GeCIP Detailed Research Plan Form

Background

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

The aims of the partnerships are:

1. To optimise:

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

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

3. To provide a sustainable thriving training environment.

The initial wave of GeCIP domains was announced in June 2015 following a first round of applications with expressions of interest in January 2015. In April 2016 we invited the inaugurated Cancer GeCIP domains to develop research plans for 'Gear 2' working closely with Genomics England. Within the Cancer Main Programme, the 'Gear 2' phase of the project refers to recruitment of specific cohorts of patients, inclusion of biopsy tissue (diagnostic/recurrence) and ctDNA in selected cohorts and the initiation of clinical trials in early stage (adjuvant/consolidation) setting. 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 required to be submitted 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.

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 as relevant (optional).

Additional members can apply to join the GeCIP domain by completing the form on our website found here: http://www.genomicsengland.co.uk/join-a-gecip-domain/.

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

Application Summary	
GeCIP domain name	**Haematological malignancies**
Project title *(max 150 characters)*	**Haematological malignancy research in the 100,000 Genomes Project**

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

GeCIP Aims

The overarching aims of the haematological malignancy GeCIP are to:

1. Use Whole Genome Sequencing (WGS) data to identify new prognostic and predictive biomarkers that inform innovative therapies and precision medicine approaches for Haematological malignancies (HM) patients.
2. Further our understanding of the biological basis of HM through local, national and international collaborations.
3. Lead the transformation of NHS processes for molecular diagnosis and clinical data collections in haematology.
4. Establish a lasting legacy for clinical 'omics and precision medicine for HM patients through training and education of the next generation of clinical academic leaders in genomics.

GeCIP Objectives

1. Collect and sequence up to 7,500 samples with associated clinical data from HM patients recruited through routine clinical care pathways or alongside the UK National Cancer Research Network (NCRN) clinical trials via the NHS GMCs and build a longitudinal life course dataset allowing long-term follow-up.
2. Define candidate prognostic markers and response predictors using an initial test cohort of highly informative patients with available clinical outcome data and/or surrogate markers of treatment response - extreme responders and patients with sequential diagnostic and relapse samples.
3. Validate candidate markers of outcome identified using samples from patients recruited prospectively into clinical trials. For four disease areas (i.e. chronic lymphocytic leukaemia, acute myeloid leukaemia, multiple myeloma and childhood acute lymphoblastic leukaemia) the WGS programme can be integrated into the already on-going adaptive clinical trials (FLAIR, AML18 and 19 MUK Nine and UKALL11) so that results from WGS can ultimately be used for stratified healthcare.
4. Enrich the whole genome data generated by this programme with already existing and future multi-omics and other molecular data from the same patients and to functionally and mechanistically characterise and define clinically relevant genetic variants and complements of variants that classify disease and help improve disease therapies including chemotherapy, immunotherapy and stem cell transplantation.
5. Develop and run alongside this programme appropriate graduate, doctoral and post-doctoral training schemes that will train future clinical and academic leaders in the field of clinical omics and stratified medicine and leaves a legacy to the NHS and UK academia.

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)*

Haematological malignancies refer to blood cancers that affect the blood, bone marrow, lymph and lymphatic system, systems that are all intimately connected by the circulatory (blood) and immune system. Collectively they represent the fifth most common cancer in the UK, and whilst some

patients respond well to treatment, the majority relapse and eventually die from disease progression or therapy-related side-effects. Genetic analyses have shown that Haematological Malignancies are not a single cancer, but a group of diseases with specific genetic features that affect how the blood cancer will behave. To better treat patients, and to develop new innovative targeted therapies requires a comprehensive understanding of the genetics and biology of the various haematological malignancies, it is hoped that the 100,000 Genomes Project data will provide this insight.

Expected start date	**Q2 2017**
Expected end date	**Q2 2020**

Lead Applicant(s)	
Name	Anna Schuh
Post	**Director of Molecular Diagnostics**
Department	Department of Oncology
Institution	University of Oxford
Current commercial links	

GeCIP domain - Expression of interest

Full proposal	
Title *(max 150 characters)*	**Haematological malignancy research in the 100,000 Genomes 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.*

