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. 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	**Pan cancer**
Project title	**Pan cancer analysis research in the 100,000 Genomes Project**

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

- to focus on the integration and analysis of cancer genome sequencing data from Genomics England.
- to engineer a federated computational architecture to enable the analysis of all available cancer genomes internationally while respecting local rules and restrictions.
- The detection of rare and noncoding drivers
- to characterise both established and novel genetic factors that affect both efficacy and toxicity by facilitating the collection of high quality cancer drug treatment toxicity data across multiple tumour types, many of which are treated with the same, or similar, drugs.
- The assessment of possible genetic differences in the HOX clusters in patients who develop a range of cancers including those of the breast, ovary, prostate, pancreas, head and neck, bladder, lung, brain and digestive tract, together with melanoma, leukemia, and lymphoma.
- to study broadly cancer symptomology with a specific focus on survivorship, supportive care in cancer and palliative care in cancer.
- To ensure that early stage disease is appropriately targeted within each cancer domain, and where appropriate, prompt for inclusion of precursor lesions.
- To explore early cancer drivers that may be common between the cancer types.
- To integrate genomic and environmental data to generate a cancer risk profile that will be used to stratify cancer-preventive interventions.

Lay summary. *Information from this summary may be displayed on a public facing website. Provide a brief lay summary of your planned research. (max 200 words)*

The domain aims to focus on the analysis of cancer genome sequencing data across all tumour types from the 100,000 Genomes Project. There are complex patterns of mutation(s) specific to tumour types and patterns of mutation(s) that can be found across tumour types. Analysis of this would be of great importance but requires sample sizes in the thousands to robustly detect. Increased sample sizes leveraged across different tumour types provide insights beyond tumour-specific analyses.

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

Lead Applicant(s)	
Name	Prof Dion Morton
Post	Head of Academic Department of Surgery
Department	Academic Department of Surgery
Institution	Institute of Cancer and Genomic Sciences, University of Birmingham
Current commercial links	

Subdomains	Subdomain leads
Pan-cancer analysis	Dr Peter Campbell
Cross-cancer analytics	Francesca D. Ciccarelli
Pan-cancer drug therapy toxicity	Dr. Jean E Abraham and Professor Paul Pharoah
HOX gene clusters in cancer	Prof Richard Morgan
Radiogenomics	Prof Neil Burnet
Cancer symptomology	Dr Ollie Minton PhD FRCP FHEA
Cancer Prevention	Prof Dion Morton

GeCIP domain - Expression of interest

Full proposal	
Title *(max 150 characters)*	**Pan cancer analysis 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.*

The Pan Cancer GeCIP domain will carry out analysis across all tumour samples collected during the 100,000 Genomes Project and use the totality of the data to investigate driver mutations across tumours, and establish whether there are signatures that affect treatment efficacy and toxicity that are common across tumour types. The domain will also investigate analytical and computational solutions to enable the co-analysis of 100,000 Genomes Project genomes with other genomes present in other datacentres, whilst respecting the local rules and restrictions around access.

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.*

Out of necessity, the domain will form close collaborations with all other cancer domains and would be interested in collaborating with cross-cutting domains (e.g. the Machine Learning domain) to share ideas on meta-analytical techniques.

Detailed research plan – Subdomain

Full proposal (total max 1500 words per Subdomain)	
Title *(max 150 characters)*	**Pan-cancer analysis**

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.*

Genomics England will sequence ~40,000 whole cancer genomes and matched germline DNA over the next 2-3 years. This represents a considerably larger collection of whole cancer genomes than any that currently exist, and will be of similar magnitude to efforts planned elsewhere in the world in the same time frame, including from the NCI and ICGC-2. There will be a GeCIP domain for each tumour type analysed by Genomics England, focusing on the genomic and clinical aspects of their particular cancer. We propose to establish a GeCIP with an aim to understand patterns of germline and somatic variation in cancer genomes across tumour types. We believe such a GeCIP would be exploring fundamentally different questions to individual tumour type domains, and would produce data that would assist the individual domains in their interpretation of the genomic data. There are many fascinating research questions that are best answered in a broader, cross-tumour context:

