GeCIP Detailed Research Plan Form

August 2015

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) – (1) QMMLFG_GeCIP_structure_approved_by_GEL_27052015.docx

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

Application Summary	
GeCIP domain name	**Quantitative Methods, Machine Learning, and Functional Genomics**
Project title *(max 150 characters)*	**Advancing Genome Interpretation in Genomics England through Quantitative Methods, Machine Learning, and Functional Genomics**

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

The GeCIP will address many of the major challenges in genome analysis and interpretation via: methods development; generation, analysis and interpretation of functional genomics data; analysis and interpretation of multi-omic data; training; provision of tools for improving genome analysis for the research and clinical community and; application of appropriate methods in partnership with GeCIPs and GMCs.

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)*

In the field of genomics, technology is outstripping our capacity to analyse the data. The 100,000 Genomes Project raises data analysis and interpretation questions that have not yet been addressed. These are the main focuses of the Quantitative Methods, Machine Learning, and Functional Genomics GeCIP. We will develop and use statistical methods to improve our understanding of the human genome, and to better understand the relationship between changes in the genome and human disease. The GeCIP concentrates complementary expertise from various disciplines and will have a strong training component in order to train the next generation of genomics analysis scientists.

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 Quantitative Methods, Machine Learning, and Functional Genomics GeCIP brings together four subdomains focused on the analysis of genomics data across two main strands: sequence variation and functional genomics. We will deploy statistical method development to address both broad questions and will include application of these methods to the generated sequence and functional genomic data. The Statistical Genomics and Genetic Epidemiology (SGGE) and Statistical Machine Learning (SML) subdomains bring together statistical, computational and genomics expertise in a bid to develop efficient and powerful methods for handling, analysing and interpreting big data. The Epigenomics (E) and Transcriptomics and RNA Splicing (TRS) subdomains aim to generate additional functional genomics data, which will serve as an excellent substrate for further integrative method development and analysis with respect to disease. Together the four subdomains will work towards improving our understanding of genetic and epigenetic variation, their functional consequences, and their relationship with human phenotypes. Expected outcomes include a better understanding of disease aetiopathology; efficient tools for data handling, analysis and interpretation; and training of the next generation of genomics scientists who are currently in short supply.

Expected start date	1st February 2016
Expected end date	31st January 2021

Lead Applicant(s)	
Name	Professor Martin Tobin
Post	Professor of Genetic Epidemiology and Public Health
Department	Department of Health Sciences
Institution	University of Leicester
Current commercial links	no commercial affiliation; collaborations include Pfizer, GSK.

Administrative Support	
Name	Miss Anna I.J. Harding, Assistant Registrar
Email	aijh1@le.ac.uk
Telephone	0116 252 2974

Subdomain leads		
Name	**Subdomain**	**Institution**
Prof. Ele Zeggini (co-lead)	Statistical Genomics and Genetic Epidemiology (SGGE)	Wellcome Trust Sanger Institute
Prof. Martin Tobin (co-lead)	Statistical Genomics and Genetic Epidemiology (SGGE)	University of Leicester
Prof. Chris Holmes (co-lead)	Statistical Machine Learning (SML)	University of Oxford
Prof. Chris Yau (co-lead)	Statistical Machine Learning (SML)	University of Oxford
Prof. Vardhman Rakyan (co-lead)	Epigenomics (E)	Queen Mary University of London
Prof. Stephan Beck (co-lead)	Epigenomics (E)	University College London
Prof. Diana Baralle (co-lead)	Transcriptomics and RNA Splicing (TRS)	University of Southampton
Prof. Ian Eperon (co-lead)	Transcriptomics and RNA Splicing (TRS)	University of Leicester

Detailed research plan

Full proposal (total max 1500 words per subdomain)	
Title *(max 150 characters)*	Advancing Genome Interpretation in Genomics England through Quantitative Methods, Machine Learning, and Functional Genomics

