GeCIP Detailed Research Plan Form

Background

The Genomics England Clinical Interpretation Partnership (GeCIP) brings together researchers, clinicians and trainees from both academia and the NHS to analyse, refine and make new discoveries from the data from the 100,000 Genomes Project.

The aims of the partnerships are:

1. To optimise:

- clinical data and sample collection
- clinical reporting
- data validation and interpretation.

2. To improve understanding of the implications of genomic findings and improve the accuracy and reliability of information fed back to patients. To add to knowledge of the genetic basis of disease.

3. To provide a sustainable thriving training environment.

The initial wave of GeCIP domains was announced in June 2015 following a first round of applications in January 2015. On the 18th June 2015 we invited the inaugurated GeCIP domains to develop more detailed research plans working closely with Genomics England. These will be used to ensure that the plans are complimentary and add real value across the GeCIP portfolio and address the aims and objectives of the 100,000 Genomes Project. They will be shared with the MRC, Wellcome Trust, NIHR and Cancer Research UK as existing members of the GeCIP Board to give advance warning and manage funding requests to maximise the funds available to each domain. However, formal applications will then be needed to individual funders. They will allow Genomics England to plan shared core analyses and the required research and computing infrastructure to support the proposed research. They will also form the basis of assessment by the Project's Access Review Committee, to permit access to data. Some of you have requested a template for the research plan which we now provide herewith.

We are only expecting one research plan per domain and have designed this form to contain common features with funder application systems to minimise duplication of effort. Please do not hesitate to contact us if you need help or advice.

Domain leads are asked to complete all relevant sections of the GeCIP Detailed Research Plan Form, ensuring that you provide names of domain members involved in each aspect so we or funders can see who to approach if there are specific questions or feedback and that you provide details if your plan relies on a third party or commercial entity. You may also attach additional supporting documents 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)

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

Application Summary	
GeCIP domain name	**Enhanced Interpretation (Previously known as Validation and Feedback)**
Project title *(max 150 characters)*	**Identification of variants underlying ultra-rare diseases and complex variant analysis**

Objectives. *Set out the key objectives of your research. (max 200 words)*

1. To identify the genetic causes of ultra-rare diseases not in other GeCIP categories.
2**. C**omprehensive re-**A**nalysi**S** of unsolved **P**aediatric-onset **A**utosomal **R**ecessive Disorders (CASPAR) – in collaboration with the paediatric GeCIP.
3. Determination of the optimum techniques for orthogonal verification of complex pathogenic variants in NHSE Genomic Hub Laboratories.

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)*

Some of the patients and families recruited to the 100,000 Genomes Project have ultra-rare conditions that do not fit into a single group e.g. nerve or heart diseases. These conditions affect a tiny number of individuals and so very large studies like the 100,000 Genomes Project and international collaboration are required to establish the specific genetic cause.
Our previous work indicates that a large number of rare disorders that affect children and are likely to be inherited in a pattern where brothers and sisters are more likely to be affected (autosomal recessive) remain unsolved. We will undertake analyses of the whole genome sequence data to identify the causes of these conditions. This work will determine the specific diagnoses for these conditions removing uncertainty, the need for unnecessary tests, informs reproductive choices and potential treatments.
We plan to work out the best techniques to accurately confirm genetic test results from whole genome sequence data in clinical laboratories. When a single letter is changed in the sequence we may not need extra checks but for complex changes there is a lot of work to do. This will ensure accurate, rapid cheap testing for at risk family members.
Sometimes changes in genes can lead to many different health problems and genes that seem very similar to each other can lead to different problems. Rather than searching for genetic changes in a specific patient group we propose to look at changes in certain selected genes across the entire the 100,000 data and see if these are present in patients with similar/overlapping health problems (genotype first approach).

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)*

1. **Ultra-rare disease:** We will undertake a comprehensive analysis of the clinical features through comparison of HPO terms to determine if unrelated individuals could have the same diagnosis. This will allow comparison of whole genome data i.e. to determine if variants in the same gene are shared in unrelated individuals or by analysis of trios (unaffected parents and children) to identify spontaneous (de novo) causative variants.

2. **CASPAR** – We will identify families with rare diseases with the following criteria for analysis. No known diagnosis following initial 100k Genomes analysis; Affected siblings (multiplex family) OR consanguineous parents OR a single affected individual with a well-defined phenotype suggestive of a specific autosomal-recessive disorder. We expect that this will include ~1500 – 1800 cases recruited across England. We will then prioritise cases for study based on recognition of shared clinical features. In cases where there is a known clinical diagnosis and only one or no pathogenic variants in the known causative gene has been identified we will screen the data for intronic variants and undertake RNA based analysis or determine if structural variants are present. For consanguineous families we will prioritise the analysis of novel or ultra rare homozygous variants – initially coding variants but then non-coding variants in or near attractive candidate genes based on known function.