How we will achieve objectives 1-3

To achieve these objectives, we will initially focus on analysing highly informative samples from clinical trial patient cohorts where early surrogate markers of clinical outcome such as minimal residual disease (MRD) monitoring are already available. This will allow us to rapidly identify strong biomarkers and will secure high impact publications in this competitive field. We will select patients who either responded extremely well or extremely poorly (i.e. extreme responders) and/or patients where sequential samples from diagnosis and relapse are available to correlate WGS with clinical outcome. We will then combine WGS results with already on-going multi-omics efforts on the same patient cohorts to establish a comprehensive multi-omics dataset in collaboration with UK researchers and international collaborators. At a minimum (and depending on future funding), this will consist of WGS, RNAseq, miRNA_Seq and methylome data.
Next, candidate outcome predictors will be chosen for prospective validation in a larger cohort of patients recruited through routine clinical care pathways in the GMCs. The size of this second cohort will depend on the specific clinical context. For CLL, MM, ALL, AML we will sequence 800-1000 patients. For HMs such as CML and MPDs where the majority of patients have an excellent outcome on current therapies, but a minority of patients still succumb to disease we will validate findings after selection of an independent second set of extreme (non)-responders. A comparative approach will be applied to overlap syndromes where the identification of common molecular markers would contribute significantly to a precise disease classification.
Candidates and/or combinations of candidates showing a strong independent positive predictive value will then be integrated into adaptive clinical trial designs to direct therapy for specific subgroups of patients.
The work of the GeCIP will be divided into different workstreams that map to its overarching aims and objectives. These workstreams also address a number of specific aspects listed below that are critical to the success of the research elements of this programme and that also involve other stakeholders outside the HM GeCIP domain.

Objective 4 will be achieved through collaboration with the entire UK research community and lies outside the scope of the current specialist programme application. It will be primarily met through on-going research by the applicants using existing funding streams or future grant applications.

Objective 5 is also outside the remit of this programme and will be delivered in partnership with Health Education England and UK academic institutions through current initiatives such as the Masters in Genomic Medicine and the NHSE clinical fellowships (applications from this GECIP are planned)

Workstreams
Overview Workstream 1: Sample cohorts, quality of DNA, clinical data collection and governance
1. Engagement and definition of contractual agreements between GEL and the clinical trial research units (CTRUs) to make clinical baseline and outcome data available without compromising the integrity of the clinical trial in question.

2. Efficient and comprehensive collection of high quality clinical trial samples in collaboration with clinical trial banks and NHSE GMCs without compromising other explorative studies using samples from the same patient.
3. Development of robust governance including a data access and publication policy in collaboration with UK HM researchers and funding bodies.

Overview Workstream 2: Bio-Informatics

Access to a rapid and robust bio-informatics pipeline in the GEL data centre that uses standardised algorithms that are validated and accepted by the wider research community. The Genomics England Data Centre is a 20 petabyte storage system with a minimum of 20,000 computer processing units that runs as a virtual private cloud with the research environment being divided into GeCIP specific embassies to which all approved GeCIP members have access. Funding support for this element and continuous data curation has been confirmed by the Government and runs until 2020.

Overview Workstream 3: Integration of WGS with clinical outcome and already existing and future multi-omics data

1. The establishment of a bio-informatics core that can be accessed and used in an inclusive manner by UK research groups that wish to use the WGS data to perform integrated analyses on existing or future multi-omics data in the data centre
2. Development and use of algorithms to analyse big data

GeCIP Analysis Plan

Starting from this core pipeline, the HM GeCIP will augment this rich dataset using additional methods/algorithms. The following analyses are currently planned (partially in close collaboration with the GEL bio-informatics team and Illumina).

Overview:

Alignment and SNV/indel calling requirements

We plan to use the Genomics England analysis pipeline for this.

Tool requirements and import

We anticipate primarily using bio-informatics tools listed within List_of_Embassy_apps.xlsx . However, we wish to retain the option to import other tools as required to meet the demands of data analysis: in particular those relevant to identification of the range of somatic changes including but not limited to copy number aberrations and low frequency variants (see details below).

Data import

We may wish to import BAM files and VCF files generated from WES or WGS of relevant patient cohorts to meta-analyse data. We may also wish to import genotype files, microarray files, RNAseq and methylome data.

Computing resource requirements

It is difficult to predict at this stage, particularly as recruitment numbers are uncertain. However, we anticipate that 30-50 cores will be required in the first instance for analysis using .vcf file-based variant and phenotype data analyses. Depending on the quality of these data, additional resources may be required if additional analysis of raw data (.bam) files is indicated.