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).*

Technologies for generating genome sequencing have evolved rapidly. Methods for analysis of the data have been developed usually in a research context and the improved methods to analyse and interpret genomic data in a clinical context are urgently required. Specific examples of the areas of development required are described in the Research Plan below. It is not only the development of the methods that is required but also dissemination of methods and tools, the integration of the new insights that methodological improvements can bring to clinical interpretation and,

crucially, training of clinicians, genome analysts and biomedical researchers that is required to realise the full benefits of the Genomics England 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.*

The Quantitative Methods, Machine Learning, and Functional Genomics GeCIP spans 4 subdomains which together address many of the major challenges in genome analysis and interpretation: Statistical Genomics and Genetic Epidemiology (SGGE), Statistical Machine Learning (SML), Epigenomics (E) and Transcriptomics and RNA Splicing (TRS). The first two particularly focus on methods development, whilst the latter two have a key role in generation of new data to enhance Genomics England (GEL). All will contribute to improvement of interpretation of genomic and multi-omic data. As a cross-cutting domain, the data interpretation will not be restricted to any particular disease area, nor necessarily to established tools for genomic analyses in order to facilitate the development of new methods. Crucially the methods aim to relate genetic variation to phenotype; should any areas of application overlap with those of disease focused GeCIP(s), we should be pleased to work collaboratively towards common goals. Details of the structure, terms of reference and an overview of cross-subdomain activities are provided in the attached Appendix. Timelines for outcomes are referred to in the Beneficiaries section.

<u>The SGGE subdomain</u> comprises 5 main streams of activity

1) Methods for relating Genotype to Disease;
2) Methods for Functional Genomics;
3) Implementation;
4) Dissemination
5) Training

The areas of methods development include:

- Integrating resequencing data across multiple individuals (few to thousands) and advancing methods for detecting and accounting for, sources of bias and uncertainty in resequencing data, and confounding from population substructure of rare variants within England.
- Methods to combine information from different “causal” alleles at a given locus;
- Approaches to improve on annotation of called variants, especially those that will exploit the availability of whole-genome sequences (annotation of combined consequences of multiple variants in an individual), and related samples, such as trios and paired germline-somatic samples;
- Areas specific to cancer will include:
    - joint calling of tumour & germline sequences (including assessing copy number variation);
    - integration of approaches to identify clinically relevant tumour subtypes by prioritising candidate driver mutations and relating germline DNA variation to tumour mutation spectra;
    - understanding the role of telomere length in cancer susceptibility;
- Statistical methods to explore pleiotropic effects and methods to inform phenomic study design utilising electronic health care records (including validation of relevant data fields) to prioritise studies in Genomics England and in external datasets;
- Methods to analyse genome-wide resequencing data together with epigenetic, transcriptomic and/or proteomic data and phenotype data;
- Methods of functional and pathway-based modelling of variant data for causal modelling of

diseases;

Implementation: We will implement novel methods, in partnership with other GeCIPs and leveraging external additional data as appropriate, to ensure rapid utility and exemplars of innovative methods, to provide supervised training opportunities and to contribute to transformed patient care through:

- insights pertinent to the role of specific loci by integrating evidence from multiple rare, low frequency and common variants within a specific locus;
- understanding pleiotropic effects of variants;
- patient clustering based upon integrative analyses of health records, biomarkers, functional clustering of variants and / or functional genomics data and quantitative phenotyping;
- target validation, drug repurposing and stratified medicine using genomic and functional patient data;
- in collaboration with cancer focused GeCIPs we will undertake functional interpretation of somatic variants using publicly available and collaborative omic datasets and functional modelling of mutational processes. We will characterise tumour subtypes and assess the impact of germline variants on these.

