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

Application Summary	
GeCIP domain name	**Functional Cross Cutting**
Project title *(max 150 characters)*	**To investigate Endogenous retroviruses, retrotransposons and infectious DNA sequences across GeCIP domains**

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

Sub-domain: Human Endogenous Retroviruses
The objectives of this sub-domain are:
1. To identify and characterize polymorphisms of Human Endogenous Retroviruses (HERVs) in the human population.
2. To test for disease associations for the above-mentioned HERV polymorphic integrations

Sub- domain: Identification of pathogens.
The objective of this sub-domain is to investigate the presence DNA sequences from infectious agents (viruses, bacteria, small eukaryotes) in WGS data that does not align to the human genome. Infectious agents are involved in the development of a variety of human cancers and some rare conditions. Our hypotheses are (a) that there are infectious agents are directly linked to cancer development, but as yet remain undefined and (b) that some rare conditions can result from infections that are not immediately recognized as sepsis. DNA sequences from infectious agents will be identified in both blood and cancer samples.

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)*

Sub-domain*: *Endogenous Retroviruses
More than 10 years have passed since the first high quality (nearly) complete version of the human genome. The most well-studied part is the one that encodes for proteins (known as exome) and accounts for 2-3%. The rest of the genome is much less understood. Around half is comprised of repetitive elements; one category of them, known as Endogenous Retroviruses (ERVs), are remnants of ancient retroviral infections of our ancient ancestors's germ-cells, and comprise around 5-8% of the human genome. Our research explores potential health effects of Human ERVs by combining bioinformatics and wet-lab approaches.

Sub-domain: Pathogens
Over 30 research groups on the Norwich Research Park, carry out research on bacteria (http://www.micron.ac.uk/). This work, linked to the BBSRC funded The Genome Analysis Centre (TGAC) (http://www.tgac.ac.uk/), is generating discoveries that have applications in the fields of human health, pharmaceuticals, agriculture, food processing and environmental monitoring. We will apply our combined expertise to the identification of human pathogens using Whole Genome DNA Sequence (WGS) Data from GeL. As part of the Pan-Cancer ICGC (International Cancer Genome Consortium) Project a bioinformatics analysis pipeline (SEPATH) has already been set up within TGAC to carry out this work, demonstrating proof of principle.

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)*

Sub-domain Endogenous Retroviruses

A significant proportion (~5-8%) of the human genome is comprised of mostly defective Human Endogenous Retroviruses (HERVs), the descendants of occasional germ line invasions by exogenous retroviruses (XRVs). They are classified into 30-40 families, each one being considered as an independent invasion of the germ line. Their role on the development of human disease is still unclear. Our research aims to explore the pathophysiologic effects of HERV replication activity and expression.

Here we are interested in exploring 2 research questions:

1) What is the extent of HERV polymorphic integrations?

2) What is the pathophysiologic effect of these polymorphic integrations?

To do so I propose to explore these questions by focusing on 2 families, namely HERV-K HML-2 (hereafter called as HK2) and HERV-H.

We will use our in-house bioinformatics approaches to mine the genomes and catalogue polymorphic integrations for these HERV families. We note that recovering HERV integrations is computationally challenging, thus all the full-genome projects up to date have either not recovered this information at all or the identification of novel HERV integrations is subject to high error rate.

We will then assess if these integrations are associated with diseases. The project will provide translational insights about the roles of HERVs in human disease.

Sub-domain: Pathogens

The assumption behind the SEarching for PATHogens (SEPATH) pipeline is that when samples are collected from tumours and blood, the DNA from any pathogens present will be collected along with the human tissue. Therefore, when this DNA is whole genome sequenced, it is possible to detect and quantify pathogens. We will apply the SEPATH analytical pipeline to selected cancer sample and bloods sequence datasets generated by Genomics England.

The SEPATH pipeline consists of two approaches:

Approach 1 – Specific and curated

Our starting point is to take the leftover reads that are found to not map to the human genome. Reads are trimmed and filtered. Metagenomic Phylogenic Analysis (MetaPhlAn) is then applied to identify and quantify the presence of bacterial, small eukaryotes and viral populations.

