GeCIP Detailed Research Plan Form v2

May 2018

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	**Respiratory**
Project title *(max 150 characters)*	**Hypothesis-driven research and development to establish the functional role of DNA sequence variants in respiratory rare diseases**

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

We propose a simple national mechanism for validation and support of individual Genomic Medicine Centres (GMCs), together with innovative use of the whole genome sequencing (WGS) data which will be generated by the *100,000 Genomes Project,* focussing on applications of most relevance to patients within our six disease-specific subdomains. We will incorporate deeper phenotyping led from those best placed to direct this, namely the Subdomain teams that are in place, well organised and already undertake this role. Close linkage to the NIHR BioResource will facilitate both this and collaboration with industry. Novel disease causing variants and genes will be identified with genomic interpretation focused at Imperial College, taking advantage of the internationally recognised experience available there. We have assembled international teams with the disease-specific genomic expertise to prioritise sequence variants identified in novel genes, rapidly evaluate function and test pathogenicity in existing *in vitro* and *in vivo* model systems, catalysing progression to diagnostic and potential therapeutic approaches for patients. Feedback will be enhanced by the subdomains' integral partnerships with patient support groups and charities.

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 Respiratory GeCIP is split into six subdomains each dealing with a specific disease group. These sub-domains are: i) Primary Ciliary Dyskinesia (PCD), ii) Familial Interstitial Lung Disease (ILD), iii) Non-CF Bronchiectasis, iv) Pulmonary Arteriovenous Malformations (PAVM) , v) Hereditary Haemorrhagic telangiectasia (HHT) and vi) Familial Pneumothorax.

Each subdomain will undertake research into two areas, i) linking the genes that are known to be disease-causing with the clinical effect it has on the patient, and ii) identifying genes that are expected to be disease-causing and confirming or refuting this in laboratory models.

The combination of both research areas will help to understand better the genetic cause of the diseases and their effects, allowing more tailored and effective screening and treatment for patients.

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)*

The goal of these proposals from the Respiratory GeCIP is to use whole genome data obtained through the *100,000 Genomes* project for **hypothesis-driven research and development in health** in six specific diseases: Primary Ciliary Dyskinesia (PCD), Familial Pulmonary Fibrosis (FPF),

Aggressive Bronchiectasis, Pulmonary Arteriovenous Malformations (PAVM), Hereditary Haemorrhagic Telangiectasia (HHT) and Familial Pneumothorax.

Our primary goal is to identify new disease genes using variant call format (vcf) text files generated through the GeL pipeline, which we will filter to prioritise candidate variants based on gene knowledge and *in silico* prediction tools. Whole genome sequence variants will be validated in source gDNA from the affected individual and affected relatives, with precise methods determined by number of variants/genes to be evaluated. Confirmation that the gDNA variant modifies the relevant gene products will be obtained using accessible patient samples that provide sufficient expression of the appropriate gene, splice variant, and/or protein. Pathogenicity checks will use functional readouts already in use in our laboratories for analyses of existing disease genes. Assays in patient-derived samples will be supplemented using expression vectors to enable structural analysis of recombinant wildtype and variant proteins *in vitro.*
In all cases, it is anticipated that once novel genes are identified through GeL, genotyping will take a prominent and cost-efficient role in the diagnostic portfolio for the respective disease.

Further research outputs differ according to the specific disease/subdomain:

a) The **PCD** subdomain will focus on advancing the understanding of the ultrastructural cilial defects using immunofluorescence, and 3D electron tomography;

b) The **Familial Pulmonary Fibrosis** subdomain will examine phenotype-genotype correlations in relation to radiological, spirometric and gas exchange indices, and evaluate the effects of manipulating gene expression in the presence and absence of anti-fibrotic and anti-inflammatory agents.

c) The **Aggressive Bronchiectasis** subdomain will categorise "extreme phenotype" cohorts, allowing clustering of sub-phenotypes, and use the initial analysis of whole genome data to produce a small panel of genes with a role in bronchiectasis. Analyses will be undertaken in collaboration with the Primary Immunodeficiency and Paediatric GeCIPs

d) The **Hereditary Haemorrhagic Telangiectasia (HHT)** subdomain will modify gene expression in zebrafish and chicken embryos to elucidate signalling pathways, and use a Fli1:EGFP vascular reporter line to examine vascular formation/stability. Novel premature termination codons (PTCs) will be screened for potential correction by read-through drugs such as ataluren using robust *in vitro* fluorescent reporter assays, and patient-derived cells.

e) The **Pulmonary Arteriovenous malformation (PAVM)** subdomain will also use embryological techniques, and examine if there are differences between PAVM genotypes in the patients' abilities to compensate for their PAVMs in phenotypes related to oxygen delivery and compromised pulmonary capillary bed filtration/function.

f) The **Familial Spontaneous Pneumothorax (FSP)** subdomain will develop cell based and cell free assays for gene product function to enable functional assessment of mutations within signalling pathways. Defects in pathways related to formation of the extracellular matrix will be analysed by examining fibroblast secretory function.

Studies will be facilitated by a strong training arm (based around the HEE-sponsored Genomic Medicine MSc and modules), close linkage to the NIHR Rare Disease TRC, industrial collaborations and integral partnerships with recruiting Genomic Medicine Centres (GMCs), national specialist groups, and patient support groups and charities.

Expected start date	**01/05/16**
Expected end date	**30/4/19**

Lead Applicant(s)	
Name	Eric Alton
Post	Professor of Gene Therapy and Respiratory Medicine, Honorary Consultant Royal Brompton and Harefield NHS Foundation Trust, Director Respiratory BRU, Respiratory GeCIP lead
Department	Gene Therapy
Institution	Imperial College
Current commercial links	ID Pharma (formally DNAVEC), Sanofi-Genzyme,

Administrative Support	
Name	Tracy Higgins
Email	t.higgins@imperial.ac.uk
Telephone	0207 594 7932

Subdomain leads		
Name	**Subdomain**	**Institution**
Claire Hogg	PCD	Royal Brompton and Harefield Foundation Trust
Gisli Jenkins	Familial ILD	University of Nottingham
Anthony De Soyza	Non-CF Bronchiectasis	Newcastle University, BronchUK partnership (www.bronch.ac.uk)
Claire Shovlin	HHT	Imperial College London
Claire Shovlin	PAVM	Imperial College London
Stefan Marciniak	Familial Pneumothorax	University of Cambridge

Detailed research plan

Full proposal (total max 1500 words per subdomain)	
Title *(max 150 characters)*	**Primary Ciliary Dyskinesia**

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).*

Primary Ciliary Dyskinesia (PCD) is a rare, complex and heterogeneous inherited disorder affecting primarily the motile respiratory cilia. The diagnostic pathway for PCD includes the assessment of clinical symptoms suggestive of PCD and where possible non-invasive measurement of nasal nitric oxide. If suggestive these investigations are followed by a nasal brush biopsy assessed by light and electron microscopy. Light microscopy assessment of cilia function is by high speed video analysis of the frequency and pattern (waveform) of cilia movement on live cells. Electron microscopy allows visualisation of the ultrastructure of cilia and can often provide a definitive diagnosis. In 15-30% of cases where ciliary ultrastructure is normal the expanding knowledge of PCD-associated gene mutations is furthering diagnostic capabilities. To date, more than 39 disease-associated mutations have been identified, which encode proteins involved in ciliary synthesis, structure and function, and are estimated to account for 72%of known PCD cases. A combination of genetics

and downstream protein analysis using advanced techniques such as 3D electron tomography and immunoflourescent antibody staining has already confirmed an array of new disease causing mutations, and this area will be greatly enhanced by the whole genome data generated through this project.

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.*

We would hope to receive datasets back from GeL within a 12 month period, with the intention of applying the generated data to our own research programmes within 6-12 months. At the forefront we plan to expand current programmes using immunofluorescence and 3D tomography to advance our understanding of the downstream effects of disease causing mutations. Immunofluorescence has been used extensively in PCD research in confirming protein absence due to gene mutations [1]. A number of antibodies to the structures defective in PCD have been developed, validated and described in this capacity. These include DNAH5 (an outer dynein arm heavy chain), DNALI1 (an inner dynein arm light chain), GAS8 (a component of the nexin dynein regulatory link), RSPH4A, RSPH9 and RSPH1 (components of the radial spoke) and IFT88 a component of the intra-flagellar transport mechanisms required for axonemal assembly. These antibodies represent the five key ultrastructural abnormalities detected by electron microscopy which are the end products of multiple gene defects. The identification of new gene mutations will allow further research in this field to identify additional ultrastructural defects.
We employ electron tomography, a high resolution electron microscope technique, to elucidate in three dimensions the ultrastructural arrangements caused by PCD gene mutations, a specific example being of the *RSPH4A* gene. We demonstrated that the central pair can be present within the cilium. In some cilia, the central pair rotates at the base of the axoneme, and it is likely that this rotation gives rise to the intermittent appearance of the central pair when viewed under conventional electron microscopy [2].

The challenges facing this area of research are both financial and practical. Data analysis and application of the generated data requires an extensive team of post-doctoral level researchers to drive these ambitions forward. Pilot studies in this field have been very successful in both experimental design and in gene discovery projects. The opportunity to take the new genomic data and apply it to these research projects on a large scale will greatly enhance our research output and we would aim to submit applications to the relevant funding bodies to fund this research.