- Personalised, real-time, data-driven cancer genome annotation – An ultimate goal of the GeCIP will be to develop a software tool that can, in a clinically relevant time frame, sift through the full catalogue of somatic and germline variants in a given patient's cancer genome and compare against the full pan-cancer database. The aim would be to identify mutations that (1) are biologically relevant to the cancer; (2) influence patient prognosis; (3) could modify therapeutic choices; or (4) have relevance for familial risk of cancer.
- Driver mutations – Cancer genes are often, but not always, mutated in multiple tumour types. Combining sequencing data from multiple cancers has identified many previously unsuspected cancer genes and documented their frequency in individual tumour types. The Genomics England initiative promises the capacity to extend this to non-coding point mutations and genome-wide structural variants.
- Mutational signatures – Cancers are characterised by distinctive mutational signatures, either from cell type-specific DNA repair defects or through carcinogen exposures. Analysing patterns of mutation across tumour types considerably improves power to detect specific signatures. Data from Genomics England will extend this considerably, enabling us to study genome-wide distribution of mutations, integrate structural variation with point mutations and identify causative genomic changes across tumour types.
- Patterns of tumour evolution – We and others have documented extensive subclonal heterogeneity in human cancers, with implications for clinical progression, treatment response and disease relapse. There appear to be differences across tumour types in the extent and patterns of cancer evolution, but this has not yet been systematically studied.
- Identification of novel pathogens – Cancer development can be linked to pathogen exposure: for viruses (cervical and liver cancer), bacteria (gastric and gall bladder cancer) and parasites (bladder cancer). We will identify infectious agents linked to cancer development using algorithms developed by the Pan-Cancer ICGC Pathogens Working Group.
- Influences of germline genotype on cancer development – Germline predisposition to cancer ranges from rare, high penetrance alleles to common, low penetrance alleles. For a few of the former, such as mismatch repair deficiency or BRCA1/2 loss, the germline variant determines patterns of mutations that emerge in the cancer, but this has not been explored for most such changes. The cancers sequenced by Genomics England will provide better statistical power for disentangling these effects than has been previously possible.

The key aspects of the Pan-Cancer QPQ computational environment will be:

- A system for assessing user credentials, protecting data security and maintaining ethical standards;
- A *quid pro quo* marketplace for trading data, analytics and access to compute / storage that will reward data generators and analysts alike;
- Access to state-of-the-art algorithms for generators of cancer genome sequencing and associated data in the data bank;
- A menu of analytic tools that can be accessed to run across up-loaded data;
- A federated computational architecture that will be physically collocated near large data repositories in several countries but will enable aggregated data analysis to users with appropriate credentials;
- A virtual machine infrastructure for developing, testing and running new informatics packages across the aggregated data sets;
- Tracking of downstream analysis outputs, including files of variant calls, mutation signatures, transcriptome features etc., with portability to external databases, such as COSMIC, EGA, dbGAP, Ensembl or the UCSC Genome Browser.

Users who generate cancer genomes and associated data will be invited to participate in the marketplace in return for access to compute resource and algorithms to run on their data. We will develop systems for handling diverse classes of associated data, including histology and radiology images linked to the patient; information on clinical features and treatment response; transcriptome and epigenome profiling. Data generators will be able to access a range of applications to run on their data – this might include variant calling algorithms, annotation, specialised analyses (telomere length, viral insertions), aggregated analyses (gene recurrence, mutational signatures) and integrative studies (transcriptome, epigenome). Bioinformaticians will be able to offer implementations of their applications to other users in the QPQ infrastructure. They will be able to access both the aggregated datasets and embedded compute storage / nodes for their research programmes.

Framing Genomics England cancer data within Pan-Cancer QPQ

We understand that the raw sequencing and clinical data generated by Genomics England will be housed behind a secure firewall in the UK, and that only analysed and summarised data (such as variant call files or data aggregated across individuals) can leave the firewall. These constraints arise from several logistic, regulatory and legal precepts relevant to the UK and the NHS. In the current ICGC pan-cancer project, we have encountered similar barriers to export of raw data in several countries, most notably USA and Germany. Our solution to this in ICGC, and what we would propose here for the Genomics England data, is to engineer a federated computational architecture that enables the initial analyses requiring access to raw sequencing data to be performed within the individual jurisdiction, with only summarised data being exported. By replicating this architecture across each jurisdiction, we enable the analyses to access all available cancer genomes internationally while respecting local rules and restrictions. One of the critical aspects to this endeavour is the process of user certification. For the current pan---cancer project, this certification has been provided by two organisations – the Data Access and Compliance Office of the ICGC and the NCI for TCGA data. The Global Alliance for Genomic Health (GA4GH) is addressing the legal and regulatory barriers to cross---border sharing of genomic data, with a view to making this certification more streamlined and consistent. We will work with Genomics England to build a certification process for researchers within the Pan---Cancer QPQ project that respects the local rules but harmonises with international ethical standards for open data access.

