath S. Distinct Trends of DNA Methylation Patterning in the Innate and Adaptive Immune Systems. Cell Rep. 2016;17(8):2101-11.
25. Simeoni I, Stephens JC, Hu F, Deevi SV, Megy K, Bariana TK, Lentaigne C, Schulman S, Sivapalaratnam S, Vries MJ, Westbury SK, Greene D, Papadia S, Alessi MC, Attwood AP, Ballmaier M, Baynam G, Bermejo E, Bertoli M, Bray PF, Bury L, Cattaneo M, Collins P, Daugherty LC, Favier R, French DL, Furie B, Gattens M, Germeshausen M, Ghevaert C, Goodeve AC, Guerrero JA, Hampshire DJ, Hart DP, Heemskerk JW, Henskens YM, Hill M, Hogg N, Jolley JD, Kahr WH, Kelly AM, Kerr R, Kostadima M, Kunishima S, Lambert MP, Liesner R, Lopez JA, Mapeta RP, Mathias M, Millar CM, Nathwani A, Neerman-Arbez M, Nurden AT, Nurden P, Othman M, Peerlinck K, Perry DJ, Poudel P, Reitsma P, Rondina MT, Smethurst PA, Stevenson W, Szkotak A, Tuna S, van Geet C, Whitehorn D, Wilcox DA, Zhang B, Revel-Vilk S, Gresele P, Bellissimo DB, Penkett CJ, Laffan MA, Mumford AD, Rendon A, Gomez* K, Freson* K, Ouwehand* WH, Turro* E. A high-throughput sequencing test for diagnosing inherited bleeding, thrombotic, and platelet disorders. Blood. 2016;127(23):2791-803.
26. Buitrago L, Rendon A, Liang Y, Simeoni I, Negri A, Filizola M, Ouwehand* WH, Coller* BS. αIIbβ3 variants defined by next-generation sequencing: Predicting variants likely to cause Glanzmann thrombasthenia. Proceedings of the National Academy of Sciences. 2015;112(15):201422238-.
27. McKerrell T, Park N, Moreno T, Grove CS, Ponstingl H, Stephens J, Group USS, Crawley C, Craig J, Scott MA, Hodkinson C, Baxter J, Rad R, Forsyth DR, Quail MA, Zeggini E, Ouwehand W, Varela I, Vassiliou GS. Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. Cell Rep. 2015;10(8):1239-45.
28. Westbury SK, Turro E, Greene D, Lentaigne C, Kelly AM, Bariana TK, Simeoni I, Pillois X, Attwood A, Austin S, Jansen SB, Bakchoul T, Crisp-Hihn A, Erber WN, Favier R, Foad N, Gattens M, Jolley JD, Liesner R, Meacham S, Millar CM, Nurden AT, Peerlinck K, Perry DJ, Poudel P, Schulman S, Schulze H, Stephens JC, Furie B, Robinson PN, van Geet C, Rendon A, Gomez K, Laffan MA, Lambert MP, Nurden P, Ouwehand WH, Richardson S, Mumford AD, Freson K, Consortium B-B. Human phenotype ontology annotation and cluster analysis to

unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. Genome Med. 2015;7(1):36.