Approach 2 – Sensitive and assembly

This approach takes the same initial steps for obtaining good quality data as Approach 1, but further filtering is performed. The first step is to remove any data that could possible have come from the human tissue by filtering the reads using Kontaminate. Non-human reads are then subjected to metagenomics de novo assembly using MetaCortex. Contigs are then assigned to species by matching the contig sequence to the NCBI nucleotide sequence database using megablast. A contig may be assigned to multiple species and so a statistical approach is used to assign the contig to the lowest common ancestor in MEta Genome ANalyzer (MEGAN).

The studies are expected to identify "smoking-gun" bacterial, viral and other infectious species that represent a starting point for examining potential mechanisms of disease induction together with the appropriate GeCIP partner.

Expected start date	1/04/2016
Expected end date	31/03/2018

Lead Applicant(s)	
Name	Gkikas Magiorkinis
Post	Senior Clinical Fellow – Honorary Consultant in Medical Virology
Department	Zoology
Institution	Oxford
Current commercial links	None
Lead Applicant(s)	
Name	Colin Cooper
Post	Chair of Cancer Genetics
Department	School of Medicine
Institution	University of East Anglia
Current commercial links	None

Administrative Support	
Name	Daniel Leeson
Email	d.leeson@uea.ac.uk
Telephone	01603 591953

Subdomain leads		
Name	**Subdomain**	**Institution**
Gkikas Magiorkinis	Endogenous retroviruses, retrotransposons	University of Oxford
Colin Cooper	Pathogens	University of East Anglia

Detailed research plan

Full proposal (total max 1500 words per subdomain)	
Title *(max 150 characters)*	**Sub-domain: Human Endogenous Retroviruses.**

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).*

Sub-domain: Human Endogenous Retroviruses.

Endogenous retroviruses (ERVs) are retroviruses that are inherited through the host germline; they are descendants of occasional germ line invasions by exogenous retroviruses (XRVs). Well-defined groups determined by phylogenetic analysis are termed "families" and generally represent a single invasion followed by expansion within a host genome. Human ERVs (HERVs) are classified into 30-40 families.

HERVs are inactivated and down-regulated through random knock-out mutations, hypermutation and silencing mechanisms, but are up-regulated in patients with cancer and other diseases. The observation that cancer cells produce virus-like antigens and particles is very old; however it is a phenomenon that has only partially been described and its role in cancer biology is still under

investigation.

HERVs can also be involved with the development of cancer through tumorigenic and immunosuppressive proteins. For example HK2 loci produce two non-standard retroviral proteins, NP9 and Rec, through alternatively spliced mRNA. Rec is analogous to the Rev protein of HIV-1 and has been shown to promote tumour development in mice. The transmembrane domains of some retroviral *env* proteins have been found to be immunosuppressive. Another possible pathogenic mechanism would be through promoter activity of LTRs (Long Terminal Repeats); for example, de-repression of one the oldest HERVs has been recently connected with the development of Hodgkin's Lymphoma.

HERVs can also be connected to the development of autoimmunity through innate immunity signalling pathways. A recent study identified a critical role for the ERV-mediated activation of cytosolic innate immune signalling pathways in B-cell responses. On the other hand HERV antigens might promote immune responses. HERV antigens are host antigens and believed not to promote immune response upon their expression. However, if an ERV antigen was not presented during the development of the immune system, a later expression could promote immune response. Immune responses against HK2 antigens have been reported in patients with up-regulated HK2. Older studies have shown that the envelope gene of HERV-K is a superantigen, although there are no solid findings with respect to health implications.

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.*

For the suggested project we will be exploring two research questions related to HERVs and their pathophysiologic potential:
1) What is the extent of HERV polymorphic integrations?
2) What is the pathophysiologic effect of these polymorphic integrations?
Polymorphisms among individuals with respect to HERVs can be considered at 3 levels:
1) Integration: existence or absence of a HERV in a specific location, 2) Viral Genome integrity: an integration can be either full-length or solo-LTR (the result of recombinational deletion that occurs after integration and results into removing the provirus apart from one of the two LTRs), 3) Mutations within the viral sequence: single nucleotide polymorphisms (SNPs) or larger deletions/insertions within the HERV sequence.