Education and training

There is an urgent need to train a new generation of bioinformaticians conversant in the world of

big data. A Pan Cancer GeCIP would be an excellent training ground for aspiring computational biologists. Behind the scenes, many of the infrastructural and logistic challenges of aggregating cancer genome sequencing data on such a large scale will have been solved, providing trainees with a system that is flexible, open, large---scale and powerful. Gathered in one federated system, the world's reference database of cancer genomes will enable research questions to be pursued with appropriate statistical power. The network of researchers involved in the Pan-Cancer GeCIP will provide strong mentorship to trainees. The senior members of the Pan-Cancer GeCIP have committed to providing opportunities for doctoral and post-doctoral computational biologists in their groups to work on Genomics England data. These trainees will be included in the face-to-face meetings, occurring twice a year, of the ICGC Pan-Cancer consortium, as well as monthly teleconference calls to discuss projects and progress. Through these interactions, not only will trainees get access to the UK leaders in cancer genomics, but also internationally recognised bioinformaticians.

ICGC Pan-Cancer QPQ – Bringing international collaborators & funding to GEL data Under the auspices of the ICGC, we are building a virtual marketplace, Pan-Cancer QPQ, for sharing and analysing cancer genome data. We will develop a funding model and formal constitution to govern the organisation and are seeking research grants from the Wellcome Trust, Cancer Research UK (CRUK), National Cancer Institute (NCI) and others. These funding agencies will fund locally, and we will need to build a federated system that enables the local funding to be leveraged for international data sharing and analysis. While not feasible in the short-term, we aspire to drive the Project towards self-sufficiency over time through building competitive pricing models for access to computational resources.

Interactions with other GeCIPs, cancer researchers, clinicians and industry

We envisage that the Pan-Cancer GeCIP will be predominantly a computational programme. We will develop new algorithms, statistical models and data portals for handling the slew of genomic and other data that will emerge from the Genomics England initiative. We believe that the resources thus developed will be of considerable interest to individual tumour type GeCIPs – for example, an up-to-date census of driver mutations in each patient's tumour will be derived and refreshed on an ongoing basis as data arrive in Pan-Cancer QPQ. This census will be especially useful to, say, the breast cancer group as it begins to analyse which driver mutations predict treatment response or patterns of relapse. By building in rapid dissemination of results to data portals, such as COSMIC20 and Ensembl, we can ensure that mutation calls reach the broad clinical and scientific community efficiently. A strong thread of the current ICGC pan-cancer project is that of improving data visualisation – in our discussions about the on-going improvement of COSMIC, for example, industry users have given useful and strong advice about what functionality they would like to see in the interface. We aim to continue this engagement with downstream, non-expert users of caner genome data.

Full proposal (total max 1500 words per Subdomain)	
Title *(max 150 characters)*	**Cross-cancer analytical subdomain to identify rare and noncoding cancer driver events.**

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 idea of an analytical subdomain specifically focused on the detection of rare and noncoding drivers comes from the observation that at present ~20% of sequenced cancer genomes have no somatic alterations in known cancer genes. We thus expect that the driver events of a considerable fraction of cancer samples sequenced in the context of the 100k Genome Project will remain undetected. The complete map of cancer driver mutations requires dedicated efforts towards the development of analytical approaches for the detection of rare and noncoding events.

This subdomain will bring together six scientific groups already actively working in the field of cancer genomics (Van Loo, de Rinaldis, Buffa, Coolen, Ponting and Ciccarelli). Our expertise will be further complemented by the contribution of International collaborators, who have been widely involved in large cancer genomics projects.

Some of us have contributed to the BRC-pilot phase and we will have direct connections with two GMCs, namely South London and Oxford GMCs. Members of this domain have been invited to contribute to other disease-oriented and analytical GeCIPs and this will favour a coordinated and complete analysis of the cancer mutational landscape.