29. Adoue V, Schiavi A, Light N, Almlof JC, Lundmark P, Ge B, Kwan T, Caron M, Ronnblom L, Wang C, Chen SH, Goodall AH, Cambien F, Deloukas P, Ouwehand WH, Syvanen AC, Pastinen T. Allelic expression mapping across cellular lineages to establish impact of non-coding SNPs. Mol Syst Biol. 2014;10:754.
30. Almlof JC, Lundmark P, Lundmark A, Ge B, Pastinen T, Cardiogenics C, Goodall AH, Cambien F, Deloukas P, Ouwehand WH, Syvanen AC. Single nucleotide polymorphisms with cis-regulatory effects on long non-coding transcripts in human primary monocytes. PLoS One. 2014;9(7):e102612.
31. Bielczyk-Maczynska E, Serbanovic-Canic J, Ferreira L, Soranzo N, Stemple DL, Ouwehand WH, Cvejic A. A loss of function screen of identified genome-wide association study Loci reveals new genes controlling hematopoiesis. PLoS Genet. 2014;10(7):e1004450.
32. Lentaigne C, Freson K, Laffan MA, Turro E, Ouwehand WH, Consortium B-B, et al. Inherited platelet disorders: toward DNA-based diagnosis. Blood. 2016 Jun 9;127(23):2814-23.
33. Greene D, NIHR BioResource, Richardson S, Turro, E. Phenotype Similarity Regression for Identifying the Genetic Determinants of Rare Diseases. American Journal of Human Genetics. 2016, 2016 Mar 3;98(3):490-9
34. Turro E, Greene D, Wijgaerts A, Thys C, Lentaigne C, Bariana TK, et al. A dominant gain-of-function mutation in universal tyrosine kinase SRC causes thrombocytopenia, myelofibrosis, bleeding, and bone pathologies. Science translational medicine. 2016 Mar 2;8(328):328ra30.
35. Stritt S, Nurden P, Turro E, Greene D, Jansen SB, Westbury SK, et al. A gain-of-function variant in DIAPH1 causes dominant macrothrombocytopenia and hearing loss. Blood. 2016 Feb 24.
36. Guerrero, J. A., Bennett, C., van der Weyden, L., McKinney, H., Chin, M., Nurden, P., McIntyre, Z., Cambridge, E. L., Estabel, J., Wardle-Jones, H., Speak, A. O., Erber, W. N., Rendon, A., Ouwehand, W. H. & Ghevaert, C. Gray Platelet Syndrome: Pro-inflammatory megakaryocytes and α-granule loss cause myelofibrosis and confer resistance to cancer metastasis in mice. *Blood* (2014). doi:10.1182/blood-2014-04-566760
37. Westbury SK, Turro E, Greene D, Lentaigne C, Kelly AM, Bariana TK, Simeoni I, Pillois X, Attwood A, Austin S, Jansen SB, Bakchoul T, Crisp-Hihn A, Erber WN, Favier R, Foad N, Gattens M, Jolley JD, Liesner R, Meacham S, Millar CM, Nurden AT, Peerlinck K, Perry DJ, Poudel P, Schulman S, Schulze H, Stephens JC, Furie B, Robinson PN, van Geet C, Rendon A, Gomez K, Laffan MA, Lambert MP, Nurden P, Ouwehand WH, Richardson S, Mumford AD, Freson K; BRIDGE-BPD Consortium. Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. Genome Med. 2015 Apr 9;7(1):36

38. Chen, L., Kostadima, M., Martens, J. H. A., Canu, G., Garcia, S. P., Turro, E., Downes, K., Macaulay, I. C., Bielczyk-Maczynska, E., Coe, S., Farrow, S., Poudel, P., Burden, F., Jansen, S. B. G., Astle, W. J., Attwood, A., Bariana, T., Bono, B. de, Breschi, A., Chambers, J. C., Consortium, B., Choudry, F., Clarke, L., Coupland, P., Ent, M. van der, Erber, W. N., Jansen, J. H., Favier, R., Fenech, M. E., Foad, N., Freson, K., Geet, C. van, Gomez, K., Guigo, R., Hampshire, D., Kelly, A. M., Kerstens, H. H. D., Kooner, J. S., Laffan, M., Lentaigne, C., Labalette, C., Martin, T., Meacham, S., Mumford, A., Nürnberg, S., Palumbo, E., Reijden, B. A. van der, Richardson, D., Sammut, S. J., Slodkowicz, G., Tamuri, A. U., Vasquez, L., Voss, K., Watt, S., Westbury, S., Flicek, P., Loos, R., Goldman, N., Bertone, P., Read, R. J., Richardson, S., Cvejic, A., Soranzo*, N., Ouwehand*, W. H., Stunnenberg*, H. G., Frontini*, M., Rendon*, A., Soranzo, N., Ouwehand, W. H., Stunnenberg, H. G., Frontini, M. & Rendon, A. Transcriptional diversity during lineage commitment of human blood progenitors. *Science (80-.).* **345,** 1251033 (2014).
39. Cvejic, A., Haer-Wigman, L., Stephens, J. C., Kostadima, M., Smethurst, P. A., Frontini, M., van den Akker, E., Bertone, P., Bielczyk-Maczyńska, E., Farrow, S., Fehrmann, R. S. N., Gray, A., de Haas, M., Haver, V. G., Jordan, G., Karjalainen, J., Kerstens, H. H. D., Kiddle, G., Lloyd-Jones, H., Needs, M., Poole, J., Soussan, A. A., Rendon, A., Rieneck, K., Sambrook, J. G., Schepers, H., Sillјé, H. H. W., Sipos, B., Swinkels, D., Tamuri, A. U., Verweij, N., Watkins, N. A., Westra, H.-J., Stemple, D., Franke, L., Soranzo, N., Stunnenberg, H. G., Goldman, N., van der Harst*, P., van der Schoot*, C. E., Ouwehand*, W. H. & Albers*, C. A. SMIM1 underlies the Vel blood group and influences red blood cell traits. *Nat. Genet.* **45,** 542–5 (2013).
40. Albers CA, Paul DS, Schulze H, Freson K, Stephens JC, Smethurst PA, et al. Compound inheritance of a low-frequency regulatory SNPs and a rare null mutation in exon-junction complex subunit RBM8A causes TAR. Nature genetics. 2012 (published online 26 February 2012).
41. Albers CA, Newbury-Ecob R, Ouwehand WH, Ghevaert C. New insights into the genetic basis of TAR (thrombocytopenia-absent radii) syndrome. Current opinion in genetics & development. 2013 Jun;23(3):316-23.
42. Albers CA, Cvejic A, Favier R, Bouwmans EE, Alessi MC, Bertone P, et al. Exome sequencing identifies NBEAL2 as the causative gene for gray platelet syndrome. Nature genetics. 2011 Aug;43(8):735-7.