Contamination and Sample checks

GEL will check the integrity of the delivered data and confirm that it meets the required quality level. Sample contamination will be checked with VerifyBamID (GEL) and custom scripts (GeCIP) including assessment of bacterial contamination in the germline DNA obtained from saliva, inference of

germline contamination in leukaemias from B-allele frequency distributions and estimation of germline mosaicism.

Annotation of coding and non-coding variants

The potential deleteriousness and functional consequences of variants will be predicted using multiple programmes. Variants will be annotated using VEP with respect to RefSeq and EnsEMBL gene annotations and we will add information about segmental duplications and conservation (e.g., GERP and phyloP scores; GEL/GeCIP). Additionally we will add deleteriousness predictions for coding and non-coding variants from SIFT, PolyPhen, MutationTaster and CADD. Variants will further be annotated with dbSNP identifiers (version 1.42), and population allele frequencies from multiple databases, including data from the Exome Aggregation Consortium (ExAC), the NHLBI Exome Sequencing Project (ESP6500) and the 1000 Genomes Project (1000G). For somatic variants, we will additionally make use of annotation data from COSMIC, cBioPortal and other publicly accessible databases. We will also employ MutSigCV and similar tools to define the statistical significance of recurrent mutations and will estimate the predicted pathogenic effects of somatic non-coding mutations using FATHMM-MK and FunSeq where required. Additionally, somatic mutations in intronic regions, UTRs, transcription factor binding sites and other non-coding elements as defined by the ENCODE and GENCODE projects will be identified on a genome-wide scale using in-house tools. Significantly mutated elements will be defined as a product of the length of the region, the number of mutations, and the mutational background.

Transition and re-alignment of the sequencing data to the newer version of the human reference genome (GRCh38) is anticipated for early 2016 as mentioned above (done by Illumina and GEL). We are preparing for this scenario by evaluating methods (such as liftover and CrossMap) for upgrading the genomic coordinates of relevant annotation datasets that are unlikely to switch until then (such as ExAC or ESP6500).

Variant filtering

Variant calls will be filtered based on these comprehensive annotations using in-house developed filtering tools and Illumina's VariantStudio software. We will filter, e.g., by minimum variant/genotype quality, minimum coverage and number of alternate-allele reads, predicted deleteriousness, rareness or occurrence in disease databases. Specifically, for somatic variant calling, we will consider only calls that are supported by at least 3 mutant reads and by ≤1% variant allele frequency in the germline. For germline variant calling we will filter common variants based on maximum allele frequencies found in multiple publicly available databases as described above.

Copy number variation calling

We will put a special emphasis on the accurate detection of copy number alterations and structural variations due to their significant role in HM. During our collaboration with GEL on the CLL pilot we experienced low concordance between different automated CNV calling approaches, particularly when working with tumour/normal pairs that were prepared using different sequencing protocols (e.g., FFPE tumour samples and fresh-frozen normal samples). As a consequence we have implemented and optimized an in-house CNV calling pipeline to work with this kind of data. Specifically: Illumina will provide CNV and SV calls using Canvas and Manta respectively and we will create additional call sets using other tools such as BIC-seq, Delly and Socrates. These different call sets will then be integrated with our CNV calling pipeline that is configured around ngbin, ngCGH, and visualisation software such as Nexus Discovery edition v7.5 (BioDiscovery, Hawthorn, CA) or a comparable platform. This will allow us to accurately call and manually post-process and annotate called somatic regions of CN gain, CN loss or copy neutral LOH (cnLOH) where required. Recurrent regions of copy number change will be determined and genes within these regions will be recorded using EnsEMBL annotations. For a number of the HM patients high-resolution SNP array analyses are available (ARCTIC and AdMiRe: n=250) (HumanOmni2.5-8 BeadChips) and these data can be integrated to normalise the WGS results. In addition to optimizing and evaluating this CNV calling pipeline, we will continue to develop own algorithms for the detection of other structural variations

including translocations and chromothripsis. We will also perform Telomere length analysis by counting and normalising reads containing telomeric repeats (TTAGGG) ×3 or (CCCTAA) ×3.

Kataegis

Kataegis is defined as a region of 5 or more mutations each with inter-mutational distances less than two standard deviations from the mean inter-mutational distance. In-house tools using inter-mutational distance and variant allele frequency as metrics will be used to identify regions of somatic hypermutation across multiple samples both within exonic regions and the whole genome. The prevalence of non-exonic mutations with regard to distance from the nearest exon will be calculated across the whole genome.