We propose to build a multidisciplinary and cross-cancer analytical subdomain dedicated specifically to the identification of rare and/or noncoding driver genomic events. Rare drivers are somatic modifications of the genome that contribute to cancer initiation and/or progression in a small number of cases and that cannot be captured using recurrence-based analytical approaches. Similarly, the contributions of noncoding mutations to oncogenesis are currently poorly understood and require dedicated investigation.

The subdomain will bring together computational cancer biologists, mathematicians and statisticians with different and complementary skills from three leading UK Institutions, namely King's College London, CRUK London Research Institute, and University of Oxford.

Two of these Institutions (King's College London and University of Oxford) are involved in the pilot phase of 100k Genome Project and have direct connections with two Genomic Medicine Centres (South London NHS GMC and Oxford NHS GMC). This ensures tight links between our domain and the clinical leads, the sample preparation facilities and the repositories of cancer tissues and clinical data.

Collaborations will be established with International leaders in cancer genomics (Ben Raphael, Brown University, Kevin White, University of Chicago, and Jan Korbel, EMBL) who will complement our expertise. Members of this group have been involved in the constitution of other GeCIP domains (see below) and this will favour a coordinated and complete analysis of the cancer mutational landscape

Full proposal (total max 1500 words per Subdomain)	
Title *(max 150 characters)*	**Pan-cancer drug therapy toxicity**

Research plans. *Give details of the analyses and experimental approaches, study designs and techniques that will be used and timelines for your analysis. Describe the major challenges of the research and the steps required to mitigate these.*

Aim: The aim of the Pan-Cancer Drug Therapy Toxicity Genomics England Clinical Interpretation Partnership (GeCIP) Consortium is to create a collaborative environment in which consortium members with an existing interest in cancer drug therapy related toxicities can work together to utilise the national resource that the 100,000 Genomes Project will provide. The consortium has

representation from many Genomic Medicine Centres, and the NRCI Clinical Study Groups.

The consortium brings together a combination of expertise, from multiple disciplines including clinical and non-clinical researchers, and patient representatives with the requisite skills to implement all aspects of research using data from across the cancer GeCIPs. In addition several groups have existing well-annotated independent data sets which can be used to validate rapidly any findings. The consortium will work with those interested in other pan-cancer phenotypes where appropriate.

Treatment-related toxicities may be functionally disabling, particularly in patients receiving chemotherapy, who may require dose reductions and early cessation of treatment. In addition, with certain toxicities the NHS incurs a financial burden and there are wider economic implications due to loss of working days for the patient.

Genetic determinants of inter-individual variation in drug responses have been established for many drug treatments. The 100,000 Genome Project presents an excellent opportunity to characterise both established and novel genetic factors that affect both efficacy and toxicity by facilitating the collection of high quality cancer drug treatment toxicity data across multiple tumour types, many of which are treated with the same, or similar, drugs.

While adverse effects to drugs have been shown in some instances to be a potential surrogate for efficacy, there is a lack of clarity about the relationship between efficacy and toxicity. Again this project affords the opportunity to explore this relationship across different tumour types and for multiple cancer treatments for both acute (as described in sub- domain 2 and 3) and late onset (leukaemia, cardiac) toxicities. Pharmacogenetics will play an important role in the individualisation of cancer therapy.

The proposed Pan-Cancer Drug Therapy Toxicity GeCIP subdomain will be organised into seven key areas:

1. Clinical Data Collection and Interpretation

The 100,000 Genome Project offers the opportunity to transform the rigor of data collection and reporting of treatment induced toxicities within the NHS. It will provide a platform to develop the optimal methodology and standard operating procedures for reporting and interpreting genomic information related to treatment toxicity.

a. Clinical reporting

We will work with individual tumour site GeCIPs that have developed subdivisions of their GeCIP that advise on the data to be fed back to individual clinicians and patients. This process will evolve as 100,000 Genomes Project progresses. Members of the Pan-Cancer Drug Therapy Toxicity subdomain are also involved in tumour site specific GeCIPs including breast, ovarian, colorectal and haematological cancers. This will allow for the development of a cohesive and collaborative approach to feeding back information to the appropriate tumour GeCIPs regarding relevant pharmacogenetic data. Additional tumour site GeCIPs will be approached in due course.

b. Clinical data collection

The Pan-Cancer Drug Therapy Toxicity subdomain will work with the relevant tumour site GeCIPs and Clinical Studies Groups to ensure that all relevant clinicopathological data are collected on patients. This will ensure adequate data availability for future potential research applications and

facilitate future data interpretation for routine clinical use. Grading of chemotherapy toxicity and cancer drug therapy generally in the UK, is most commonly completed using the National Cancer Institute (NCI) Common Toxicity Criteria for Adverse Events (CTCAE).