The other barrier is on a practical level. Immunoflourescent antibody work, which has been an essential part of gene discovery to date, requires the development of commercial antibodies. Improved understanding of the underlying genetic defects will hopefully lead to the development of further antibodies and thereby improved clinical uptake that will benefit more rapid and easily applied diagnostics.

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.*

PCD sits within the Respiratory GeCIP and will collaborate with the Paediatric GeCIP, in particular the Ciliopathy sub-domain. A letter of support is to follow.

We also collaborate closely with the PCD family Support Group. Please see Fiona Copeland's Lay Summary.

We have worked closely with Dr Hannah Mitchison, Geneticist at University College London, and have many collaborative research projects in progress.

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

In recognition of the transformation of medicine and clinical care that is occurring as a consequence of advances in 'omic' technologies and the expanding educational need in this area, Imperial College is one of the successful bidders in providing an HEE-funded MSc course in Genomic Medicine. Members of the proposed Domain are key members including the Course Director (Lovett), and Shovlin who is the national Clinical Champion for the RCP on the management and delivery of Respiratory Genomics Education. The course had its first intake of eight students in October 2015 and will have a further ~20 students joining in March 2016. All of these students will be eligible to be part of the proposed GeCiP domain. The course aims to provide the skills required for the analysis and interpretation of genome-scale data. It is taught in a modular format to allow certificate, diploma or full MSc degrees to be obtained, and allows for either full or part time attendance. It is taught by leaders in genomic medicine across the entire faculty of medicine at Imperial College. For the full degree course, a key component is the research project that can be laboratory based, purely analytical, or a combination. We anticipate that projects within the Domain will be highly subscribed to by the MSc students. This course will provide a core resource for the Domain (and others), standardising the transfer of expertise to the next generation. Individual modules within the course will also be available as short courses and for CPD for other researchers and clinicians within the proposed domain. The MSc complements existing analytically themed MScs already offered by the Domain partners including the IC MSc in Bioinformatics and Theoretical Systems Biology and the UCL MSc in Computational and Genomic Medicine. The Subdomains will provide disease-specific courses, for example via the UKBronch and COST Action PCD Summer Training Schools and the FPF National Training Course, and continue to develop higher degree training, including patient interaction, to catalyse clinical transformation. Transfer of trainees between Domain centres will be encouraged and new funding sought to catalyse this.

Our PCD research programme employs an array of clinicians and clinical scientists in a research and training capacity. We have 1 MD student, 1 Clinical PCD Fellow, 4 scientists, 1 of whom is also enrolled in the Imperial College MSc in Genomic Medicine and is now engaged in teaching on the course, and a new young researcher joining next year who has also enrolled in the MSc in Genomic medicine. In addition we have a cell scientist engaged in our immunoflourescent antibody research, working alongside the team outlined above. We also have a further PhD student collaborating cross site from Dr Hannah Mitchison's lab at UCL working within our genotyping research programme.

Claire Hogg is also the Head of the BEAT-PCD Training school, funded by the COST Action grant. The first training School will provide access to existing and novel areas of research to young post-doc researchers in the field of PCD and the first training School hosted up to 100 young researchers in April 2016, in Paris. The following 2 Training Schools have grown rapidly and hosted 140 young researchers in 2017 [Valencia] and 160 in 2018 [Lisbon]. The aim is to expand further in 2019 [Poznan] to accommodate the growing interest in this field.

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

The PCD group is well placed to undertake this research on several fronts. The PCD service is a highly regulated Nationally funded service that diagnoses and looks after all patients with PCD in England. The 4 centres work closely together to provide highly audited and standardised diagnostics for PCD, and as a result we have the most well defined patient cohort in the world. This is, therefore, a unique cohort, accessible only through the specialist teams at defined centres. The 4 centres collaborate on research already and we are part of a larger European PCD Taskforce and have successfully partnered an FP-7 grant [BESTcilia] and a COST action [BEAT-PCD]. We have also been a member of both EU PCD Taskforces producing International Guidelines on the diagnosis and management of PCD. These partnerships have led to the first clinical trial in PCD, the first set of diagnostic guidelines, the first International Registry, and the first PCD specific QOL questionnaire. Through these partnerships we have collaborated with teams in Germany to publish several new gene mutations, and continue to work on new genes and ultrastructural defects currently.

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 GeCIP subdomain will liaise closely with the Variant Interpretation, and Validation and Feedback domains, to identify which sequence variants are suitable to report back to the specific GMC teams. Our joint MDT within our GMC will report the outcome of the datasets back to the specific teams. The disease-specific variants will be fed back to patients via the specialist teams within the PCD national Service.

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

Patients and clinicians will benefit from this research. Improved, faster and more accessible diagnostic tests should drive age of diagnosis down. Starting treatment early, aimed at preventing or delaying progression of lung disease, will have enormous benefit in terms of patient morbidity and mortality with the knock-on reduction in Health Care costs. Improved understanding of the molecular genetics of any inherited condition enhances our understanding of a disease, enabling more targeted therapies and the possibility of gene replacement or modifying treatments.

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?*

We have no current commercial partners, but have had a past collaboration with Sigma relating to our research in immunoflourescent antibodies. We have sought intellectual property and commercialisation advice from Imperial Innovations regarding the antibody assay. We are in the final stages of an i4i NIHR grant with a commercial partner (Cosmonio) to develop artificial intelligence interpretation of TEM defects.

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

1. Mislocalisation of DNAH5 and DNAH9 in Respiratory cells from patients with primary ciliary dyskinesia. *Am J Respir Crit Care Med. 171(12) 2005.* Fliegauf M, Olbrich H, Horvath J, Wildhaber JH, Zariwala MA, Kennedy M, et al.

2. Characterizing the Ultrastructure of Primary Ciliary Dyskinesia transposition defect using Electron Tomography. Cytoskeleton 03/2014. T Burgoyne, A Lewis, A Dewar, P Luther, C Hogg, A Shoemark, M Dixon.

3. Accuracy of immune fluorescence in the diagnosis of Primary Ciliary Dyskinesia.
4. *Am J Resp Crit Care Med* Blue-201607-1351OC.R1: Contributing Authors: Shoemark, A; Frost, E; Dixon, M; Ollosson, S; Kilpin, K; Patel, Mitchison, H, Bush, Hogg, C.X-linked primary ciliary dyskinesia due to mutations in *PIH1D3*, a new player in cytoplasmic assembly of outer and inner arm axonemal ciliary dyneins:

Nature Communications.2017. doi:10.1038/ncomms14279. C Olcese, M Patel, A, S Kiviluoto, M Legendre, H Williams, CVaughan, J Hayward, A Goldenberg, R Emes, M Munye, L Dyer, T Cahill, J Bevillard, C Gehrig, M Guipponi, S Chantot, P Duquesnoy, L Thomas, L Jeanson, B Copin, A Tamalet, C Thauvin, JF Papon, A Garin, I Pin, M Fassad, L Jenkins, C Boustred, T Cullup, M Dixon,A Onoufriadis, A Bush, E Chung, S Antonarakis, M. Loebinger, R Wilson, M Armengot, E Escudier, Hogg, C. UK10K, S Amselm, Z Sun, L Bartoloni, JL Blouin, H Mitchison.

5. High prevalence of *CCDC103* p.His154Pro mutation causing primary ciliary dyskinesia disrupts protein oligomerisation and is associated with normal diagnostic investigations.

Thorax. 2018, 73(2):157-166. Shoemark A[1], Moya E[2], Hirst RA[3], Patel MP[4], Robson EA[2], Hayward J[4,5], Scully J[4,6], Fassad MR[4,7], Lamb W[4], Schmidts M[8,9], Dixon M[1], Patel-King RS[10], Rogers AV[1,11], Rutman A[3], Jackson CL[12,13], Goggin P[12,13], Rubbo B[12,13], Ollosson S[1], Carr S[1], Walker W[12,13], Adler B[14], Loebinger MR[11], Wilson R[11], Bush A[1,15], Williams H[16], Boustred C[5], Jenkins L[5], Sheridan E[17], Chung EMK[18], Watson CM[17], Cullup T[5], Lucas JS[12,13], Kenia P[19], O'Callaghan C[3,20], King SM[10,21], Hogg C[1], Mitchison HM[4].

6. Primary ciliary dyskinesia with normal ultrastructure: three-dimensional tomography detects absence of DNAH11.

Eur Respir J. 2018 Feb 21;51(2). Shoemark A[1,2], Burgoyne T[3], Kwan R[3], Dixon M[3], Patel MP[4], Rogers AV[3], Onoufriadis A[4], Scully J[4], Daudvohra F[3], Cullup T[5], Loebinger MR[3], Wilson R[3], Chung EMK[4], Bush A[3,6], Mitchison HM[4,7], Hogg C[3,6,7]

Detailed research plan

Full proposal (total max 1500 words per subdomain)	
Title *(max 150 characters)*	**Functional effects of genetic variants associated with Familial Pulmonary Fibrosis (FPF)**

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).*

Interstitial Lung Diseases (ILDs) are a heterogenous group of lung diseases characterised by inflammation and fibrosis of the alveolar interstitium. The commonest and most serious ILD is Idiopathic Pulmonary Fibrosis (IPF), which is characterised by progressive breathlessness and cough which ultimately leads to respiratory failure and death with a median survival of patients of approximately 3 years (1-2).