Identification of mutation signatures

R packages SomaticSignatures from BioConductor and pmsignature will be used to identify mutational signatures of whole genomes and exonic regions. The number of mutational signatures found will be chosen based on the method of Brunet et al. Mutational signatures found in various samples will be correlated with tumour type, as well as patient features such as age, sex, outcomes and responses.

Analysis of transposable elements (TE)

We plan to analyse the TE somatic and germline landscape and expression profiles using TEA and retroseq packages.

Pathway analysis

We will identify pathways with a significant enrichment of mutations. Pathways will be obtained in BioPax format from Pathway Commons, which includes a number of collated databases including KEGG and PANTHER. Genes identified using this process will be cross-referenced with kataegis regions. Pathways implicated will be correlated with disease phenotype and clinical outcome.

Verification of clinically relevant variant calls

We will collaborate with the GEL validation and feedback GeCIP to establish appropriate methods for verification of clinically relevant variants. For example, recurrent SNVs will be confirmed using existing IsoStandard accredited targeted NGS panels or conventional diagnostics methods (pyrosequencing, fragment analysis, PCR, etc.). Substitutions and indels will be verified using orthogonal sequence data which also includes data produced on different sequencing platforms (e.g.: IonTorrent) or data from related nucleotide samples (RNA-seq). RNA-seq data will also be used to confirm translocations. CNVs will be checked by whole genome array if required.

Longitudinal samples for Gear 2 Consideration

Contrary to management of solid tumours, it is standard of care for HMs to re-biopsy patients at relapse or high-grade transformation to obtain more tissue for diagnostics.

1. Relapse

The value of longitudinal sample collections is generally appreciated across the cancer programme. Overall, more then 50% of patients recruited into GEAR 1 will relapse and eventually succumb to their disease. This is why capturing these patients and re-sequencing samples at relapse represents a unique opportunity. We therefore propose that all patients recruited into GEAR 1 should be eligible for sequencing at relapse.

2. High-grade Transformation

Patients recruited into GEAR 1 might also suffer transformation into a more aggressive phenotype. Some patients might present with aggressive disease during GEAR 2 and “legacy” samples of low-grade disease are available for sequencing from their previous presentation. These patients would also be eligible for GEAR 2 (e.g. follicular lymphoma progressing to DLBCL; or MDS progressing to AML; CLL progressing to Richter’s Transformation).

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.

The domain will collaborate closely with the GeCIP for rare inherited haematological and immunological diseases for germ-line analyses and interpretation. A first inaugural joint meeting was held with the two groups in October at Madingley Hall in Cambridge. Through collaboration with Willem Ouwehand's group and the BRIDGE project, we will gain access to genome-wide epigenomic, ATAC and Hi-C data from normal blood cell precursors.
We also have already on-going conversations with a number of cross-cutting GeCIPs, in particular: GeCIPs for bio-informatics pipeline development, the "Machine Learning, Quantitative Methods and Functional Genomics" GeCIP and the health-economics GeCIP.
Importantly, the HM domain has already formed a number of on-going international collaborations across its respective disease areas (German, French and Skandinavian Study Groups, Bio-Informatics Hub in Barcelona (E Campo; ICGC, BluPrint), Dana Farber (C Wu), British Columbia Cancer Agency (A Karsan) and the ICGC consortium (NHL, MM, CML, childhood leukaemia).

Data and informatics requirements
All GeCIP Domains have contributed to the construction of the data model for their tumour type and have contributed to ongoing efforts within Genomics England to develop the clinical and research informatics infrastructure.

Data access and security	
GeCIP domain name	Haematological Malignancies
Project title *(max 150 characters)*	**Haematological malignancy research in the 100,000 Genomes Project**

Applicable Acceptable Uses. *Tick all those relevant to the request and ensure that the justification for selecting each acceptable use is supported above.*

X *Clinical care*

X *Clinical trials feasibility*

X *Deeper phenotyping*

X *Education and training of health and public health professionals*

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

X *Hypothesis driven research and development in health and social care - interventional*

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

X *Non hypothesis driven R&D - health*

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

X *Other health use - clinical audit*

X *Public health purposes*

X *Tool evaluation and improvement*

Information Governance

X The lead for each domain will be responsible for validating and assuring the identity of the researchers. The lead may be required to support assurance and audit activities by Genomics England.

Any research requiring access to the embassy will be required to complete IG Training and read and sign a declaration form. Access will only be granted once these requirements have been met.