We will explore the two research questions by focusing on two families, namely HERV-K HML-2 (hereafter called HK2) and HERV-H.
HERV-K HML-2 (HK2), the most recently active family: We have shown that over the last 10 million years HERV replication activity appears to have slowed down within the human genome. We also showed that this can be at least partially attributed to the increase of our body size. The vast majority of HERVs in the human genome have acquired many deleterious mutations. The family that retained replication activity until at least 250,000 years ago, as evidenced by multiple human specific insertions, and shows polymorphism within the human population is HK2, and for this reason some HK2 integrations are almost intact. Recently, we analyzed whole human genome sequencing (WGS) projects, identified 17 un-catalogued HK2 polymorphic integrations. Crucially, we found two of these integrations within gene introns, one within an intron of the RASGRF2 gene and one in the predicted Transmembrane Protein 117.
HERV-H, the superspreader of the human genome: In 2012 we published a paper in PNAS showing that endogenous retroviruses, which lost their env genes and evolved into replicating intracellularly (effectively becoming retrotransposons) are more abundant than ERVs who

continued to replicate via an extracellular life-cycle. HERV-H is such a superspreader family (or Mega-family) in the human genome. There is evidence that HERV-H had replication activity in the human genome within the last 10 million years, thus it is possible that some HERV-H integrations are polymorphic as it has been documented in 2 studies, but the extent and the role of HERV-H polymorphic integrations is unknown. We also showed that the apparent recent slow-down of retroviral activity in the human genome coincided with the exceptional deceleration of HERV-H activity. Crucially, HERV-H sequences have recently been shown to be part of long non-coding RNAs (lncRNA) playing a major role in the pluripotency of stem cells.
To summarise, HK2 and HERV-H provide a good starting point to explore the pathophysiologic effects of ERVs in humans because: 1) they cover the range of ERV evolutionary strategies: the re-infecting family (HK2) with full-length retroviral coding capacity and the retrotransposing Mega-family (HERV-H) with partial retroviral coding capacity, 2) at least one of them is polymorphic with respect to integrations, 3) We have shown that ERV replication activity poses a deleterious burden to the host which scales-up with body size; more specifically our model predicts that families which, like HK2 and HERV-H, have retained their replication activity until very recently are more likely to be virulent.

We will screen human WGS data and identify uncatalogued integrations using our previously described bioinformatics approach. Briefly, our approach analyzes sequence reads to find evidence in support of integrations that exist in the WGS data but not in the human reference genome. The pipeline includes 3 steps to prove that an ERV is integrated in a specific location, the final step being the visual inspection of the alignment of sequence reads spanning the integration site and part of the retroviral LTR. We have shown that this approach is more accurate than other algorithms alone (e.g. Retroseq), which tend to provide false positive integrations compared to the human reference genome.

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 establish associations between diseases and the polymorphic integrations using the Genomics England dataset and in close collaboration with the other GeCIP domains. We have already established a link with the neurodegenerative and the colorectal cancer GeCIP domain, and we will endeavour to establish links with other GeCIP domains as our research moves forward.

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

We expect that the domain will work as a hub for students that wish to pursue research degrees (PhD or Masters by research) in the area of endogenous retroviruses and paleovirology. Other opportunities for training can be leveraged through existing fellowship schemes. We also foresee that there can be opportunities for short-term training positions (e.g. 3-6 months). An example would be to train researchers from other GeCIP domains (e.g. a disease-specific domain) as they could potentially be interested in testing the ERV polymorphisms with respect to their specific disease domain.

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

The domain aims to cover a wide range of areas with respect to Human Endogenous Retroviruses and pathophysiology:

1) Paleovirology (Katzourakis, Magiorkinis, Tristem, Stoye, Kassiotis)
2) Retrovirology (Stoye, Bangham, Katzourakis, Magiorkinis, Kassiotis)
3) Immunology (Hurst, Bangham, Stoye, Kassiotis, Klenerman)
4) NGS technologies and Bioinformatics (Mbisa, Karamitros, Melamed, Kanapin, Samsonova)
5) Molecular Pathology (Tomlinson)
6) Population genetics (Hein, Tomlinson)

The expertise and skills are evidenced by the CVs of the respective researchers.

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.*

We will work in close collaboration with the disease-specific and the cross-cutting Validation and Feedback GeCIPs to ensure that the clinical interpretation of these novel polymorphic integrations will be efficiently translated in clinical applications.

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

Our research will attract the attention of many researchers working on the area of genomic susceptibility of diseases as well as retrovirologists.