IPF is a complex, heterogeneous genetic disorder associated with rare and common sequence variants in many genes (MUC5B, SFTPC, SFTPA2, RTEL1, TERT, and hTR; (1-4)), over 10 novel loci (5), and multiple emerging epigenetic (6-8) and transcriptional (9-11) profiles. In the past five years investigators have found that: 1) genetic risk variants play major and similar roles in the development of both FPF and IPF (5), accounting for 35% of the risk of a disease process previously thought to be idiopathic; 2) a promoter variant in MUC5B, rs35705950, is the strongest risk factor for the development of Familial Pulmonary Fibrosis (FPF) or IPF, however, rs35705950 has a low penetrance (5); 3) rs35705950 can potentially be used to identify individuals earlier in the course of disease (12); 4) rs35705950 is associated with unique biological (13) and clinical (14) IPF phenotypes; and 5) IPF is a complex genetic disease with at least 16 independent loci contributing to the development of this disease (5), pronounced changes in DNA methylation (8), and transcriptional subtypes (11). TERT mutations are the most common single genetic defect found in FPF, and these variants have increasing penetrance in males, with increasing age and have positively associated with fibrogenic environmental exposures (15).

However, 65% of patients do not have known genetic variants explaining their disease. Similarly the genotype phenotype interactions for FPF remain unclear and importantly the effects of known mutations on disease progression and lung function are unknown. Therefore, these studies will determine the effect of different mutations on disease phenotype and progression. Initial genetic analysis will be followed up with functional genomic and proteomic analysis using tissues derived from patients with ILD and FPF.

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.*

We hope to receive datasets back from GeL within a 6 months period, applying the data to our research programmes within 6-12 months. Initial studies will focus on identifying novel genetic mutations associated with different FPF phenotypes and determining whether there are any associations between both established and novel genes associated with Familial Pulmonary Firbosis and the nature of the disease at presentation based on High Resolution Computerised Tomography (HRCT) scan and lung function.

Subsequent studies will focus on generating preliminary data to determine whether genotype predicts disease outcome as measured by change in Forced Vital Capacity (FVC) and Diffusing Capacity of Carbon Monoxide (DLco) after one year and death. Patients will be defined as stable if FVC <10% and DLco <5% from baseline and progressive if FVC >10% or DLco>5% or they have died.

Longer-term studies (24-48 months) will assess functional effects of genetic mutations associated with FPF. Peripheral blood will be obtained from patients with disease-associated mutations and transcriptional profiles will be measured from monocytes and proteins from serum. Results from these studies will be compared with data obtained from the PROFILE study (16) and UKILD GWAS studies which have assessed genetic, transcriptional and protein signatures from patients with IPF. Protein expression associated with genetic variants will be assessed in immunohistochemical tissue sections from patients who have donated lung tissue through the Nottingham Molecular Pathology Node.

Functional genetic studies will be undertaken in fibroblasts obtained from patients with pulmonary fibrosis. Fibroblasts from the UKILD collaborators (>100 lines available) will be screened for genetic variants and gene expression will be manipulated using siRNA strategies. Functional endpoints including cellular proliferation, collagen deposition, alpha smooth muscle actin and TGFbeta activation will be measured. Similarly, luciferase reporter constructs will be used to assess the effect of genetic variants on transcriptional responses. Where genetic variants are associated with abnormalities in protein expression, the effect of inhibiting protein expression using antibody/small molecular inhibitor studies will be measured.

Having defined the effect of genetic variants on disease phenotype and cellular function, functional pharmacogenomic studies will be performed. Cells expressing mutant or wild-type genetic variants (either disease associated or genetically engineered) will be cultured in the presence or absence of anti-inflammatory or anti-fibrotic molecules such as prednisolone, pirfenidone and nintedanib.

The challenges facing this area of research are both financial and practical. The opportunity to take the new genomic data and apply it to the described research projects on a large scale will greatly enhance our research output and we would aim to submit applications to the relevant funding bodies to fund this research

Practical challenges include the power to generate meaningful observations from the initial FPF cohort. These will be mitigated by international collaboration with other groups. Jenkins and Parfrey are already invloved in the Global IPF Network that has collected over 5,000 DNA samples and clinical data from cases of IPF, and the FPF registry, established by Schwartz, has DNA specimens from over 16,000 subjects from 1,400 families with two or more cases of pulmonary fibrosis. If initial studies do not generate sufficient power for definitive results, we will perform appropriately powered prospective studies to validate initial discoveries.

Similarly it is possible that even a collection fibrotic cell lines as large as that available to the UKILD collection will not reveal genetic mutants associated with the observed genetic varients obtained from GeL. This will be mitigated by international collaborations as described before.

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 collaborate with the Musculoskeletal and Machine Learning, Quantitative Methods and Functional Genomics GeCIPs and with the Royal Brompton Clinical Genetics and Genomics Laboratory.

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

Imperial College are bidding to establish an MSc in Genomic Medicine targeted at clinical professionals and the Nottingham MPN is developing an MSc program in Molecular Pathology. Members of the proposed Domain are key members including the Course Director (Lovett), and Shovlin at Imperial, and Jenkins and Johnson (NMPN). These courses will provide the skill sets required for the analysis and interpretation of sequencing data, are modular to allow certificate, diploma or full MSc degrees to be obtained, and allows for either full or part time attendance. Transfer of trainees between Domain centres will be encouraged and new funding sought to catalyse this.

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

The Familial Pulmonary Fibrosis Sub domain includes all ILD clinicians working in centres approved by NHS England to prescribe pirfenidone, and are therefore secondary referral centres for patients with complex or familial pulmonary fibrosis. Furthermore, all 11 Genomic Medicine Centres are represented in the UKILD consortium. The FPF GeCIP subdomain is led by Jenkins who is the Clinical Lead for the Academic Interstitial Lung Disease Unit at the Nottingham University Hospitals Trust, the Pulmonary Fibrosis workpackage lead at the MRC/EPSRC Nottingham Molecular Pathology Node, PI of the UKILD consortium and Co-PI of the PROFILE study (the largest prospective observational study of IPF in the world). Other key members include the clinical ILD leads from the East of England NHS GMC (Parfrey) and West Midlands NHS GMC (Thickett) as well as Tobin who leads the Machine Learning, Quantitative Methods and Functional Genomics GeCIP.

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.*

For Familial Pulmonary Fibrosis genetic data will become part of the Interstitial Lung Disease (ILD) Multi-Disciplinary Team (MDT). We would envisage the incorporation of these data into regular MDT discussions which would be interpreted with a clinical geneticist alongside clinical, radiological and pathologic data. Follow up consultations with the individual would take place with a defined plan for family screening in addition to the individual's clinical care. Ideally this would occur jointly with a respiratory clinician and clinical geneticist.

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

Multiple beneficiaries are likely to result from this work:
Patients and families. We are seeing an increasing number of referrals for family members of patients dying from IPF. The, currently low, expectation of a precise diagnosis of familial pulmonary fibrosis (FPF) conferring, prognostic information and potential family screening will be of significant value and benefit.
Clinicians Treating Patients with ILD. It is anticipated that the data generated from these studies will aid both screening (relatives) and treatment (patients) algorithms by determining the genetic risks associated with the different genetic variants as well as the effect on disease prognosis and possible pharmacogenomics interactions.

Academic researchers. Knowledge of the prevalence of currently known causes of FPF and understanding the molecular genetic variants underlying as yet unknown causes of FPF will inform research into new mechanisms of fibrogenesis with implications for both FPF and IPF. Understanding genetic variance will have an enormous impact on disease stratification approaches.
Pharma. It is likely that at least some of the genetic abnormalities underlying FPF will lie in a pharmacologically tractable area. Current initiatives for repurposing compounds for rare diseases are likely to fit well into this area. Similarly genetic variants that may alter disease progression are particularly for informing clinical trial design.

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?*

There are excellent links with pharma for pre-competitive research. Sub-Domain members have active collaborations on pulmonary fibrosis studies with GSK, Biogen Idec, Galecto, Gilead Boehringer, Creabilis and MedImmune.

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

1. G. Raghu et al., An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. Am J Respir Crit Care Med 183, 788 (Mar 15, 2011).
2. R. Wei et al., Association between MUC5B and TERT polymorphisms and different interstitial lung disease phenotypes. Transl Res, (Dec 17, 2013).
3. T. E. Fingerlin et al., Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. Nat Genet 45, 613 (Jun, 2013).
4. C. Kannengiesser et al. Heterozygous RTEL1 mutations are associated with familial pulmonary fibrosis. Eur Respir J 46, 474 (Aug, 2015).
5. P. W. Noble, C. E. Barkauskas, D. Jiang, Pulmonary fibrosis: patterns and perpetrators. J Clin Invest 122, 2756 (Aug 1, 2012).
6. R. Vij, I. Noth, Peripheral blood biomarkers in idiopathic pulmonary fibrosis. Transl Res 159, 218 (Apr, 2012).
7. H. Ohnishi et al., Comparative study of KL-6, surfactant protein-A, surfactant protein-D, and monocyte chemoattractant protein-1 as serum markers for interstitial lung diseases. Am J Respir Crit Care Med 165, 378 (2002).
8. F. Zuo et al., Gene expression analysis reveals matrilysin as a key regulator of pulmonary fibrosis in mice and humans. Proc Natl Acad Sci U S A 99, 6292 (2002).
9. Pardo et al., Up-regulation and profibrotic role of osteopontin in human idiopathic pulmonary fibrosis. PLoS Med 2, e251 (Sep, 2005).
10. H. Ishii et al., High serum concentrations of surfactant protein A in usual interstitial pneumonia compared with non-specific interstitial pneumonia. Thorax 58, 52 (Jan, 2003).
11. K. E. Greene et al., Serum surfactant proteins-A and -D as biomarkers in idiopathic pulmonary fibrosis. Eur Respir J 19, 439 (Mar, 2002).
12. D. C. King et al., Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome Res 15, 1051 (Aug, 2005).
13. R. C. Boucher, Idiopathic pulmonary fibrosis--a sticky business. N Engl J Med 364, 1560 (Apr 21, 2011).