3- Validation techniques: It is clear that genome sequence data is of high quality in detecting single nucleotide variants (SNVs) and that validation is not technically challenging. However,
Working with colleagues across all GMCs we will consider **complex mutational mechanisms** e.g. structural variants like inversions, intronic variants that create cryptic splice sites to determine criteria around the optimum approaches for validation. These will result in best practice guidelines to ensure optimal clinical testing strategies.

4. Genotype driven research: Previously we and others have demonstrated that systematic re-analysis of the genomic data can help in the identification of novel disorders and reveal new disease mechanisms. In this project we will analyse genomic data of patients recruited into the rare diseases arm of the 100,000 Genomes project. The patients for this study will be selected on the basis of their genotype e.g. variants in a specific gene or pathway. The genes will be selected on the basis of their involvement in a particular pathway, process or knowledge of other related genes associated with inherited disorders. We expect that the phenotypes will be very broad and span a range of phenotypes across different age ranges and so it is not possible or appropriate to focus on a single clinical participant set. With recruiting physicians, we will undertake reverse phenotypic (assess whether certain features perhaps originally not reported are present - **reverse phenotyping**). Where relevant, for selected variants/genes of interest we will interrogate the 100,000 Genomes Data to identify additional patients from other GMCs.

Expected start date	**1st December 2018**
Expected end date	**30th November 2021**

Lead Applicant(s)	
Name	Prof William Newman
Post	Professor of Translational Genomic Medicine
Department	Manchester Centre for Genomic Medicine
Institution	University of Manchester
Current commercial links	Nil

Administrative Support	
Name	Dr Rachel Mahood
Email	Rachel.mahood2@mft.nhs.uk
Telephone	0161 701 9139

Subdomain leads		
Name	**Subdomain**	**Institution**
Dr Siddharth Banka	Project 1 and 2, 4	University of Manchester
Dr Richard Scott	Project 1,2, and 4	Genomics England and GOSH
Prof Sian Ellard	Project 1-4	University of Exeter
Dr Caroline Wright	Project 1 and 2, 4	University of Exeter
Dr Emma Baple	Project 1-4	Genomics England, University of Exeter
Dr Helen Firth	Project 1,2 and 4	Addenbrookes Hospital, Cambridge
Dr Steve Abbs	Project 3	Addenbrookes Hospital, Cambridge
Mr Dominic McMullan	Project 3	Birmingham Women's Hospital
Dr Hywel Williams	Project 3,4	UCHL

Detailed research plan

Full proposal (total max 1500 words per subdomain)	
Title *(max 150 characters)*	**Identification of variants underlying ultra-rare diseases and complex variant analysis**

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).*

In this GeCIP we will focus on solving the most challenging cases. After much discussion, we have selected to focus on three groups of patients or variants: the non-specific categories within 100,000 Genomes Project which have been recruited to include:
· Single autosomal recessive mutation in rare disease (125 participants from 51 families)
· Undiagnosed monogenic disorder seen in a specialist genetics clinic (802 participants from 317 families)
· Ultra-rare undescribed monogenic disorders that do not fit into any other clinical category (3454 participants from 1327 families)

In addition we will determine best practice guidelines for the validation of complex variants. This latter objective will be of benefit to participants through improved diagnosis and to the health system, through improvements in the efficient use of resources.

Patients with ultra-rare disorders often wait the longest time before their diagnosis is made. This is in some degree due to lack of funding, but also the challenge of independently confirming that variant(s) in a gene result in a specific phenotype in more than one family as the families are so difficult to ascertain. **This is a major clinical unmet need.** We will invest significant energies to determine the causes of these disorders. Within the 100K Genomes project, a subset of patients do not fulfil specific inclusion criteria for other disease groups but their phenotypes are highly suggestive of monogenic disorders. Dissection of the underlying cause is going to be more challenging in these patients because of poor statistical power due to the unique nature of their phenotypes. We have shown that focussed phenotype-guided interrogation of the genomic data can successfully solve a significant proportion of these cases. **This fulfils a major unmet clinical need and has the potential to result in important scientific breakthroughs.**

Accurate genetic diagnosis of autosomal recessive disorders (AR) provides an option for preventing recurrence, which is usually not possible for *de novo* disorders. In addition, a large proportion of **treatable** genetic disorders are AR (reviewed in van Karnebeek 2012). Rare AR disorders disproportionately affect children born to consanguineous ethnic minority populations in the UK (Sheridan 2013). This is especially important for regions in England with ethnically diverse populations. Hence, **the lower than expected diagnostic yield for AR disorders, especially in large sequencing studies, is an important problem to address to ensure that the transformative potential of 100,000 Genomes Project (100K GP) is truly fulfilled and that the project is benefits all section of the society.** This part of the project will require the collaboration with the paediatrics GeCIP and links into other major clinical sequencing initiatives including DDD.