HERV polymorphisms are another source of genomic diversity to be connected with the pathophysiology of human diseases. However, HERVs are far more complex than Single Nucleotide Polymorphisms, which are mainly considered in Genome Wide Association Studies (GWAS). Thus the community interested in genomic susceptibility of disease would be interested in the links discovered through the suggested project. We will communicate the results first through the Genomics England project and then with publications and conferences.

Retrovirologists such as John Coffin (Tufts University) and Jonathan Stoye (The Crick Institute) have worked on HERVs. Thus, the potential links between HERVs and human diseases would be of interest for their groups as well. Given that many polymorphisms remain to be discovered the project will provide new insights into the co-evolution of the host and these retroviral parasites

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 group will seek for links with partners from industry. We foresee that two aspects of the domain activities would be of interest. The first would be the development of therapies and diagnostics (or biomarkers) based on ERV products (proteins, nucleic acids) if they are found to be associated with disease states. The other could be that the genome analyses might be of interest for bioinformatics' companies on a service provision basis as the processes and algorithms are expected to be computationally intensive.

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

1. Boeke JD & Stoye JP (1997) Retrotransposons, Endogenous Retroviruses, and the Evolution of Retroelements. *Retroviruses.*, eds Coffin JM, Hughes SH, & Varmus HE (Cold Spring Harbor Laboratory Press), pp 343–436.
2. Tristem M (2000) Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. *J Virol* 74(8):3715-3730.
3. Magiorkinis G, Belshaw R, & Katzourakis A (2013) 'There and back again': revisiting the pathophysiological roles of human endogenous retroviruses in the post-genomic era. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 368(1626):20120504.
4. Magiorkinis G, Blanco-Melo D, & Belshaw R (2015) The decline of human endogenous retroviruses: extinction and survival. *Retrovirology*.
5. Barbulescu M*, et al.* (1999) Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. *Curr Biol* 9(16):861-868.
6. Hughes JF & Coffin JM (2004) Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. *Proc Natl Acad Sci U S A* 101(6):1668-1672.
7. Marchi E, Kanapin A, Magiorkinis G, & Belshaw R (2014) Unfixed endogenous retroviral insertions in the human population. *J Virol* 88(17):9529-9537.
8. Belshaw R*, et al.* (2005) Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): implications for present-day activity. *J Virol* 79(19):12507-12514.
9. Marchi E, Kanapin A, Byott M, Magiorkinis G, & Belshaw R (2013) Neanderthal and Denisovan retroviruses in modern humans. *Curr Biol* 23(22):R994-995.
10. Agoni L, Golden A, Guha C, & Lenz J (2012) Neandertal and Denisovan retroviruses. *Curr Biol* 22(11):R437-438.
11. Lu X*, et al.* (2014) The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nature structural & molecular biology* 21(4):423-425.
12. Wang J*, et al.* (2014) Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature*.
13. Zeng M*, et al.* (2014) MAVS, cGAS, and endogenous retroviruses in T-independent B cell responses. *Science* 346(6216):1486-1492.

Detailed research plan

Full proposal (total max 1500 words per subdomain)

Title *(max 150 characters)*	**Sub-domain: Pathogens**

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).*

Hunting for Human Infectious Agents at the Norwich Research Park (HHIAN)

Infectious agents are involved in the development of a variety of human cancers. The involvement in papilloma virus infection in the development of cervical cancer, of hepatitis virus in the development of liver cancer, and EBV in the development of nasopharyngeal cancer are well established[1]. Similarly the links between *Helicobacter pylori* infection and stomach cancer and

Schistosomiasis infection and bladder cancer have been well documented[2,3]. Strong links between infection by *Salmonella* Typhi and the development gall bladder cancer have also been reported[4]**. We hypothesise that other infectious agents are directly linked to cancer development, but as yet remain undefined.** Indeed for some cancer types there are already clues that new infectious agents may be involved in cancer development. Prostate cancer incidence, for example, seems at least in part to be linked to genital infections[5]. Confirmation of the view that many of the underlying causes of cancer remain to be identified has been provided analysis of mutations from 7,042 cancer samples across 30 different cancer types[6]. 21 distinct mutational signatures were identified, but the causes of only nine these signatures are currently know**.**