<u>The SMMLE subdomain</u> aims to:

1. facilitate the open exchange and sharing of statistical ideas and methods across GeL and ensure best practice in clinical interpretation analysis.
2. work with disease groups and provide access to world-leading experts for the development of novel statistical and computational data analysis algorithms.
3. provide training and support for junior researchers and developers to leave a legacy of world-leading expertise in statistical machine learning in genomics-based healthcare.

The key areas of methods development include:

- *Linking genomics to clinical phenotypes.* We will develop machine learning based methods to infer associations between genetic traits and complex multivariate clinical phenotypes. This partnership will allow cooperation and sharing of knowledge to identify the best approaches to be used for GeL.
- *Statistical approaches for causal variant prioritisation, identification and functional validation.* We will develop statistical methods to identify and prioritise potential disease-causing or treatment targets and work with experimental and clinical groups to elucidate functionality. Members have particular interests in rare variants where novel machine learning approaches that integrate data from many individuals and multiple 'omics are likely to be beneficial.
- *Modelling heterogeneity and genomic instability in cancer.* Members have expertise in cancer sequence analysis from prior research projects including copy number analysis, deconvolution of tumor heterogeneity and evolutionary analyses. Applying similar methodology to the analysis of 100KGP cancer genome data, we will develop novel approaches that make use of the scale of data to be generated.
- *Core statistical and computational methodologies for dealing with "Big Data".* Our application-driven activities will lead to fundamental mathematical-statistical challenges which require development of novel robust and scalable machine learning algorithms.

Following progress on the above, additional future research activity will include: (i) *Statistical machine learning for precision medicine to* underpin state-of-the-art clinical decision making support systems which iteratively learn from new data and supervised clinical guidance and; (ii) *modelling tumour evolution and progression through the integrated analysis of spatiotemporal tumour samples and liquid biopsies*.

<u>The epigenomic subdomain</u> will focus, in partnership with GEL, on generation of blood DNA

methylomic datasets and on the subsequent computational and bioinformatic analysis of these data to identify epigenetic variation of clinical relevance. The activity will focus on the following questions:
(i) Are there disease specific profiles in blood and can they be used for biomarker discovery?
(ii) Are there genetic-epigenetic interactions that could be clinically informative?
(iii) What is the relationship between expression and epigenetic marks?
(iv) How clinically informative is the analysis of epigenetic variation at retroelements?
(v) What are the commonalities in epigenetic profiles among different related diseases and can this information be used for clinical benefit?
(vi) How do patient derived epigenomic profiles compare with other 'healthy', richly phenotyped populations contributed by executive and associated members and can such comparisons uncover antecedents of disease?
Building on the initial analyses, we envisage additional research questions that will take advantage of the power of multi-omic approaches in GEL data, and leverage evidence from parallel functional studies in cell or model organisms (e.g. epigenomic engineering approaches).

The TRA subdomain will provide service and research activities. The service activities involve (i) finding and interpretation of transcriptome variants associated with disease and (ii) identifying where possible how they affect gene function, whether through cis- or trans-acting effects on gene expression. The subdomain will also optimise the processes for assessing variants with regards to effects on the transcriptome through developing laboratory testing, bioinformatics or transcriptomics. This will provide a resource for future discovery research aimed at better understanding and prediction of the effects of mutations. The above will be achieved through 3 main areas of activity:

- *Clinical*: variant interpretation with regards to splicing, working closely with the Validation and Feedback domain and clinical domains to optimise return of findings;
- *Translational clinical research* - including development of processes, algorithms and tools for improved prediction of phenotypic consequences of intronic and exonic variants that alter RNA processing, with follow-up studies of key 'suspicious' variants of unknown significance (abbreviated VUS) splicing pharmacogenomics and, importantly, transcriptomics;
- *Discovery research* – including the identification of system-wide effects, investigations of the mechanisms and predictability of deep intronic mutations, identification of splicing events affected by drugs and identification of common splicing switches that mediate disease which could be targeted therapeutically.