14. L. M. Boettger, R. E. Handsaker, M. C. Zody, S. A. McCarroll, Structural haplotypes and recent evolution of the human 17q21.31 region. Nat Genet 44, 881 (Aug, 2012).
15. M.S. Devine, C.K.Garcia. Genetic Interstitial Lung Disease. Clin Chest Med 33, 95 (Mar, 2012).
16. R.G. Jenkins et al. Longitudinal change in collagen degradation biomarkers in idiopathic pulmonary fibrosis: an analysis from the prospective, multicentre PROFILE study. Lancet Respir Med. 3, 462 (June, 2015).

Detailed research plan

Full proposal (total max 1500 words per subdomain)

Title *(max 150 characters)*	**Non-CF Bronchiectasis (Bronchiectasis, Informatics and Genetics study (BIGs))**

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).*

Bronchiectasis is a life-limiting chronic infection and airway inflammation syndrome. To date no specifically licenced therapies are available for bronchiectasis reflecting the very limited understanding of the pathophysiology of bronchiectasis. It is highly likely that bronchiectasis is due to a number of tractable underlying defects including novel immunodeficiencies, ciliopathies and epithelial function/ sodium channelopathies. Understanding the frequency of genetic abnormalities within the more "extreme" phenotypes of bronchiectasis will allow further studies such as *ex vivo* new pathway analysis (novel target validation in patients cells), application of new therapies using personalised medicines/clinical stratification (biomarker directed therapies) and better inform prognosis. We will also aim to study how the novel genetic data discovered within more "extreme phenotypes" may apply to the wider bronchiectasis population.

Building on prior work: There are few if any prior genetic studies in bronchiectasis and those that are available have had conflicting results and looked at SNPs only. Likely abnormalities may however include immune protein defects, for example the PI3 kinase pathway and epithelial ion transport such as sodium channelopathies. Structural genes role in the airways cartilage leading to bronchiectasis are suggested by the association of bronchiectasis with Marfans' syndrome and also reports of matrix metalloproteinases in bronchiectasis progression.

We aim to understand the frequency of genetic abnormalities in the following

1) Those with aggressive early onset bronchiectasis
2) Sodium Channelopathies (those with abnormal sweat test sodium levels but no evidence of cystic fibrosis))
3) Familial bronchiectasis
4) Rare syndromes likely to yield discovery eg Mounier-Kuhn tracheobronchomegaly

After identifying possible disease variants we will validate these findings in replication cohorts and in future work funded elsewhere *ex vivo/in vitro* models using siRNA/ knockdown/ drug targeting etc.

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.*

Our analyses will follow the standard process of genetic/clinical phenotyping correlations. The study designs will separate into 2 cohorts of "extreme phenotypes" that will include clustering of sub-phenotypes as follows: abnormal Na channelopathy, early onset aggressive bronchiectasis and a second smaller cohort of familial bronchiectasis. We will use standard techniques as noted below under Bioinformatics.

Our future work will require both analysis of candidate genes /SNPs etc in a wider bronchiectasis population (eg applying for funds to access the BronchUK biobank) and experimental medicine approaches including further characterisation of the pathway defect (eg via project funding applications to MRC/ Wellcome/other) and stratified medicines approaches (DPFS/ DCS/ EME applications). Extension of the findings in rare cases and assessment of these pathway defects in the wider population of patients with less "extreme" bronchiectasis phenotypes will require further applications to MRC/ Wellcome/other for such studies

The major challenges will be recruiting those with extreme phenotypes which will be mitigated by applying to and accessing the BronchUK registry (interrogating 1500 patients' data for "eligible phenotypes"). Furthermore collaboration with UK centres recruiting into the parallel EU registry for Bronchiectasis (EMBARC) offers further possibilities. Patients with possible phenotypes suitable for GeCIP projects will be identified and made known to the host institution by name or by EMBARC identifier. We also aim to have Patient Public Involvement in disseminating the project and its aim to facilitate patient engagement; this would be in collaboration with charities such as the British Lung Foundation.

We expect the first 5 families with familial bronchiectasis to be entered into system within year one. This assumes the process for multi-GMC collection of any families dispersed across England goes live rapidly. We expect over 100 recruits to the "aggressive bronchiectasis" subdomain. Data analysis would be expected in 3 months and applications to access data within 3 months

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.*

Bronchiectasis will work with the Paediatric, Primary Ciliary Dyskinesia and Immunology/ Haematology Domains. It is possible that we will also work with ophthalmology and renal due to associations between these organ systems and pulmonary diseases. *We propose collaborations with the UK Biobank and BronchUK to cross validate any potential new genotypic variants identified as potentially pathogenic.*

Training. *Describe the planned involvement of trainees in the research and any specific training that will form part of your plan.*
In recognition of the transformation of medicine and clinical care that is occurring as a consequence of advances in 'omic' technologies and the expanding educational need in this area, Imperial College is one of the successful bidders in providing an HEE-funded MSc course in Genomic Medicine. Members of the proposed Domain are key members including the Course Director (Lovett), and Shovlin who is the national Clinical Champion for the RCP on the management and delivery of Respiratory Genomics Education. The course had its first intake of eight students in October 2015 and will have a further ~20 students joining in March 2016. All of these students will be eligible to be part of the proposed GeCiP domain. The course aims to

provide the skills required for the analysis and interpretation of genome-scale data. It is taught in a modular format to allow certificate, diploma or full MSc degrees to be obtained, and allows for either full or part time attendance. It is taught by leaders in genomic medicine across the entire faculty of medicine at Imperial College. For the full degree course, a key component is the research project that can be laboratory based, purely analytical, or a combination. We anticipate that projects within the Domain will be highly subscribed to by the MSc students. This course will provide a core resource for the Domain (and others), standardising the transfer of expertise to the next generation. Individual modules within the course will also be available as short courses and for CPD for other researchers and clinicians within the proposed domain. The MSc complements existing analytically themed MScs already offered by the Domain partners including the IC MSc in Bioinformatics and Theoretical Systems Biology and the UCL MSc in Computational and Genomic Medicine. The Subdomains will provide disease-specific courses, for example via the UKBronch and COST Action PCD Summer Schools and the FPF National Training Course, and continue to develop higher degree training, including patient interaction, to catalyse clinical transformation. Transfer of trainees between Domain centres will be encouraged and new funding sought to catalyse this.

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

The Bronchiectasis UK network funded by the MRC "BronchUK" has already identified genetic study of bronchiectasis as a key priority and developed a UK biobank to underpin extending discovery from "extreme phenotypes" within GeCIP into the broader bronchiectasis population. This multidisciplinary network already includes Cookson and De Soyza plus 9 clinical centres with the MRC funded UK national registry and a biobank of 1500 patients indevelopment (www.bronch.ac.uk).

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 clinical interpretation will be underpinned by using the expertise of the MRC funded multidisciplinary BronchUK network which already includes immunology (Kelleher, Imperial), microbiology (Winstanley, Liverpool), genetics (Cookson, Imperial) and 9 clinicians. We propose the BronchUK network will host "data clinics" to aid the clinical interpretation of emerging data. Patients will benefit from rare variants perhaps seen only in one or two clinics being presented to a multicentre audience- this may elicit further referral in from centres. Our work will be shared via the Validation and Feedback domain. As BronchUK has a patient registry with "consent to contact" we anticipate clinical trials will be accelerated through the GeCIP linkage with BronchUK.

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

Patients will benefit from a greater understanding of their condition, as will the NHS through better diagnostics (earlier diagnosis) and better targeted therapies (personalised medicines). These will be downstream of the initial discovery made herein. Pharma are likely to benefit from

better understanding of the defects causing bronchiectasis and the ability to better stratify patients into appropriate trials of targeted therapies.

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?*

The BronchUK bronchiectasis partnership has 5 pharma partners (Bayer, Chiesi, GSK, AstraZeneca and Gilead). These will benefit from better understanding of abnormal inflammation/immunity/ epithelial function in bronchiectasis. For example, several companies are developing novel anti-inflammatories and understanding the frequency of innate immune pathway defects, such as Pi3K pathway defects in aggressive bronchiectasis, may lead to stratified medicines approach. There is also significant pharma interest in Na^+channelopathies as a target.

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

1: Gould CM, Freeman AF, Olivier KN. Genetic causes of bronchiectasis. Clin Chest Med. 2012 Jun;33(2):249-63. doi: 10.1016/j.ccm.2012.03.002. Review. PubMed PMID: 22640844.

2: Pajusalu S, Reimand T, Uibo O, Vasar M, Talvik I, Zilina O, Tammur P, Õunap K. De novo deletion of HOXB gene cluster in a patient with failure to thrive, developmental delay, gastroesophageal reflux and bronchiectasis. Eur J Med Genet. 2015 Jun-Jul;58(6-7):336-40. doi: 10.1016/j.ejmg.2015.04.002. Epub 2015 Apr 20.PubMed PMID: 25907420.