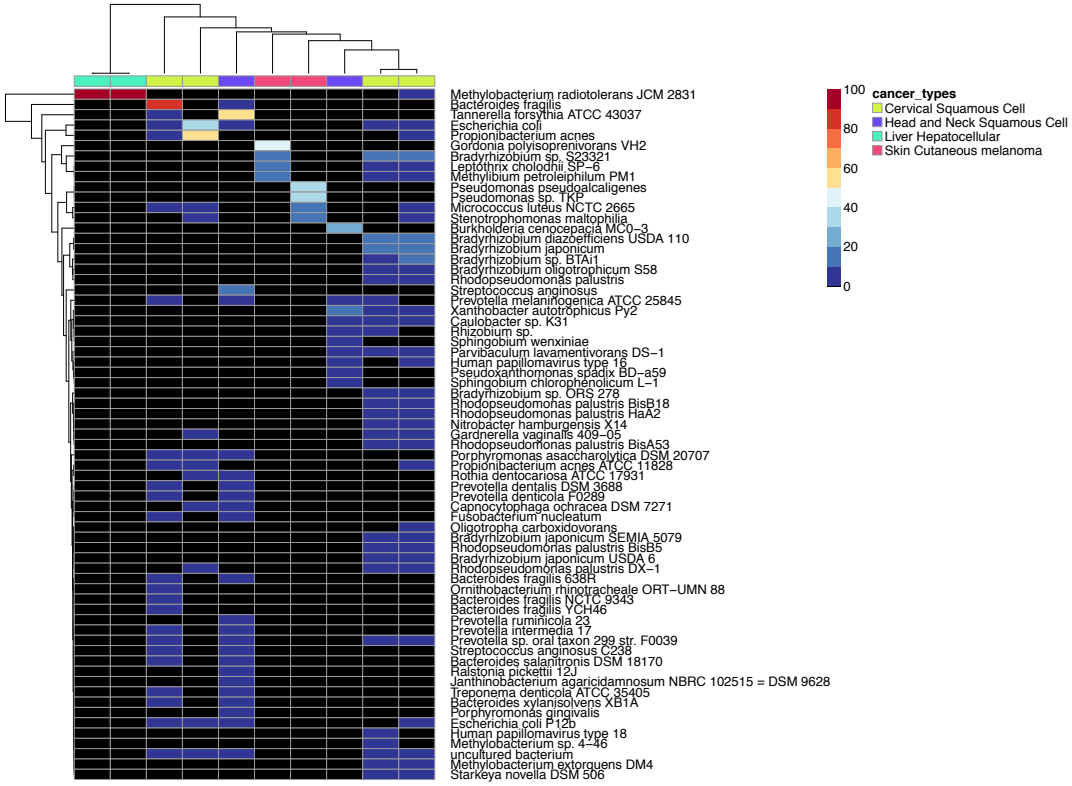

Figure 1. Detection the DNA sequences of infectious agents in GWS data from human cancers

Blood is a particularly challenging matrix to perform NGS based infectious diseases diagnostics due to the vast amount of human vs pathogen cells present. However, recent studies have demonstrated the potential of this approach. Clinically relevant concentrations of bacteria and viruses were detected in blood samples using a direct NGS approach[7]. Researchers at the Broad Institute demonstrated that the malaria parasite could be diagnosed by NGS of infected blood[8] and Depledge *et al*. demonstrated sequencing based detection of viruses directly from clinical blood samples[9]. **Our hypothesis is that some rare conditions can result from infections that are not immediately recognized as sepsis.** We have developed novel strategies for detecting infectious agents in blood (paper under review).

The availability of Whole Genome DNA Sequence (WGS) from cancer and blood genomes via Genomics England provides an important opportunity to discovery new pathogens linked to human health. The SEPTH pipeline is already in place, working and currently being improved. We have applied SEPTH to non-aligned whole genome DNA sequences from 10 cancers representing 4 cancer types to demonstrate poof of principle. As expected we detect papilloma virus in cervical cancer, and several bacteria known to be present in the mouth in head and neck cancer (Figure 1).

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.