We will focus on germline variants associated with toxicities in the commonly used cancer treatments and ensure that the phenotype data collected are broad enough to encompass toxicities from developing/new treatments. The list of toxicities within the sub-domains is not exhaustive and maybe expanded after discussion within the consortium. The sub-domains are defined below.

2. Chemotherapy

a. Chemotherapy-induced Neuropathy

Chemotherapy agents such as taxanes and platinums have been in routine use in cancer therapy for various cancers for many decades. They are used in the neo-adjuvant, adjuvant and metastatic setting. The rates of severe (CTCAE grade 3 or worse) neuropathy which causes significant functional deficit range from ~5-10% and moderate (CTCAE grade 2 or worse) neuropathy occurs in 20-50% of patients in a variety of cancers. Several groups have investigated candidate genes and single nucleotide polymorphisms (SNPs) associated with toxicity related to a variety of neuropathic drugs. But often the sample sizes used in studies have been too small to allow for a definitive association to be found. This will require collection of clinical data on patients with no toxicity and those with grade 2 or above, as functional deficit in present in these patients. In addition the risk of long-term neuropathy may be increased given that a proportion of patients develop irreversible neuropathy.

Chemotherapy-induced Neuropathy will focus on:

- Analysis of germline data to verify existing work on candidate genes and common variants associated with taxane-related sensory neuropathy (TRSN) already identified by the PGSNPS (Abraham/Caldas) breast cancer chemotherapy study, in conjunction with validation work already underway with the TACT (Orr) trial
- Analysis of germline data to verify common variants identified within a TRSN meta-analysis is currently underway involving PGSNPS and three large US clinical trials patient cohorts (~4400)
- Investigation of rare deleterious and/truncating variants associated with TRSN in the ~4000 breast cancer cohort collected through the 100,000 Genomes Project. This will be validated in the existing patient sample cohorts.
- For each chemotherapy agent that causes neuropathy we will identify new common and rare variants associated with severe and moderate neuropathy, in breast cancer, ovarian cancer, lymphoproliferative disease and other tumour cohorts in which these drugs are used. We will also compare germline variants that increase neuropathy risk across different tumour types for chemotherapy agents used in multiple sites.

b. Chemotherapy-induced Neutropenia

Chemotherapy-induced neutropenia (CIN) and febrile neutropenia (FN) are serious and frequent complications in patients receiving chemotherapy. CIN and FN can result in hospitalisation and chemotherapy dose reductions or delays impacting on treatment outcome and short-term mortality. In addition CIN creates an additional financial burden on the NHS due to the complications associated with CIN. For CIN we will consider extreme phenotypes. This will require clinical data on patients with no toxicity and those with grade 3 or above, as these patients require hospitalisation.

Chemotherapy-induced Neutropenia (CIN) will focus on:

- Analysis of germline data to verify existing work on candidate genes and/ common variants associated with CIN already identified (Pettengell; PGSNPS (Abraham/Caldas) breast cancer chemotherapy study
- Analysis of germline data to verify common variants identified within a CIN meta-analysis currently underway involving PGSNPS, in collaboration with the Japan Biobank cohorts (~2700 breast cancer; ~6000 all cancers (Abraham/Caldas)).
- Investigation of rare deleterious and/truncating variants associated with CIN in the ~4000 breast cancer cohort (Abraham/Caldas). This will be validated in the existing patient sample cohorts.
- Pettengell in collaboration European colleagues (Wildiers et al) have investigated treatment related factors and genetic factors predicting neutropenia in a cohort of breast cancer patients (~900). Neutropenia was reported to be associated with specific SNP allele variants.
- Pettengell et al has also investigated the impact of neutropenia as part of the chemotherapy-European Study Group (INC-EU). This work included a multivariate analysis of febrile neutropenia occurrence in patients with non-Hodgkin lymphoma.
- For each chemotherapy agent that causes CIN we will identify new common and rare variants associated with severe CIN, in breast, ovarian, haematological and other tumour cohorts in which these drugs are used. We will also compare germline variants that increase CIN risk across different tumour types for chemotherapy agents used in multiple sites.