A system of bioinformatics analyses (that combine investigation of splice site, enhancer and silencer programs), 'in vivo' wet lab analysis of RNA or minigene analyses will be employed to assess relevance of sequence variants found. We envisage members of this subdomain working with those developing the pipelines to optimise return of pathogenic variants that could be affecting the splicing process.

We will develop clinically useful tools and knowledge:
1) Development of processes, algorithms and tools for improved prediction of phenotypic consequences of intronic and exonic variants that alter RNA processing. This will include development of splicing bioinformatics programs and will take into account deviations from splicing patterns expected in particular tissues.
2) Follow-up studies of key 'suspicious' variants of unknown significance (VUS) to establish pathogenicity, working with appropriate international collaborations.
3) Long-term follow-up of mutation carriers to better establish phenotypic risks, spectrum and penetrance, working with appropriate international collaborations.
4) Transcriptome sequences needed to identify changes in gene expression in disease; predictive potential at all levels, including possible reasons for changes in splicing of RNA Binding Proteins or

chromatin modifiers affected.
5) Identification of splicing changes linked to drug treatments and predictions of side effects.

We will work closely with Genomics England in pilots to evaluate protocols, platforms and analysis for epigenomic and transcriptomics.

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 disease-focused GeCIPs and GMCs. In all our collaborations we will employ best practices regarding sharing of tools, expertise and knowledge of mutations, consistent with principles and established practice of Genomics England as well as, for example, the Global Alliance for Genomics and Health (GA4GH). Our cross cutting GeCIP has memberships in several domains and GMCs, through which we will provide access to our member capabilities. We aim to make our GeCIP developments available to all GeCIPs and to extend our member expertise to as many GeCIPs as possible. Our colaborations will extend beyond GeCIPs to the broader academic, healthcare and biomedical research communities, including industry collaborators (see Beneficiaries section).

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

We recognise rich opportunities for training within our proposed domains. It is anticipated that many clinical trainees will engage in the research we have detailed and that PhD students in the members' laboratories will be also involved. We have access to experts career development fellowship panels for national funding bodies (Holmes, Rattray, Westhead, Tobin, Balding) and therefore have the insight, experience and capability to develop extremely strong research training programmes within the Partnership within the following framework:

We will work with Genomics England and other partners e.g. ELIXIR, to develop proposals to discuss with potential funders with an aim to provide new training opportunities from the list below:

- National workshops (similar to statistical framework "http://www.ilike.org.uk/)funded by EPSRC
- Short courses aimed at NHS research cadres (these will, where appropriate be blended with modules in):
- MSc programmes to include MSc in Genome Medicine (already initiated at 7 universities across UK)
- PhD studentships and Fellowships for postdoctoral researchers and clinical scientists(Projects for PhD studentships and fellows will be aligned to the priority areas listed from the areas above.

mobility fellowship: we will also organise "mobility" fellowships in which trainees spend an allotted time in partnered GeCIP (rare disease or alternative "omics" domain) to crossfertilise ideas . Conversely, analytical researchers initially embedded within disease domains can exploit the mobility fellowships to "take time out" to learn new machine learning/analytical approaches within the QMMLFG domains

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

The QMMLFG domain brings together world-leading experts across the 4 integrated sub-domains of statistical genomics and genetic epidemiology, statistical machine learning, transcriptomics and RNA splicing and epigenomics. The experts will facilitate data generation, development and application of methodologies, close working with other GeCIPs and GMCs and training. Further details are provided in the Detailed Research Plan above and in the supporting documents (see "QMMLFG_GeCIP_structure_approved_by_GEL_27052015.docx").

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.*

An important functional step in the genomics process will be to establish pathogenicity or clinical relevance of sequence variants found for maximal patient benefit. The quantitative methods, machine learning and functional genomics (QMMLFG) GeCIP plans to apply its skill set to address variant interpretation in the following ways:

- We will work with disease groups providing access to world-leading experts in genomic statistics and machine learning, quantitative methods, data analysis algorithms, epigenomics, splicing and transcriptomics. We will establish a clinical subgroup within the GeCIP that can not only respond to queries from other GeCIPs but feedback developments to improve interpretation.