Key applicants (Cooper, Brewer) are members of the ICGC PanCancer Pathogens working group that have already applied this approach for detecting pathogens to 2500 WGS datasets from cancer and blood samples. The two pipelines to be used in these analyses are as follows

MetaPhlAn Pipeline: Our starting point is to take the leftover reads not mapping to the human genome. Low quality bases (q<30) are trimmed from the read ends, reads less than 32bp are discarded, and additional filtering is performed to remove reads containing more than 5% of Ns or those with low complexity. Metagenomic Phylogenic Analysis (MetaPhlAn) is then applied to identify and quantify the presence of bacterial, small eukaryotes and viral populations. MetaPhlAn comes with a curated database of sequences each unique for a clade.

Assembly Pipeline: This approach takes the same initial steps for obtaining good quality data as Approach 1, but further filtering is performed. The first step is to remove any data that could possible have come from the human tissue by filtering the reads using Kontaminate. Non-human reads are then subjected to metagenomics de novo assembly using MetaCortex. Contigs are then assigned to species by matching the contig sequence to the NCBI nucleotide sequence database using megablast. A contig may be assigned to multiple species and so a statistical approach is used to assign the contig to the lowest common ancestor in MEta Genome ANalyzer (MEGAN). Preliminary analyses from the assembly pipeline is show in Figure 1.

The output is a matrix of the presence/absence of significantly detected clades per sample plus the percentages of each clade present in a sample. We will examine count matrices to answer the following questions:

1 Are any pathogens significantly over-represented in a particular disease compared to normal tissue and other disease types. We will use Segata *et al*.'s LEfSe, a non-parametric Kruskal-Wallis test followed by subsequent Wilcox rank-sum tests on subgroups, and metagenomeSeq.
2 Are any pathogens increasingly common as the risk category increases or are they associated with other clinical categories.
3 Does the diversity of the microbiome/virome vary across clinical categories. We will calculate the alpha diversity and Shannon index.

This phase of the study will identify “smoking gun” pathogens that we be further investigated as described in the Clinical Interpretations Section below.

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 were late to start on this process due to an error in an email address sent by GeL. However the past 2 weeks we have emailed all of the GeCIP heads to start discussions. Our model of this GeCIP is quite simple and will be presented an appropriate context to each GeCIP. Namely we wish to carry out analysis of bacterial, viral and small eukaryote DNA sequence as part of the standard analysis pipeline performed by GeL. We are currently in discussion with Matthew Parker as to how this might be achieved. All approved samples will then be run through the SEPATH pipeline and the output data will be made available to the appropriate GeCIP. The analyses are quite straight-forward when set up and could in principle be applied to all data produced by GeL. Discussions with then occur with each GeCIP to determine whether links can be established

between individual pathogens and disease. We then have a number of technologies available that could be used to confirm any associations in more focused studies. Specific responses that we have received so far are listed below. Each expression of interest will be followed up to formulate a specific research strategy.

Rare Diseases
Renal (Daniel Gale). Expression of interest. Teleconference planned.
Gastroenterology and Hepatology (Guy Chung-Faye, Patrick Dubour). Email discussion in progress.
Cardiovascular (Bernard Keavney). Expression of interest received.
Sight and Hearing (Andre Webster). Expression of interest received.
Endocrine and Metabolism (Stephen O"Rahilly). Expression of interest received
Inherited Cancer Predisposition (Clare Turnbull). Expression of interest received,
Cancer
Prostate Cancer (Johnn De-Bono). Expression of `Interest. To be presented at their next GeCIP telephone conference.
Childhood Cancer (Josef Vormoor). Expression of interest received.
Renal Cell Carcinoma (James Larkin). Telephone conference planned.
Colorectal Cancer (Ian Tomlinson). Email discussion in progress.
Breast Cancer (Nick Turner). Expression of Interest received.

We have an advisory collaboration with Karen Sfanos at the Johns Hopkins University.

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

With the development of ICGC and Genomics England initiatives there is an urgent requirement to train new bioinformaticians that can translate their knowledge to clinical benefit. All 4 partners on the Norwich research Park (TGAC, UEA, IFR, NNUH) are committed to providing opportunities to doctoral and postdoctoral workers to train with Genomics England data. Therefore an essential component of this proposal is a commitment to deliver a programme of training and outreach such that the community can gain the maximum value from our efforts. Through TGAC, we will deliver training courses and workshops that will enable researchers to explore and analyse the relevant datasets. The TGAC Scientific Training and Skills Programme is geared to offer high quality training to current and the next generation of life scientists in advanced bioinformatics and next generation genomics including courses in RNA-Seq, genomics, SNP calling/genetics, programming, Linux and python for life scientists, metagenomics, ChIP-Seq and more.