3: McDonnell MJ, Anwar GA, Rutherford RM, De Soyza A, Worthy S, Corris PA, Lordan JL, Bourke S, Afolabi G, Ward C, Middleton P, Middleton D. Lack of association between KIR and HLA-C type and susceptibility to idiopathic bronchiectasis. Respir Med. 2014 Aug;108(8):1127-33. doi: 10.1016/j.rmed.2014.05.017. Epub 2014Jun 18. PubMed PMID: 24986480.

4: Payandeh J, McGillivray B, McCauley G, Wilcox P, Swiston JR, Lehman A. A Clinical Classification Scheme for Tracheobronchomegaly (Mounier-Kuhn Syndrome). Lung. 2015 Oct;193(5):815-22. doi: 10.1007/s00408-015-9757-z. Epub 2015 Jul 19. PubMed PMID: 26189148.

5: Hsieh MH, Chou PC, Chou CL, Ho SC, Joa WC, Chen LF, Sheng TF, Lin HC, Wang TY, Chang PJ, Wang CH, Kuo HP. Matrix metalloproteinase-1 polymorphism (-1607G) and disease severity in non-cystic fibrosis bronchiectasis in Taiwan. PLoS One. 2013 Jun 11;8(6):e66265. doi: 10.1371/journal.pone.0066265. Print 2013. PubMed PMID: 23776649; PubMed Central PMCID: PMC3679085.

6: Salerno T, Peca D, Rossi FP, Menchini L, Danhaive O, Cutrera R. Bronchiectasis and severe respiratory insufficiency associated with a new surfactant protein C mutation. Acta Paediatr. 2013 Jan;102(1):e1-2. doi: 10.1111/apa.12043. PubMed PMID: 23025826.

7: Angulo I, Vadas O, Garçon F, Banham-Hall E, Plagnol V, Leahy TR, Baxendale H, Coulter T, Curtis J, Wu C, Blake-Palmer K, Perisic O, Smyth D, Maes M, Fiddler C, Juss J, Cilliers D, Markelj G, Chandra A, Farmer G, Kielkowska A, Clark J, Kracker S, Debré M, Picard C, Pellier I, Jabado N, Morris JA, Barcenas-Morales G, Fischer A, Stephens L, Hawkins P, Barrett JC, Abinun M,

Clatworthy M, Durandy A, Doffinger R, Chilvers ER, Cant AJ, Kumararatne D, Okkenhaug K, Williams RL, Condliffe A, Nejentsev S. Phosphoinositide 3-kinase δ gene mutation predisposes to respiratory infection and airway damage. Science. 2013 Nov 15;342(6160):866-71. doi: 10.1126/science.1243292. Epub 2013 Oct 17. PubMed PMID: 24136356; PubMed Central PMCID: PMC3930011.

8. Mutesa L, Azad AK, Verhaeghe C, Segers K, Vanbellinghen JF, Ngendahayo L, Rusingiza EK, Mutwa PR, Rulisa S, Koulischer L, Cassiman JJ, Cuppens H, Bours V. Genetic analysis of Rwandan patients with cystic fibrosis-like symptoms: identification of novel cystic fibrosis transmembrane conductance regulator and epithelial sodium channel gene variants. Chest. 2009 May;135(5):1233-42. doi: 10.1378/chest.08-2246. Epub 2008 Nov 18. PubMed PMID: 19017867.

9: Papatheodorou A, Makrythanasis P, Kaliakatsos M, Dimakou A, Orfanidou D, Roussos C, Kanavakis E, Tzetis M. Development of novel microarray methodology for the study of mutations in the SERPINA1 and ADRB2 genes--their association with Obstructive Pulmonary Disease and Disseminated Bronchiectasis in Greek patients. Clin Biochem. 2010 Jan;43(1-2):43-50. doi: 10.1016/j.clinbiochem.2009.08.026. Epub 2009 Sep 10. PubMed PMID: 19747908.

10: Fajac I, Viel M, Sublemontier S, Hubert D, Bienvenu T. Could a defective epithelial sodium channel lead to bronchiectasis. Respir Res. 2008 May 28;9:46. doi: 10.1186/1465-9921-9-46. PubMed PMID: 18507830; PubMed Central PMCID: PMC2435537.

Detailed research plan

Full proposal (total max 1500 words per subdomain)	
Title *(max 150 characters)*	**GENE-STOP-HHT**

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).*

Hereditary haemorrhagic telangiectasia (HHT) affects approximately 1 in 5,000 people, and is extremely challenging to manage. The hallmarks of HHT are arteriovenous malformations (AVMs), and nasal/gastrointestinal telangiectatic vessels. Once present, vascular abnormalities are usually too numerous to treat using conventional medical approaches, and patients require life-long management to reduce risks of major haemorrhage, pregnancy-related deaths, high output cardiac failure, and strokes. Disease-causing genes such as *ENG*, *ACVRL1* and *SMAD4* modify signalling by TGF-beta superfamily members, but at least 20% of HHT patients do not have disease-causing variants in the known and emerging genes. Despite significant advances since the HHT gene assignments of ENG in 1994, and ACVRL1 in 1996, molecular therapies are only now being developed for the HHT population.

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.*

AIM 1) IDENTIFICATION OF NEW DISEASE-CAUSING GENES

Challenge 1:
Following the steps outlined in the Data Analysis plan (below), up to 100 variants are likely to remain per whole genome.
Mitigation steps: The assembled international GeCIP subdomain panel of HHT gene experts including Aldred, Bayrak-Toydemir, Berg, Bernabeu, Botella, Brusgaard, Govani, Letarte, Marchuk, McDonald, Mollet, and Shovlin will prioritise variants for validations, based on gene knowledge and in silico prediction tools (Mollet; Govani), supported by other GeCIP members.

***Challenge 2*:**
To reduce the number of incidental coinheritances for gDNA validations
Mitigation steps:

i) To facilitate capturing unrecruited, distant family members, participants will complete selected questions in the HHT Survey developed by Shovlin.
i) International GeCIP colleagues will be alerted to facilitate examination of gDNA from other HHT families in which disease genes have not been identified, particularly if the survey suggests the presence of family branches in the relevant country.

Challenge 3:
Confirming that the gDNA variant modifies relevant RNA transcript and/or protein which requires accessible patient material with sufficient expression of appropriate gene, splice variant, and/or protein. To date, HHT disease-causing gene variants have been ascribed to endothelial-expressed genes, but it is difficult to directly obtain endothelial cells from HHT patients

Mitigation steps:

i) The Shovlin and Botello groups have developed methods to evaluate known HHT genes in monocytes and activated monocytes. Shovlin group techniques permit transcript evaluations in same-day blood samples before and after *ex vivo* treatment with nonsense mediated decay (NMD) inhibitors to reveal PTC-containing transcripts.

ii) The Botello group have also optimised analyses of RNA and protein in activated monocytes of blood samples, and are also able to obtain primary cultures of endothelial cells from 50 ml of peripheral blood.

iii) Accessory techniques such as tissue biopsies are under evaluation.

iv) Pathogenicity checks using existing readouts for known HHT genes will be performed in non-patient samples:

- Biochemical characterisation of encoded proteins will be performed by generating expression vectors, and performing structural analysis of the recombinant and endogenous proteins using expertise available through the Bernabeu and other subdomain laboratories.
- Functional characterisation of the recombinant and endogenous proteins will be performed using assays extending from those employed for known HHT genes in the Aldred, Bernabeu and Botella laboratories.

Anticipated timescales:

Variant calling within 2-4 weeks of availability of WGS data; gDNA validations 8 weeks; RNA/protein evaluations: 4-6 weeks if adequate expression in blood samples and commercial antibodies available, 8-12 weeks if endothelial progenitor cells (EPCs) need to be derived; pathogenicity checks 8-16 weeks

AIM 2) IMPROVING UNDERSTANDING OF THE MECHANISMS BY WHICH AVMs DEVELOP

Challenge:

To permit rapid *in vivo* assessments, ideally within the 6 month GeCIP window.

Mitigation steps:

i) Site directed mutagenesis will utilise CRISPR/Cas technology to improve efficiency and speed, allow study of the effects of gene manipulation on the embryo/cells, and determination of gene pathway changes.

ii) Screening of candidate genes in chicken and zebrafish embryos to determine their expression patterns throughout development (Vargesson)

ii) Using their expertise in gene misexpression technologies, the Vargesson group will also determine roles/functions in embryogenesis as well as elucidate signalling pathways that the genes may be influencing. This will include gene knockdown in zebrafish using morpholino, RNAi, CRISPR/Cas or overexpression in chicken embryos using the RCAS viral system. Fli1:EGFP vascular reporter lines available in the lab will be used to examine if candidate gene misexpression alters vascular formation/stability *in vivo*.

Anticipated timescale: Within 6 months.

AIM 3) EXAMINATION OF DISEASE-CAUSING PTCs THAT HAVE THE POTENTIAL FOR RAPID TRANSLATION TO PATIENT BENEFIT

Challenge:

To rapidly screen novel PTCs for potential therapeutic correction

Mitigating steps:

i) The Aldred lab has developed a robust *in vitro* reporter assay for the known HHT genes to rapidly screen the efficiency of read-through drugs (e.g. ataluren, amlexanox). This would be focused on HHT-associated nonsense point mutations, introducing nonsense substitutions and N- and C-termini fluorescent reporter tags into expression constructs for the gene of interest, and measuring the efficiency of restoring reporter activity by different read-through using flow cytometry.

ii) Where alterations in TGF-beta signalling are identified, *in vitro* correction of these abnormalities by read-through drugs will also be analysed in patient monocytes or EPCs as appropriate.