Data requirements

Data scope. *Describe the groups of participants on whom you require data and the form in which you plan to analyse the data (e.g. phenotype data, filtered variant lists, VCF, BAM). Where participants fall outside the disorders within your GeCIP domain, please confirm whether you have agreement from the relevant GeCIP domain. (max 200 words)*

We welcome data on all patients recruited under haematology and haemostasis disorders eligible under criteria for the GEL pilot, Main Programme or the NIHR BioResource pilot phase, together with information about family members. We will also be happy to consider patients recruited to other domains with complex phenotypes that include a blood and haemostasis component. In addition we have a reciprocal arrangements with the Malignant Haematology domains re data-sharing where relevant. We will require all available quantitative and qualitative phenotype data, the appended HPO codes, BAM and VCF files. We also seek access to the BAM and VCF files and the high level appended HPO codes of all other samples analysed by GEL to select a subset for use as ‘controls’

Data analysis plans. *Describe the approaches you will use for analysis. (max 300 words)*

We will use the variant, ethnicity and relatedness output files provided by the GEL analysis team. We will cluster patients in a non-supervised and supervised manner to create groups of cases with an increased likelihood of having variants in the same or closely connected genes. The performance of our methods relies on variant effect predictor tools and knowledge from other databases like the mouse phenotype ontology (MPO) project, OMIM/Orphanet, EBI-NIHR GWAS catalogue. Our analysis methods can be conditioned for expected mode of inheritance pattern, segregation within families, and similarity to variants in other affected cases with similar phenotypes. We have developed a suite of applications to illustrate the presence of candidate variants in the context of the clinical and laboratory phenotypes. Variants within known disease genes will be interrogated first using HGMD, ClinVar and gene-specific reference databases as well as allele count catalogues such as GnomAd, UK10K and other large scale genome sequencing projects. We will take into account metrics such as the gene damage index (degree to which a given gene tolerates deleterious mutations within a population), and biological factors including known interactions with relevant biomolecules. After completing the Tier 1 analysis we will consider putative novel association signals including those within noncoding space. We are beginning to define the levels of functional annotation required to overlay the genome using cell-type specific IHEC reference epigenome maps and regulome builds that relate promoters and regulatory elements.

Key phenotype data. *Describe the key classes of phenotype data required for your proposed analyses to allow prioritisation and optimisation of collection of these. (max 200 words)*

All available clinical and laboratory phenotypic data related to patients and their family members within our domain (according to the data models we have helped to develop and data which are being copied from the NIHRBR-RD).