UEA and the NNUH participate in the NIHR Academic training programme. This has clinical lecturers and academic clinical fellows in microbiology, gastroenterology, paediatric gastroenterology, respiratory medicine, medicine for the elderly, cardiology and endocrinology. Our proposed programme of research will give excellent research training opportunities in genomic medicine for these trainees. The networks set up to run the HHIAN project will provide strong mentorship for translation of knowledge from TGAC to clinical benefit in the NNUH.

Additionally the Norwich Research Park is planning to set up postgraduate modules specifically focused on bioinformatics and its application to clinical medicine.

People and track record. *Explain why the group is well qualified to do this research, how the investigators would work together.*
We have put together a group of experts, each of whom is an established expert in their chosen

field and has a proven track record of working as part of multi-disciplinary teams (see individual CVs for details). We have expertise to run and develop our analytical pathway (Brewer,Di Palma). We have expertise in interpreting the clinical relevance of the infection data that we will generate (O' Grady, Livermore, Mithen, Narbad, Walshaw, Wain,Eeles, Kote-Jarai, Sfanos) and expertise that will allow us to further investigate discoveries using clinical datasets and other technologies as required (Fraser, Frenneaux, Bowles, Wilson, Cooper). Additionally we will link to the specialist interested of other GeCIPs as listed in the sections above. There is an enormous resource of knowledge on bacteria at Norwich that can be accessed as required. Cooper and Brewer have papers in Nature and Nature Genetics describing results from the ICGC Prostate Project, O'Grady has a recent publication in Nature Biotechnology describing the application of nanopore DNA sequencing to bacteria. And novel techniques for detecting bacteria in biological materials have been developed. If significant viral or other pathogen DNA sequences are identified we will build links as appropriate to investigate the significance of our observations.

We will run the project in the same way that we run other multicentre studies. Namely, through minuted telephone conferences, probably once every 6 weeks initially. A critical component of the study at this stage will be to interact with Genomics England to get the components of our SEPATH pipeline integrated into the Genome Analysis carried out by Genomics England.

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.*

It should be noted that the discovery of infectious-bacterial viral or other DNA sequences, associated with a particular cancer type or disease, does not fulfill either classic or molecular Koch criteria for causation. Our studies are not expected to establish a causal link between the presence of bacterial infection and disease. However they will establish in an objective fashion, through use of appropriate study design, whether there really is a link between presence of an infectious agent and a particular disease. The studies are expected to identify "smoking-gun" bacterial viral and other infectious species that represent a starting point for examining potential mechanisms of disease induction together with the appropriate GeCIP partner.

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

The work will eventually lead to the recognition of new links between infection by particular agents and specific human diseases that may suggest novel treatment strategies.

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 will search for opportunities to file patent protection for novel discoveries.

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

1. Bergonzini, V., Salata, C., Calistri, A., Parolin, C. & Palù, G. View and review on viral oncology research. *Infect. Agents Cancer* **5,** 11 (2010).
2. Chang, A. H. & Parsonnet, J. Role of bacteria in oncogenesis. *Clinical microbiology reviews* (2010).
3. Badawi, A. F., Mostafa, M. H. & Probert, A. Role of schistosomiasis in human bladder cancer: evidence

of association, aetiological factors, and basic mechanisms of carcinogenesis. … *journal of cancer …* (1995).
4. Nagaraja, V. & Eslick, G. D. Systematic review with meta-analysis: the relationship between chronic Salmonella typhicarrier status and gall-bladder cancer. *Aliment Pharmacol Ther* **39,** 745–750 (2014).
5. Wright, J. L., Lin, D. W. & Stanford, J. L. Circumcision and the risk of prostate cancer. *Cancer* **118,** 4437–4443 (2012).
6. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).
7. Frey, K. G., Herrera-Galeano, J. E. & Redden, C. L. Comparison of three next-generation sequencing platforms for metagenomic sequencing and identification of pathogens in blood. *BMC …* (2014).
8. Melnikov, A. *et al.* Hybrid selection for sequencing pathogen genomes from clinical samples. *Genome Biol.* **12,** R73 (2011).
9. Depledge, D. P. *et al.* Specific capture and whole-genome sequencing of viruses from clinical samples. *PLoS ONE* **6,** e27805 (2011).