Anticipated timescale: 3 months for vector construction of new reporter vectors; 5-14 weeks for drug screening and signalling studies depending on whether EPC isolation is required.

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 collaborate with the Paediatric, Cardiovascular, and Haematology GeCIPs, PTC Therapeutics (through Ataluren), and anticipate dovetailing with existing Pharma partners, and US Department of Defense with their specific calls for Vascular Malformations in 2015-6.

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

We anticipate that projects within the Domain will be highly subscribed to by the MSc students on the HEE-funded MSc course in Genomic Medicine, for their research project. The Imperial course (Director Lovett; Module Lead Shovlin) had its first intake of students in October 2015 and all students will be eligible to be part of the GeCiP domain. This course will provide a core resource for the Domain (and others), standardising the transfer of expertise to the next generation.

Individual modules within the Imperial MSc course will also be available as certificate and diploma degrees, short courses and for CPD for other researchers and clinicians within the GeCIP domain. Additionally, the Subdomain will provide HHT-specific courses via British Thoracic Society On Line Courses (discussions in progress through the Pulmonary Vascular Special Advisory Group), and via the Royal College of Physicians genomics initiatives (Shovlin is the Clinical Champion for the RCP on the management and delivery of Respiratory Genomics Education).

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

The GeCIP has recruited an unprecedented international scientific HHT Faculty spanning necessary areas of expertise including

- HHT gene identification [Marchuk (US), Bayrak-Toydemir (US); McDonald (US), Berg (UK), Brusgaard (DK), Pyeritz (US), and Shovlin(UK)];
- HHT clinical genetics [(Bayrak-Toydemir (US); Berg (UK); Ousager (DK); Thompson (UK)];
- HHT and generic bioinformatics [Mollet (S); Govani (UK)];
- HHT endothelial modelling [Bernabeu (E) Letarte (C), Botella (E)]; and

- HHT clinicians with non-UK families awaiting HHT gene identification [(Geisthoff (D), Kjeldsen (DK), Buscarini (I), Mager (NL)];

Many have collaborated on international HHT projects for more than 20 years, meet at least biannually at HHT meetings, and already operate an informal partnership through email support and advice to international colleagues.

To facilitate "quick wins" for the 100,000 Genomes Project, we will particularly focus *in vivo* models on chickens and zebra fish, and therapeutics on the emergent premature termination codon (PTC) read-through agents that are predicted to be suitable for 1 in 10 patients with nonsense codon substitutions in their HHT-causing gene. The subdomain therefore specifically recruited Vargesson (UK) and Aldred (US/UK), who already collaborate with subdomain members.

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.*

All patients stand to benefit through the close planned interactions between the GeCIP and GMCs. Individual GMCs have their own MDTs to process disease-causing variants called in known HHT genes (eg, West London GMC MDT, chaired by Shovlin). Our GeCIP subdomain will offer support to all GMCs recruiting HHT patients, to ensure up-to-the-minute advice on pathogenicity, which is particularly important for HHT gene missense substitutions.

Variants will also be entered in a timely fashion into existing HHT mutation databases.

To allow specific "news" items to be rapidly promulgated to the HHT communities, a GeCIP Subdomain webpage will be developed, and linked to existing HHT sites (eg NHS Choices). Links will also be placed on our patient partners' websites (HHTUK, TSHG, CureHHT), Twitter and Facebook feeds.

We look forward to meeting with members of the cross cutting domains in January, and to developing interactions to ensure incorporation of new insights into gene interactions that may facilitate prioritisation of sequence variants.

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

Patients and clinicians will benefit from this research with improved diagnostic testing, and improved targeting of healthcare resources, screening and treatment regimes to relevant individuals.

The 100,000 Genomes project also provides major opportunities to address key outstanding needs of the research and pharma communities. These include improving understanding of the mechanisms by which AVMs develop, and triggering development of new therapies to reduce the burden of disease, most immediately for nonsense read-through strategies expected to benefit ~1 in 10 patients with nonsense substitutions.

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?*

We have no commercial partners for the current proposal which is likely to generate commercially exploitable results. GeCIP members already collaborate with Pharma in development and trials of existing pharmaceuticals. Recent interactions with Pfizer, GSK, Vernalis and PTC Therapeutics suggest there will be enthusiasm for development of diagnostics and therapeutics within the HHT field.

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

McDonald et al Front Genet. 2015 Jan 26;6:1. doi: 10.3389/fgene.2015.00001
Shovlin CL.Front Genet. 2015 Apr 9;6:101 doi: 10.3389/fgene.2015.00101

Detailed research plan

Full proposal (total max 1500 words per subdomain)	
Title *(max 150 characters)*	**Genes, and AVM Development, Physiology, and Compensations in the Pulmonary Circulation (GADP-PAVMs)**

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).*

Pulmonary arteriovenous malformations (PAVMs) are abnormal vascular communications between a pulmonary artery and a pulmonary vein leading to an intrapulmonary right-to-left shunt. They affect ~50% of people with hereditary haemorrhagic telangiectasia (HHT, usually due to ENG, ACVRL1 (encoding ALK-1) or SMAD4 sequence variants), and display other familial patterns, unrelated to HHT. PAVMs require interventional and medical management to prevent ischaemic stroke, brain abscess, severe hypoxaemia, and pulmonary haemorrhage which is the main cause of maternal death in pregnancy (maternal death rate 1% per pregnancy).
Identification of new disease-causing genes and gene variants will improve understanding of pathophysiological mechanisms and consequences, leading to improved patient care.

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.*

AIM 1) IDENTIFICATION OF NEW DISEASE-CAUSING GENES

Challenge 1: Following the steps outlined in the Data Analysis plan, up to 100 variants are likely to remain per genome.

Mitigation steps:
The assembled international GeCIP subdomain panel of PAVM/HHT gene experts including Aldred, Bayrak-Toydemir, Berg, Bernabeu, Botella, Brusgaard, Govani, Letarte, Marchuk, McDonald, Mollet, and Shovlin will prioritise variants for validations, based on gene knowledge and *in silico* prediction tools (Mollet; Govani), supported by other GeCIP members.

Challenge 2*:* To reduce the number of incidental coinheritances for gDNA validations
Mitigation steps:

i) To facilitate capturing unrecruited, distant family members, participants will complete selected familial questions using Survey methodologies developed by Shovlin.
ii) International GeCIP colleagues will be alerted to facilitate examination of gDNA from other families in which disease genes have not been identified, particularly if the survey suggests the presence of family branches in the relevant country.

Challenge 3: Confirming that the gDNA variant modifies relevant RNA transcript and/or protein requires accessible patient material with sufficient expression of appropriate gene, splice variant, and/or protein. To date PAVM disease-causing gene variants have been ascribed to endothelial-expressed genes, but it is difficult to obtain pulmonary endothelial cells (EC) from PAVM patients particularly as PAVMs are usually treated by endovascular methods, not surgery.
Mitigation steps:

i) The Shovlin and Botello groups have developed methods to evaluate known PAVM genes in monocytes and activated monocytes. Shovlin group techniques permit transcript evaluations in same-day blood samples before and after *ex vivo* treatments .
ii) The Botello group have also optimised analyses of RNA and protein in activated monocytes of blood samples, and are also able to obtain primary cultures of circulating endothelial cells from 50 ml of peripheral blood.
iii) Accessory pulmonary biopsy techniques will be evaluated.
iv) Pathogenicity checks using existing readouts for known genes will be performed in non-patient samples:
    - Biochemical characterisation of encoded proteins will be performed by generating expression vectors, and performing structural analysis of the recombinant and endogenous proteins using expertise available through the Bernabeu and other subdomain laboratories.
    - Functional characterisation of the recombinant and endogenous proteins will be performed using assays extending from those employed for known HHT genes in the Aldred, Bernabeu and Botella laboratories.

Anticipated timescales:
Variant calling within 2-4 weeks of availability of WGS data; gDNA validations 8 weeks; RNA/protein evaluations: 4-6 weeks if adequate expression in blood samples and commercial antibodies available, 8-12 weeks if EPCs need to be derived; pathogeneicity checks 8-16 weeks

AIM 2) IMPROVING UNDERSTANDING OF THE MECHANISMS BY WHICH AVMS DEVELOP
Challenge: To permit rapid *in vivo* assessments, ideally within the 6 month GeCIP window.
Mitigation steps:

i) Site directed mutagenesis will utilise CRISPR/Cas technology to improve efficiency and speed, allow study of the effects of gene manipulation on the embryo/cells, and determination of gene pathway changes.

ii) Screening of candidate genes in chicken and zebrafish embryos to determine their expression patterns throughout development (Vargesson)

ii) Using their expertise in gene misexpression technologies, the Vargesson group will also determine roles/functions in embryogenesis as well as elucidate signalling pathways that the genes may be influencing. This will include gene knockdown in zebrafish using morpholino, RNAi, CRISPR/Cas or overexpression in chicken embryos using the RCAS viral system. Fli1:EGFP vascular reporter lines available in the lab will be used to examine if candidate gene misexpression alters vascular formation/stability *in vivo*.

Anticipated timescale: Within 6 months.

AIM 3) EVALUATION OF GENOMIC INFLUENCES ON PAVM COMPENSATIONS.