Alignment and calling requirements. *Please refer to the attached file (Bioinformatics for 100,000 genomes.pptx) for the existing Genomics England analysis pipeline and indicate whether your requirements differ providing explanation. (max 300 words)*

In the first instance we anticipate using the GEL pipeline.

Tool requirements and import. *Describe any specific tools you require within the data centre with particular emphasis on those which are additional to those we will provide (see attached excel file List_of_Embassy_apps.xlsx of the planned standard tools). If these are new tools you must discuss these with us. (max 200 words)*

Development of a test ('sandbox') system that mirrors the setup of the real data centre, but does not contain the data of the Main Programme and thus need not be restricted to the same level with respect to data and code import and export. This test system should contain benchmarking data (e.g. the Illumina's platinum genomes, the 8,000 WGS-BAM files from the NIHR BR-RD pilot phase or Genome In A Bottle (GIAB) consortium data) that may then be used for measuring the accuracy of the developed analysis pipeline and would enable incremental pipeline development and optimization.

Data import. *Describe the data sets you would require within the analysis environment and may therefore need to be imported or accessible within the secure data environment. (max 200 words)*

NIHR BR-RD
TCGA
UK Biobank/INTERVAL GWAS datasets for blood cells (Astle et al, Cell 2016)
BLUEPRINT (Chen et al, Cell 2016; Stunnenberg et al, Cell 2016), Roadmap, FANTOM and other IHEC Epigenome datasets
Interaction data between promoters and regulatory elements (Javierre et al, Cell 2016)
Expression QTL data for blood cells (joint Oxford-Cambridge initiative)
A fast suite of analytical application tools which have been developed by the NIHR BioResource and GEL bioinformaticians

Computing resource requirements. *Describe any analyses that would place high demand on computing resources and specific storage or processing implications. (max 200 words)*
We do envisage issues that will be significantly different from the rest of the rare disease community because of the extensive information that we need about the functional annotation of the regulatory builds of the genome in tens of different blood cell types. We look forward to working with Augusto Rendon and team in the new compute environment to implement this functionality.

Omics samples

Analysis of omics samples. *Summarise any analyses that you are planning using omics samples taken as part of the Project. (max 300 words)*

We have obtained a first example where analysis by Metabolom (http://www.metabolon.com/) has aided gene discovery and may therefore seek access to EDTA plasma for further studies. In certain circumstances, linked samples of plasma or whole blood RNA might be helpful in interrogating specific aspects of individual phenotypes. However the main blood component of interest to us is its cells, which are not currently being collected as part of the Project. We are performing a pilot study of enhanced cellular characterisation on primary cells and IPSCs using genomics and proteomics (with Prof Lamond FRS (Dundee) and Profs Durbin (Sanger), Prof Vallier (Sanger/University of Cambridge) and Watt (Kings) but these studies are best performed with freshly obtained cells.

Data access and security

GeCIP domain name	**Non-malignant haematological and haemostasis disorders**
Project title *(max 150 characters)*	**Towards a comprehensive genetic architecture for heritable blood and haemostasis disorders**

Applicable Acceptable Uses. *Tick all those relevant to the request and ensure that the justification for selecting each acceptable use is supported in the 'Importance' section (page 3).*

X Clinical care

□ Clinical trials feasibility

X Deeper phenotyping

X Education and training of health and public health professionals

X Hypothesis driven research and development in health and social care - observational

□ Hypothesis driven research and development in health and social care - interventional

X Interpretation and validation of the Genomics England Knowledge Base

X Non hypothesis driven R&D - health

□ Non hypothesis driven R&D - non health

□ Other health use - clinical audit

□ Public health purposes

□ Subject access request

X Tool evaluation and improvement

Information Governance

X The lead and sub-leads of this domain will read and signed the Information Governance Declaration form provided by Genomics England and will submit by e-mail signed copies to Genomics England alongside this research plan.

Any individual who wishes to access data under your embassy will be required to read and sign this for also. Access will only be granted to said individuals when a signed form has been processed and any other vetting processes detailed by Genomics England are completed.

Other attachments

Attach other documents in support of your application here including:

- a cover letter (optional)
- CV(s) from any new domain members which you have not already supplied (required)
- other supporting documents as relevant (optional)