- Each subdomain within this GeCIP will collaborate and work on developing systems for improved clinical interpretation, including:

1. Machine learning based tools to infer associations between genetic traits and complex multivariate clinical phenotypes.
2. Statistical methods to identify and prioritise potential disease causing variants.
3. Approaches to improve on annotation of called variants, especially those that will exploit the availability of whole-genome sequences (annotation of combined consequences of multiple variants in an individual), and related samples, such as trios and paired germline-somatic samples.
4. Patient clustering based upon integrative analyses of health records, biomarkers, functional clustering of variants and / or functional genomics data and quantitative phenotyping;
5. Target validation, drug repurposing and stratified medicine using genomic and functional patient data;
6. In collaboration with cancer focused GeCIPs we will undertake functional interpretation of somatic variants using publicly available and collaborative omic datasets and functional modelling of mutational processes. We will characterise tumour subtypes and assess the impact of germline variants on these.
7. Consider and analyse variants that cause disease through aberrant splicing. It is well known that without considering splicing a large number of missense, intronic and even 'silent' variants will not be classed as pathogenic, and so overlooked. A system of bioinformatics analyses (using up to date programs that combine investigation of splice site, enhancer and silencer programs), 'in vivo' wet lab analysis of RNA or minigene analyses will be employed to assess relevance of sequence variants found. In addition we envisage members of this domain working with those developing

the pipelines to optimise return of pathogenic variants that could be affecting the splicing process.

Where new insights are pertinent to a "genetic re-diagnosis", we will communicate these to Genomics England via the disease focused GeCIPs and Validation and Feedback domain, where we will provide representation. We will also ensure timely feedback of guidance relevant to potential improvements in the pipeline for rapid WGS analysis and reporting.

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

Timelines are specified in years (Y1-Y4, Year 1 to Year 4, where Year 1 commences on the date that the GECIP first access data)

NHS Patients: (i) some of the work of the domain may inform the pipelines utilised by GEL for initial reporting (Y1 and beyond) or the interpretation of the reports by the GMCs, disease-focused GeCIPs or Validation and Feedback Domain; (ii) application of the new methods and effective data integration will in some cases inform a "genetic re-diagnosis" (likely Y2 and beyond); we will communicate these via the Validation and Feedback Domain together with disease focused GeCIPs/GMCs as appropriate; (iii) research use of the data may inform understanding of the mechanisms and prediction of disease (and potentially response to treatment), informing the development of new approaches to prevention and treatment of disease for NHS patients (Y3 and beyond). Some of the above improvements could lead to improved models of provision of NHS care e.g. for clinical genetics services (Y1 and beyond) or approaches to stratified medicine across a range of NHS providers in primary and secondary care (likely Y4 and beyond, with some potential examples from Y3).

The biomedical research community, including clinical researchers, will benefit from improved methods, tools and algorithms for data analysis (likely Y2 and beyond, although advice available via GeCIP experts and completion of tools already in development could benefit the research community as early as Y1).

The training our members will provide will benefit the clinical and research community in both the near (from Y1) and distant future (e.g. Y5 and beyond), this is of key importance in addressing the skills shortage in these areas and will bring economic benefits.

The partnership of GeCIP members with industrial partners and the biomedical advances made feasible via the knowledge generated from insightful analysis using the methods developed in, and expertise of, the GeCIP will benefit the UK economy.

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?*

Technologies under development are in some cases of strong interest to industry. Where discovery directed research is likely to be performed that has reward based milestones, we will seek to establish the impact of access to Genome England data and develop agreements accordingly. All existing partners will be declared at initiation.