Our third focus will be to examine if there are differences between PAVM genotypes in the patients' abilities to compensate for their PAVMs such as i) maintaining total arterial oxygen content and/or the oxygen pulse (oxygen delivered per heart beat), and ii) avoidance of brain abscess, ischaemic stroke, and migraines despite compromised pulmonary capillary bed filtration/ function.[1] To date, such studies have been restricted to environmental factors. We will utilise the captured 100,000 genomes datasets, if necessary supplemented by formalised survey approaches to patients, to evaluate if there are any differences between the PAVM genotypes, once adjusted for age and iron status, and plan to incorporate a health economics component following the workshop on 20th January 2016.

Anticipated timescale: Within 18-24 months.

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 collaborate with the Paediatric, Cardiovascular, Haematology and Health Economics GeCIPs, and anticipate dovetailing with existing Pharma partners, and the US Department of Defense with their specific calls for Vascular Malformations in 2015-6.

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

We anticipate that projects will be highly sought by MSc students on the HEE-funded MSc course in Genomic Medicine, for their research project. The Imperial course (Director Lovett; Module Lead Shovlin) had its first intake of students in October 2015 and all students will be eligible to be part of the GeCiP domain. This course will provide a core resource for the Domain (and others), standardising the transfer of expertise to the next generation.

Individual modules within the Imperial MSc course will also be available as certificate and diploma degrees, short courses and for CPD for other researchers and clinicians within the GeCIP domain. Additionally, the Subdomain will provide PAVM-specific courses via British Thoracic Society On Line Courses (discussions in progress through the Pulmonary Vascular Special Advisory Group); the European Respiratory Society (Sept 2016), and via the Royal College of Physicians genomics initiatives (Shovlin is the Clinical Champion for the RCP on the management and delivery of Respiratory Genomics Education).

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

The GeCIP subdomain includes a broad UK and international scientific Faculty spanning necessary areas of expertise (see accompanying CVs) including

- PAVM clinicians [Chilvers (UK), Corris (UK), Hall (UK), Jackson (UK), Kiely (UK), Lomas (UK), Mager (NL), Peacock (UK), Shovlin (UK), Whyte (UK)], including core members of the British Thoracic Society PAVM Specialist Advisory Group;
- PAVM/HHT gene identification [Marchuk (US), Bayrak-Toydemir (US); McDonald (US), Berg (UK), Brusgaard (DK), Pyeritz (US), and Shovlin(UK)];
- Clinical genetics [(Bayrak-Toydemir (US); Berg (UK); Ousager (DK); Thompson (UK)];
- Bioinformatics [Mollet (S); Govani (UK)];
- Endothelial modelling [Bernabeu (E) Letarte (C), Botella (E)];
- Clinicians with non-UK PAVM families awaiting HHT gene identification [Mager (NL)].

To facilitate "quick wins" for the 100,000 Genomes Project, we will particularly focus on *in vivo* models on chickens and zebra fish. The subdomain therefore specifically recruited Vargesson (UK), who already collaborates with subdomain members.

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.*

All patients stand to benefit through the close planned interactions between the GeCIP and GMCs. Individual GMCs have their own MDTs to process disease-causing variants called in known PAVM genes (eg, West London GMC MDT, chaired by Shovlin). Our GeCIP subdomain will offer support to all GMCs recruiting PAVM patients to ensure up-to-the-minute advice on pathogenicity (particularly important for missense substitutions).

To allow specific "news" items to be rapidly promulgated to PAVM patient communities, a GeCIP Subdomain webpage will be developed, and linked to PAVM websites already in existence (eg Imperial) and planned through the British Thoracic Society (many GeCIP members are in the PAVM specialist advisory group [SAG]).

We look forward to meeting with members of the cross cutting domains in January, and to developing interactions to ensure incorporation of new insights into gene interactions that may facilitate prioritisation of sequence variants, and explore Health Economic considerations.

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

Patients and clinicians will benefit from this research with improved diagnostic testing, and improved targeting of healthcare resources, screening and treatment regimes to relevant individuals. Currently, In the absence of a gene test of exclusion, first and sometimes second degree relatives undergo regular clinical assessments, involving exposure to ionising radiation through CT scans, and costly modifications to pregnancy management.

The 100,000 Genomes project also provides a major opportunity to address key outstanding needs of the research and pharma communities by improving understanding of the mechanisms by which AVMs develop, and triggering development of new therapies to reduce the burden of disease.

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?*

We have no current commercial partners, but recent interactions with Pfizer, GSK, Vernalis and Bioxyden suggest there will be enthusiasm for development of diagnostics and therapeutics within the PAVM field.

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

1. Shovlin CL. Pulmonary arteriovenous malformations. Am J Respir Crit Care Med. 2014 Dec 1;190(11):1217-28. doi: 10.1164/rccm.201407-1254CI.

Full proposal (total max 1500 words per subdomain)

Title *(max 150 characters)*	**Functional effects of genetic variants associated with Familial Spontaneous Pneumothorax (FSP)**

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).*

Primary spontaneous pneumothorax (PSP) occurs when the lung deflates in the absence of obvious underlying lung disease. This occurs in 20 males per 100,000 population; women are affected 5 times less often. In 10% of males affected by PSP and 25% of affected women, at least one first-degree family member has also been affected and in these instances the condition is called Familial Spontaneous Pneumothorax (FSP) [1]. Rather than being a single entity, FSP comprises an increasing number of genetic disorders often with severe extra-pulmonary manifestations ranging from renal cancer [*FLCN1* mutation in Birt-Hogg-Dubé (BHD) syndrome] to aortic rupture [mutations affecting extracellular matrix formation including *COL3A1*, *FBN1*, or *TGFBR2 etc.* in vascular Ehlers Danlos, Marfan, and Loeys Dietz syndromes respectively] [2]. Because pneumothorax frequently occurs early in life, a definitive genetic diagnosis permits targeted interventions that can avoid or delay morbidity and to prolong life [2]. For example, in BHD, pneumothorax tends to precede the development of renal cell carcinoma by two decades. This enables effective cancer screening programmes, since renal cell carcinoma can be cured if diagnosed before reaching 3 cm in size. In Marfan and Loeys Dietz syndromes, randomised controlled trails have proved that pharmacological intervention reduces the rate of arterial rupture, an important cause of morbidity and mortality in these disorders.

Despite the large heritable component in FSP, up to half of cases remain unclassifiable following thorough investigation in a specialist Pneumothorax Genetics service. By recruiting individuals with FSP who fulfil the following criteria, it is hoped that many of the remaining causative mutations will be identified: (i) primary spontaneous pneumothorax, (ii) one or more affected relatives, and (iii) prior testing for *FLCN* or *FBN1* mutations if suggested by the clinical and

radiological findings. In this way, we hope eventually to develop inexpensive, targeted gene panels for routine screening of FSP families.

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.*

Initial studies will aim to identify new genetic mutations associated with FSP and to determine whether these are associated with novel phenotypes (clinical, radiological and where appropriate pathological) thus allowing the description of new syndromes.
Where genes are identified in pathways related to known pneumothorax syndromes, for example the modulation of mTOR signalling as occurs in BHD and tuberous sclerosis, deep phenotyping will be performed as indicated, e.g. renal imaging. We will develop biochemical and cell-based assays of the function of the affected protein to help classify variants of uncertain significance, as we have done previously [3].
When mutations are identified in genes plausibly implicated in the integrity of the extracellular matrix (and therefore the vasculature) as occurs in a substantial number of FSP syndromes, we will perform long-term studies (5+ years) to monitor cardiovascular physiology and changes of aortic root dimensions. This will be facilitated by existing close links with the Cambridge Inherited Cardiovascular Disease service. Fibroblasts will be obtained and used to generate induced pluripotent stem cells, which will subsequently be differentiated towards appropriate smooth muscle subtypes for detailed phenotypic analysis. Expression of the gene produces will be assessed in archived tissues from the Papworth Biobank.
Having defined the effects of genetic variants on disease phenotypes and cellular function, functional pharmacogenomic studies will be performed. Cells expressing mutant or wild type genetic variants (either disease-associated or engineered) will be cultured in the presence or absence of appropriate drug libraries, e.g. modulators of the mTOR pathway.
The challenges facing this area of research are both financial and practical. The opportunity to take the new genomic data and apply it to the described research projects on a large scale will greatly enhance our research output and we would aim to submit applications to the relevant funding bodies to fund this research
Practical challenges include the power to generate meaningful observations from the initial FSP cohort. These will be mitigated by international collaboration with other groups. We already benefit from close links with the Raby (pulmonary genetics) and Henske (mTOR-related pulmonary disease) groups at Harvard [2].

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 collaborate with the Musculoskeletal, Cardiovascular, and Renal Cell Cancer GeCIPs.

Training. *Describe the planned involvement of trainees in the research and any specific training that will form part of your plan.*
In recognition of the transformation of medicine and clinical care that is occurring as a consequence of advances in 'omic' technologies and the expanding educational need in this area, Imperial College is one of the successful bidders in providing an HEE-funded MSc course in Genomic Medicine. Members of the proposed Domain are key members including the Course Director (Lovett), and Shovlin who is the national Clinical Champion for the RCP on the management and delivery of Respiratory Genomics Education. The course aims to provide the

skills required for the analysis and interpretation of genome-scale data. It is taught in a modular format to allow certificate, diploma or full MSc degrees to be obtained, and allows for either full or part time attendance. It is taught by leaders in genomic medicine across the entire faculty of medicine at Imperial College. For the full degree course, a key component is the research project that can be laboratory based, purely analytical, or a combination. We anticipate that projects within the Domain will be highly subscribed to by the MSc students. This course will provide a core resource for the Domain (and others), standardising the transfer of expertise to the next generation. Individual modules within the course will also be available as short courses and for CPD for other researchers and clinicians within the proposed domain. The MSc complements existing analytically themed MScs already offered by the Domain partners including the IC MSc in Bioinformatics and Theoretical Systems Biology and the UCL MSc in Computational and Genomic Medicine. Transfer of trainees between Domain centres will be encouraged and new funding sought to catalyse this.