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)*

Sub-domain: Endogenous Retroviruses

We plan to handle and analyse data in BAM format. With the C++ software we have developed to find integrations in WGS samples we aim to scan at least 1000 whole genomes for discovering novel integrations. We will then interrogate the other genomes for presence absence of the annotated catalogue of our novel integrations.

Sub-domain: Pathogens

Groups of participants: All groups of participants, in particular, prostate cancer, colorectal cancer and hematological malignancy as well as the following non-cancer conditions: respiratory, cardiovascular, gastroenterological, musculoskeletal, and hematological disease.

Form of data: BAMs with unaligned reads and phenotype data

This is a cross-domain GeCIP. We have no agreement in place with GeIP-specific domains but are very happy to share interesting findings specific to the projects.

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

Sub-domain: Endogenous Retroviruses

We will process whole genomes in BAM format with our C++ program, which can run on 8 cores. Remapping of reads of interest will be carried out with a sensitive aligner such as *NovoAlign*. We will analyse presence and absence, as well as novel integration discovery, using third-party tools such as BEDtools. Determination of polymorphic integrations will be executed using in-house written R scripts. We will mark which polymorphic integrations are solo-LTRs or full-lengths to determine the viral integrity of these unfixed sites and perform variant calling on polymorphic integrations using the reads collected by our algorithm.

Sub- domain: Pathogens

The data analysis plans are described in the research plans section above.

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)*

Sub- domain: Endogenous Retroviruses
We are interested only on the type of the disease. We will collaborate with the disease specific domains to assess whether the polymorphic integrations can be associated with the disease of their interest.

Sub-domain: Pathogens
Disease Type and all phenotype data collected by individual GeCIPs. History of use of medications (antibiotics, anti-viral etc)

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)*

Sub-domain: Endogenous Retroviruses
Our requirements do not differ from the Genomics England analysis pipeline. We will use the BAM files of WGS for processing.

Sub-domain: Pathogens
Our requirements do not differ from the Genomics England analysis plan. Our only request is that all reads are retained in the BAM even if they do not map to the human genome.

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)*

Sub-domain: Endogenous Retroviruses
We will be using a more sensitive aligner, such as *NovoAlign*, to align the collected reads from our algorithm. If *NovoAlign* will not be made available, we would prefer a more sensitive than BWA aligner to be provided amongst the list of available tools.

Sub-domain: Pathogens
Cutadapt
Prinseq
Metaphlan
Kontaminate
MetaCortex
MEGAN
LEfSe
metagenomeSeq (R/Bioconductor package)

Other software/tools might be required as the pipeline and analyses develop.

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)*

Sub-domain: Endogenous Retroviruses
We will require to have lists of RefSeq genes, of RepeatMasker elements, and of custom compiled lists of HERV integrations to mark presence and absence of integrations as well as to characterise them in regards to their proximity to genes and repetitive elements.

Sub-domain: Pathogens
Metaphlan database (comes with software)
NCBI nucleotide sequence database in blast format (nt)
NCBI protein non-redundent sequence database in blast format (nr)

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

Sub-domain: Endogenous Retroviruses
Our algorithm takes about 50 hours (clock-time) on 8 cores for a 40X-50X coverage whole genome. A single node consisting of 24 cores can have 3 genomes processing at a time in parallel. Our resulting outputs should not take up more than 10 GB per genome.

Sub-domain: Pathogens
We will require storage space for all the unmapped reads.
De novo assembly can be resource intensive (memory)

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)*

Sub-domain: Endogenous Retroviruses
We do not plan do perform analyses with omics samples during this part of the project, but it is highly likely that we will need access in a second wave.

Sub-domain: Pathogens
There may be a downstream requirement to look at specific antibodies or DNA sequences present in -omics samples.

Data access and security	
GeCIP domain name	**Functional Cross Cutting**
Project title *(max 150 characters)*	**To investigate Endogenous retroviruses, retrotransposons and infectious DNA sequences across GeCIP domains**

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

X *Clinical care*

☐ *Clinical trials feasibility*

X *Deeper phenotyping*

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

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

X *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*

X *Public health purposes*

☐ *Subject access request*

☐ *Tool evaluation and improvement*

Information Governance

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

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

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)