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

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)*

Our cross-cutting domain would request eventual access to the entire cohort to maximise our ability to develop and apply novel methodologies as well as providing expertise in areas of transcriptome and epigenomics analysis. Individual members of our GeCIP belong to a variety of disease-focused domains, however, if data access is restricted to those domains, this would severely limit the potential long-term contribution of our group.

The variety of work undertaken by members of our group will require access to both low-level (BAMs) and high-level (filtered variant lists) data. The exact depend on the nature of the methodological or analytical work being undertaken. In order to avoid redundancy, we would like to engage with GeCIPs and industrial partners to devise a set of core genomic analysis pipelines to minimise repetition of effort. Phenotype data are of particular interest to many of our group with a focus on genotype-phenotype relationships but also modelling deep phenotype information alone.

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

The group aims to develop novel approaches that improve upon existing methods with respect to three main areas: (i) variant calling and genome annotation, (ii) patient stratification and molecularly-defined phenotypes and (iii) 'omics-phenotype data integration. The principle novelty will be the opportunity to exploit the scale of the collection of 'omics (genomes initially, then transcriptomes and methylomes) and phenotype data that will be available. This will allow us to build and improve empirical models of technical noise, variation within- and between individuals and to build robust predictive algorithms.

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)*

The QMMLFG GeCIP is not focused on a specific disease. Across the whole domain, we expect to make use of a broad range of phenotype data and evaluate methods to make use of extensive phenotype data collected during recruitment of Genomics England participants. In addition, our GeCIP also aims to make use of linked routine health care records from consenting participants as it becomes available.

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 would like to work with other GeCIPs, GEL and industry partners to define common, core analysis pipelines in four interrelated areas:

(i) a cancer specific analysis pipeline, that is able to handle multiple biopsies per patient or longitudinal samples,
(ii) a sequence variant functional annotation pipeline
(iii) a rare variants pipeline, to specifically detect rare alleles or de novo events by sharing information across genomes, that might not ordinarily pass standard filtering and QC metrics,
(iv) establishing protocols for future transcriptomics/methylomics analysis pipelines.

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)*

Our main request is an environment that **enables** methods development and the deployment of novel software and algorithms. This will require access to standard software deployment software (e.g. compilers, SDKs) and also high-level programming environments (e.g. R, MATLAB, Python). However, in order to support the variety of approaches that might be utilised, we believe that the ability to run virtual machines (VMs) or the deployment of software containers (e.g. Docker) would best provide the versatility and flexibility required. This will allow our developers to devise and maintain their own working environments reducing the effort by GeL to support the needs of individual researchers and groups and facilitate the improved sharing of software and algorithms across GeL.

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)*

Cancer: TCGA/ICGC data?
1,000/10,000 Genomes?
UK Biobank?

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

High-demand activities:

(i) Re-analysis of genomes,
(ii) Cross-sample variant calling or comparisons,
(iii) High-dimensional numerical computations underpinning statistical methods (e.g. massive matrix factorization operations),
(iv) RNAseq and methylation data analysis.
(v) Analysis of high-dimensional –omics and miulti-omics data, for example to develop approaches to infer associations between genetic traits and complex multivariate clinical phenotypes.

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)*

The focus of the The GeCIP will address many of the major challenges in genome analysis and interpretation via: methods development; generation, analysis and interpretation of functional genomics data; analysis and interpretation of multi-omic data; training; provision of tools for improving genome analysis for the research and clinical community and; application of appropriate methods in partnership with other GeCIPs and with GMCs.

Data access and security	
GeCIP domain name	**Quantitative Methods, Machine Learning, and Functional Genomics**
Project title *(max 150 characters)*	Advancing Genome Interpretation in Genomics England through Quantitative Methods, Machine Learning, and Functional Genomics

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 signed the Information Governance Declaration form provided by Genomics England and will submit by e-mail signed copies to Genomics England alongside this research plan.

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

Other attachments

Attach other documents in support of your application here including:

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