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

One of us (Marciniak) runs the Cambridge Pneumothorax Clinic, is currently a specialist service within the NHS and comprises services within Cambridge University Hospitals NHS Trust and Papworth Hospital Trust [2, 3]. This takes advantage of existing relevant specialist services including the Pneumothorax Genetics Multi-Disciplinary Team (MDT), the Inherited Cardiovascular Disease service (run by Dr Rosemary Rusk) and the vascular Ehlers Danlos / Loeys Dietz syndrome MDT. Prof Johnson runs the National Centre of LAM in Nottingham University Hospitals NHS Trust [4].

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.*

Familial Spontaneous Pneumothorax genetic data will become part of the Pneumothorax Genetics MDT. We envisage the incorporation of these data into regular MDT discussions that would be interpreted with a clinical geneticist alongside clinical, radiological and pathologic data. Follow-up consultations with the individual would take place with a defined plan for family screening in addition to the individual's clinical care. Ideally this would occur jointly with a respiratory clinician and clinical geneticist.

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

Multiple beneficiaries are likely to result from this work:
Patients and families. We are seeing an increasing number of referrals of patients with family histories of FSP. Currently, only half receive a genetic diagnosis conferring prognostic information and potential family screening.
Clinicians Treating Patients with FSP. It is anticipated that the data generated from these studies will aid both screening and treatment algorithms by determining the genetic risks associated with the different genetic variants as well as the effect on disease prognosis and possible pharmacogenomics interactions.
Academic researchers. Knowledge of the prevalence of currently known causes of FSP and understanding the molecular genetic variants underlying as yet unknown causes will inform

research into new mechanisms of lung integrity. This has implications for both FSP and idiopathic PSP. Since many pneumothorax syndromes involve alterations in cardiovascular and long bone growth, it is likely the results from the FSP studies will have broader scientific relevance.

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?*

Identification of novel drug targets is likely to improve long term cardiovascular and cancer outcomes. Existing links with, among others, Astra Zeneca, will help facilitate this.

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

[1] Abolnik IZ, Lossos IS, Zlotogora J, Brauer R (1991) On the inheritance of primary spontaneous pneumothorax. Am J Med Genet 40: 155-158

[2] Scott RM, Henske EP, Raby B, Boone PM, Rusk RA, Marciniak SJ (2018) Familial pneumothorax: towards precision medicine. Thorax 73: 270-276

[3] Chambers JE, Dalton LE, Subramanian DN, et al. (2015) Spontaneous pneumothorax can be associated with TGFBR2 mutation. Eur Respir J 46: 1832-1835

[4] Johnson SR, Cordier JF, Lazor R, et al. (2010) European Respiratory Society guidelines for the diagnosis and management of lymphangioleiomyomatosis. Eur Respir J 35: 14-26

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)*

(1) Our primary data analyses will be of phenotypic and genotypic data from patients recruited within the Respiratory subdomains: Primary Ciliary Dyskinesia (PCD), ii) Familial Interstitial Lung Disease (ILD), iii) Non-CF Bronchiectasis, iv) Pulmonary Arteriovenous Malformations (PAVM) v) Hereditary haemorrhagic telangiectasia (HHT) and vi) Familial Pneumothorax. Our Subdomain members span all of the GMCs, and we envisage notification of all samples progressing through the GeL pipeline with these conditions. If additional samples from similar phenotypic spectra within the 100,000 Genomes Project samples are considered important, we will liaise with the relevant GeCIP at that stage.

(2) We will request access to filtered variant lists for review of the material provided to GMCs, and for cross reference to our own data analyses.

(3) While we have the technical expertise to analyse FASTQ and BAM files, the analyses by the Respiratory GeCIP fall more within variant prioritisation and functional evaluations rather than read mapping, and we anticipate that for the purposes of our analyses (see below) we will use VCF files. This will in turn limit the need for very high computational resources.

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

GeL's VCF outputs will be searched for pathogenic variants. Based on comparable sequence projects, we expect each whole genome to provide >100 premature termination codons (PTCs), and indels overlapping coding sequence, ~3 million single nucleotide substitutions, and ~10,000 missense substitutions in open reading frames (ORFs).

gDNAs with PTC and indel variants in known disease-causing genes should not have been entered into the project. If identified, these would be immediately prioritised for validation. Otherwise, customised scripts will:

- Remove all variants with allelic frequencies exceeding the maximum estimated frequency for a new disease-causing sequence variant in that disease;
- Remove any variants that do not segregate with disease;
- Select coding and deep intronic sequence variants in known disease causing genes or interacting genomic regions
- Select rare sequence variants present in multiple families, and any PTC, indel or splice site substitution segregating in the pedigree.

We will initially focus on coding region variants. These will be analysed by ANNOVAR and other tools that interrogate multiple software packages that predict mutational effects before import into a relational database. For copy number variation we currently run CNVnator and RDXplorer, for structural variant identification, Breakdancer and CLEVER for ORF indels, and EXPASY to identify PTC positions. Final manual analysis of the combined output is frequently necessary. Initially, GeCIP subdomain expertise will be utilised, but in the longer term we expect to be able to combine gene network/pathway prediction with variant calling, to further prioritise candidates in a more streamlined and rapid manner (for example via VARELECT). We will collaboratively interact with GeL bioinformaticians to ensure that we do not provide redundant analytical methods.

gDNA sequence variants identified by WGS will be validated by conventional methods in gDNA. The precise methods will be determined according to the number of different variants/genes to be evaluated.

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)*

Primary Ciliary Dyskinesia:

- Situs abnormalities – Situs inversus, situs ambiguous [heterotaxy]
- Neonatal respiratory distress in otherwise healthy term infant
- Persistent rhinitis from early days of life, persistent through life, even when well
- Sinusitis – from mid childhood through adult life
- Chronic middle ear disease, hearing impairment, otorrhoea after perforation or grommet insertion
- Chronic productive cough, often starting in early days/weeks of life
- Bronchiectasis [favouring lower lobes]
- Infertility/subfertility – immotile sperm or ineffective ciliary motion in fallopian tubes
- Associated syndromic ciliopathy features – polydactyly, visual defects, hydrocephalus, congenital heart disease [such as isomerism], renal or hepatic cystic disease

Aggressive Bronchiectasis:

- Sweat Sodium levels,
- CT evidence of multilobar disease, tracheobronchomegaly or recurrent/ multispecies non-tubercular mycobacterial infection

Familal Pulmonary Fibrosis:

- High resolution computerised tomography (HRCT) reports and images
- Pulmonary function indices (Forced vital capacity (FVC); diffusing capacity of carbon monoxide (DLCO) ideally at several time points to evaluate disease progression

Hereditary Haemorrhagic Telangiectasia

- Radiological reports (+/- imaging) for arteriovenous malformations
- Nosebleed severity (ideally by the epistaxis severity score)
- Haemoglobin and iron indices, ideally at several time-points to evaluate disease progression

Pulmonary arteriovenous malformations

- Computerised tomography (CT) and pulmonary angiography reports and images
- Oxygenation indices (SaO_2, haemoglobin, iron indices) ideally at several time points to evaluate disease progression

Familial Spontaneous Pneumothorax

- High resolution computerised tomography (HRCT) reports and images
- Echocardiography: ascending aortic dimensions, valvular function
- Clinical: ocular [ectopia lentis], musculoskeletal [dental crowding, high palate, arachnodactyly, disproportionate tall stature, reduced upper to lower segment ratio, scoliosis, hypermobility, chest wall deformity, hind foot deformity], skin [striae distensae, soft thin skin, cutis laxa, atypical scarring], family history of neoplastias [adrenal cortex, angiofibromas, any renal tumour type].

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)*

We note that the pipeline includes validations of BAM and VCF files, and calculations of read depth, Mendelian errors, inbreeding estimates, IBD estimation and ancestry. All will be used in the prioritisation of variants for validations. We also note the planned move to GRCH38 later this month (January 2016). We do not intend to repeat alignments and therefore do not have additional requirements here. However, we would be keen to use CellBase to supplement our in-house variant calls.

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)*

We plan conventional analyses and therefore no specific new tools are required. We are particularly excited by the integration with phenotypic interrogation tools through OpenCGA, and will request access to the APIs for the later aspects of the proposal.

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)*

As outlined above, the specific datasets are the phenotypic entries post OpenClinic (ideally through OpenCGA), and VCF files.

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

We have the infrastructure and trained personnel in place to cope with data analysis and interpretation. We will consider investing in an additional server for data storage and analysis.

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)*

For the current proposals, the 'omics' samples are not specifically required, but we anticipate that hypotheses generated in the course of the proposed analyses may be tested, particularly in RNA samples, in later proposals.

Data access and security	
GeCIP domain name	Respiratory
Project title *(max 150 characters)*	**Hypothesis-driven research and development to establish the functional role of DNA sequence variants in respiratory 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).*

✓ *Clinical care*
✓ *Clinical trials feasibility*
✓ *Deeper phenotyping*
✓ *Education and training of health and public health professionals*
✓ *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

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

As per discussions with Genomics England we understand this will be made available at a later date to sign.

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)