Validation of single nucleotide variants is not techinally challenging. However more complex variants can be detected by genome sequence analysis. Therefore, it is vital that the results indicating possible structural variations (chromosomal inversions) or deep intronic variants resulting in splicing defects can be validated in an independent way to 1. Ensure the accuracy of

the bioinformatics tools analysing the WGS data and to allow testing of other at risk family members. This work will have implications for all seven of the new Genomic Laboratory Hubs within NHSE as well as diagnostic laboratories worldwide wishing to confirm the analyses of WGS.

We already have evidence that a genotype driven - reverse phenotype approach can reveal novel diagnoses in large datasets (e.g successful in DDD). We wish to apply this approach to the 100,000 Genomes data set. This will lead to important diagnoses for patients with rare disease where the diagnosis has not previously been achieved.

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.*

Case selection criteria:

For Project 1 we will include all patients recruited in the ultra-rare category in the 100,000 Genomes Project. In addition individuals where there is a known pathogenic copy number (i.e. deletion) or single nucleotide variant and the second causative variant on the trans allele is unidentified.

The clinical characterisation of cases for prioritisation based on the likelihood of diagnostic yield will be undertaken. We will select families/individuals where the phenotype is consistent across a number of families/individuals as this will increase the likelihood of successful identification of a causative gene. Additionally, in-depth genome homozygosity measures will be undertaken to assess the degree of autozygosity in families, to identify and prioritise founder sequence variants located in these regions. A number of factors will be taken into consideration e.g. family structure, number of available samples from affected and unaffected family members.

Analysis: We will screen all de novo tier 3 variants and all biallelic tier 3 variants in affected individuals and then undertake a systematic analysis of variant allele frequency analysis in (ethnically matched and general population) controls, segregation, *in silico* analysis and candidate determination (i.e. likelihood that the identified gene is relevant to the clinical phenotype) in these cases. Follow up functional studies will be undertaken as appropriate.

Additional cases will be sought through international initiatives e.g. Genematcher and clinical interactions and community focussed genomic research groups.

For Project 2 we will include patients with the CASPR inclusion criteria set out above.

For Project 3 we will include cases identified by individual GMCs or other GeCIPs with complex variants for validation.

There will be a number of individual projects that will be led by members across different GMCs and dependent on expertise in different clinical phenotypes/molecular pathways. It is likely that a number of these projects will take 2 – 3 years (however some may be quicker dependent upon availability of data about the putative causative gene and the numbers of ascertained cases).

The major challenge is the interpretation of rare/novel variants in the context of a very rare condition where potentially only a single family is available for study.

With regard to the determination of orthogonal tests to confirm in a clinical setting complex variants e.g. structural variants, approaches to RNA analyses these will be iterative approaches that will develop across the entire GMC constituency dependent upon the variants identified and local expertise in RNA splicing, minigene assays, droplet PCR, FISH analysis etc.

We expect this work will continue through the lifespan of the GeCip as new bioinformatics approaches will determine different variant types that will require validation.

For Project 4:

We will select certain genes based many different approaches to find the causes of rare disease.

For example data from ongoing research in our own groups has identified many families where we have identified a compelling genetic explanation for the disorder (it has not been possible to identify other families with changes in the same gene with a similar phenotype through collaboration or matching tools like Genematcher). The 100K Genomes dataset provides an enormous resource of genomic data that can be interrogated to identify other individuals with variants in the same gene potentially associated with the same/overlapping/different phenotypes. All of these outcomes provide important information about disease association with the gene and variant pathogenicity.

Knowledge about a specific biochemical pathway or disease process or gene family will indicate genes that may be more likely to be associated with disease - however the specific phenotype will not be obvious. Therefore genes belonging to defined groups based on known function, paralogues or similarities to known disease causing genes will be prioritised.

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.*

We will work with a number of other GeCIPS
1. Enabling rare disease translational genomics via advanced analytics and international interoperability
2. Paediatrics
3. Quantitative methods, machine learning and functional genomics

In addition, we have had discussions with the new leads Prof David FitzPatrick, MRC Human Genetics Unit, University of Edinburgh and Dr Fiona Cunningham, at EBI of the Transforming Genomic Medicine Initiative funded by the Wellcome Trust. We will work with them to define optimum transcripts for clinical testing and establish a dialogue to determine further how we can work closely together.

The approaches will mirror some of those that members of this Domain have already applied successfully within the DDD project.