c. Cardiotoxicity

Anthracyclines remain a key component of chemotherapy regimens for breast cancer and lymphoma but their use is associated with significant cardiotoxicity. Anthracycline-induced cardiotoxicity (ACT) is reported to be present as an asymptomatic cardiac dysfunction in up to 57% of the patients and the cumulative incidence of acute cardiotoxicity (during treatment) was 3 to 5% in patients treated with anthracyclines. Long-term follow-up studies have demonstrated cardiotoxicity in up to 70% of children treated with anthracyclines. The severity of cardiotoxicity ranges from a mild, transient condition with asymptomatic electrocardiographic changes, e.g. arrhythmias, to more severe toxicities including left ventricular dysfunction and congestive heart failure. The underlying pathological process in ACT seems to be irreversible, due to dose-dependent death of heart muscle cells leading to cardiac remodelling, mainly hypertrophy of the left ventricle which causes myocardial wall stress and contractile dysfunction and eventually cardiac failure.

Although a linear dose-response relationship with ACT has been suggested there is clearly a patient-dependent dose-response relationship. Much less is known about pharmacogenetic risk factors, but small studies have addressed the influence of single-nucleotide polymorphisms (SNPs) on cardiotoxicity. Trastuzumab impairs myocyte shortening, blocking the action of HER2 on cardiomyocytes as well as impairing cardiac stem cell differentiation and formation of microvascular networks. Identifying patients at risk is key as if identified early this toxicity may be reversible.

Cardiotoxicity will focus on:

- Centres involved in the St. George's cardiology study (Pettengell/Sharma/Gaze/Kelleher) will provide information from 2D and 3D echos, which will be correlated with WGS data collect via 100,000 Genomes Project. The study will initially run at St. George's and then be opened to recruitment from other interested centres after appropriate ethical approval. This would permit investigation of known biomarkers associated with anthracycline-related cardiotoxicity and also provides the potential to discover new variants common or rare associated with the same phenotype. These findings may be validated in previous clinical trials cohorts from breast

cancer patients treated with anthracyclines and/ Trastuzumab. Anthracyclines are used in a variety of cancers and the 100,000 Genomes Project would allow investigation of cardiotoxicity in different tumour sites and between male and female patients.

- Earl/Abraham have clinical data and samples (~3800) who have received adjuvant Trastuzumab and/ athracyclines as part of a clinical trial in whom validation studies could be completed.

d. Chemotherapy-Induced Fatigue and Cancer-Related Fatigue

Chemotherapy-induced fatigue is a common but poorly researched area of pharmacogenetics. We have investigated the common variants associated with chemotherapy-related fatigue in two independent clinical trial cohorts (~4000 patients (Abraham/Caldas; Orr)). The data from the 100,000 Genomes Project will be used to investigate common and rare variants associated with this phenotype and the cohorts mentioned may be used as validation sets.

Identification of genetic variants associated with an increased risk of severe fatigue, sleep disturbance and pain (i.e. oncology patients at risk of experiencing high chronic symptom burden) is currently under investigation by Miaskowski et al. They are funded by the National Cancer Institute (USA) to look for common variants associated with these phenotypes using a genome-wide association study. The 100,000 Genomes Project would enable validation of those findings and enable the possibility of identifying new variants, which may be cross validated.

It is the intention of the consortium to approach all the GeCIPs for each tumour site including prostate, colorectal and lung. We have already agreement from members of the breast, ovary and haematological GeCIPs for involvement in the collection of toxicity data.

3. Endocrine Therapy

a. Germline Genetic Variation contributing to Tamoxifen and Aromatase Inhibitor efficacy and toxicity

There are many publications investigating the relationship between genetic variants and efficacy of both Tamoxifen and Aromatase Inhibitors (Abraham/Caldas/Pharaoh). But despite many studies the results are often conflicting and inconclusive. The 100,000 Genome Project will allow investigation of common and rare variants associated both toxicity and efficacy. We have available cohorts in which any new variants identified may be validated (PGSNPS - Abraham/Caldas; SEARCH - Pharoah). Using a large cohort (~3000) from an observational study (SEARCH) we have assessed the association of genotype for candidate genes associated with tamoxifen metabolism with survival and have shown for example that SNPs in the gene CYP2D6 alone do not affect tamoxifen efficacy.