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

We envisage that STPs in Genomics and genomic scientists on the HSST program will undertake projects within this domain. Many potential supervisors are GeCIP members and this analysis will be ideal for clinically orientated projects.
Rare disease gene discovery work has formed the basis of a number of Masters and PhD studentships and this program will facilitate additional opportunities in this area. We would encourage students to undertake PhDs/masters studying in these research areas and apply for funds through Charities, Research Councils and HEE.

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

We have extensive experience in gene discovery for rare and ultra-rare diseases using exome and whole genome sequence analysis. Previous work has identified non-coding and complex variants underlying such conditions, which are likely to be enriched in the 100K dataset.

The **key investigators in the group have been meeting regularly over the past three years to support Genomics England** in the design of the pipeline, interpretation of results and feedback of these to health professionals and patients. This group will now act as a focus to link with other GeCIP members to develop the research proposals set out above.

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.*

Our group members have been integral to the adoption of the American College of Medical Genetics guidance for variant interpretation across the UK. We have provided updates and modifications to these guidelines and Ellard and Baple have led monthly webexes for clinical scientists and clinicians from all laboratories across the country to work through practical examples and implement this new practice.
Our members have extensive experience in variant interpretation in both known genes and in the robust evidence required to define a novel disease-causing gene associated with a rare disorder, with a number of seminal publications in this field.

There is a member of this GeCIP in each of the GMCs to ensure that all Centres are able to seek support and input in variant interpretation and a member of our GeCIP is already in place in each of the other rare disease GeCIPs.

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

The major benefit of our work will be **to increase the diagnostic yield** in the especially challenging patient group within the ultra-rare monogenic category. These are often orphan disorders with very little research activity or clinical focus as they are so rare or have not previously been delineated or clinically recognised as a discrete entity. Defining these disorders will be of value predominantly to the affected individual(s) and their families so that they have an accurate diagnosis and explanation for their health problems. It will facilitate risk estimation to other family members, inform screening and reproductive decision-making and potentially inform clinical management and provide new therapeutic avenues.

All of these benefits listed above directly benefit the NHS and the health care providers within it. The work on determining the optimal orthogonal tests for clinical validation of genomic sequencing results will directly benefit the genomic laboratories to ensure best practice.

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?*

It is unlikely that this will be relevant. However, there is a possibility that a gene will be identified that when altered results in a rare disease and presents a tractable therapeutic target. We will explore this in detail.

For project 3 we are happy to work with diagnostic companies to develop different approaches to validate variants identified through WGS. We would expect to act as testing partners to access the sensitivity and clinical applicability of orthogonal approaches.

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

Sheridan E et al. Risk factors for congenital anomaly in a multiethnic birth cohort: an analysis of the Born in Bradford study. Lancet. 2013;382:1350-9.
van Karnebeek CD, Stockler S. Treatable inborn errors of metabolism causing intellectual disability: a systematic literature review. Mol Genet Metab. 2012;105:368-81

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 will require access to the three non-specific categories within 100,000 Genomes Project which have been recruited: · Undiagnosed monogenic disorder seen in a specialist genetics clinic · Single autosomal recessive mutation in rare disease · Ultra-rare undescribed monogenic disorders that do not fit into any other clinical category
Data analysis plans. *Describe the approaches you will use for analysis. (max 300 words)* We will undertake intra-familial, intersection and trio based approaches to determine rare variants that segregate with specific phenotypes. There are a number of academic and commercial bioinformatics pipelines that can assist with such analysis and members of the group have many approaches available to undertake this work on bam/vcf files.
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)* This needs to be determined on an individual basis as by its nature this group will likely encompass all phenotypic categories.
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)* No additional requirements – this may evolve
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)* No additional requirements at this stage
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)* Small numbers for each analysis of VCF/bam files from individuals with ultra-rare disorders

Computing resource requirements. *Describe any analyses that would place high demand on computing resources and specific storage or processing implications. (max 200 words)*

Because of the variable phenotypes We will only be undertaking analysis of small collections of genomes at any one time and so we will not be placing large compute requests on the system.

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)*

To study the effects of intronic variants that may create splicing alterations access to PAXgene samples for RNA extraction and creation of cDNA for sequence anlaysis would be very helpful.

Data access and security	
GeCIP domain name	**Validation and Feedback**
Project title *(max 150 characters)*	**Identification of variants underlying rare diseases**

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*

X *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*

☐ *Interpretation and validation of the Genomics England Knowledge Base*

☐ *Non hypothesis driven R&D - health*

☐ *Non hypothesis driven R&D - non health*

☐ *Other health use - clinical audit*

☐ *Public health purposes*

☐ *Subject access request*

☐ *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.