4. Chemotherapy Toxicities and Clinical Outcome

a. Toxicities as Potential Surrogate Markers of Clinical Outcome

We have investigated the relationship between clinical outcome and chemotherapy-related toxicities in ~6000 breast cancer patients (Abraham/Caldas). This work has identified toxicities associated with clinical outcome. The clinical data from the 100,000 Genomes Project would help to validate these results and also investigate the same questions in other cancer cohorts.

b. Non-genetic factors increasing toxicity risk

There is evidence to suggest that non-genetic factors also contribute to the risk of treatment related toxicities. Using data from analysis in existing sample cohorts (listed Appendix B), non-genetic factors associated with an increased risk of specific toxicities have been identified. For each toxicity specific groups within the consortium will investigate these factors before

developing any related statistical models: neutropenia (Pettengell et al/Abraham/Pharoah); neuropathy (Abraham/Pharoah); cardiotoxicity (Pettengell/Rajan/Gaze/Kelleher); fatigue, sleep, pain (Ream et al) and chemotherapy-related fatigue (Ream et al/Abraham/Pharoah). Many of these groups have existing data on phenotypes For example; Risk factors for febrile neutropenia including: older age, lower weight, higher planned dose of chemotherapy, higher number of planned chemotherapy cycles, prior chemotherapy, performance status the number of co-morbidities, lower baseline white blood cell count (WBC), lower platelet and neutrophil count and higher baseline bilirubin, abnormal liver or renal function, low WBC, higher chemotherapy intensity and planned delivery were identified as risk factors for neutropenic complications in a prospective US study of patients with different types of cancer. Poor performance status and low lymphocyte and neutrophil counts were risk factors in a European study of solid tumour patients, as were tumour stage and number of co-morbidities in elderly patients with solid tumours.

5. Patient-related toxicity reporting

We are aware that involving patients more often in their own care is empowering for the patient and useful for clinicians. Ream, Kearney et al have secured a substantial FP7 programme grant known as eSMART, to refine and evaluate across Europe (England, Ireland, Austria, Norway and Greece) a system for remote monitoring and managing symptoms during chemotherapy. The system has previously been tested during both chemotherapy and radiotherapy. It was developed by a number of the e-SMART collaborative team in conjunction with cancer clinicians and people with cancer. It is a mobile phone based remote monitoring system that enables real time measurements of a patient's symptoms, facilitating immediate tailored management of cancer treatment-related toxicities in the home care setting, and automatic and immediate triaging of care where patient symptoms exceed clinical norms. Patients using this system complete an e-symptom questionnaire on a mobile phone and take their temperature twice a day and at any time they feel unwell, and send this information in real time to the study server. Based on evidence-based algorithms, symptom reports that are of clinical concern trigger alerts to cancer care clinicians based in acute care, who on receipt of an alert can view the patient's symptom reports on a secure web page, and contact the patient directly at home, facilitating the initiation of real-time clinical interventions. Patients immediately receive evidence-based self-care advice on the mobile phone (text, audio, video) for the specific symptoms that they have reported, facilitating self-management of symptoms.

In the cohort of UK patients within the eSMART study, the trials team in each recruiting centre will flag to the central study team if a patient has be enrolled in both eSMART and the 100,000 Genomes Project. The eSMART study cohort will have excellent patient-led follow up data allied with clinical and WGS data (from the 100,000 Genomes Project), which would be unique as currently no other study can offer that.

6. Statistical Modelling

A key aim for this consortium is the development of predictive models for chemotherapy toxicities in particular. There is strong expertise in this area for example:

- Schwenkglenks/Pfeil: Are developing risk models for the occurrence of neutropenia in the first cycle and any cycle or chemotherapy, initially in breast and lymphoma patients receiving chemotherapy, based on patient- related, chemotherapy-related and genetic characteristics. The study will correlate genotypic data with chemotherapy data from the SACT database and both neutropenic and non-neutropenic related infection.
- Pharoah/Abraham: Have been developing polygenic risk scores which incorporate the genetic variants found from candidate genes studies and genome-wide association studies into a predictive model which includes non-genetic factors which increase the risk of toxicity, for phenotypes such as neuropathy and neutropenia.

These models will be expanded in other toxicity phenotypes, if appropriate. Both groups have existing validation cohorts in which to validate their models and also to validate any models developed using the 100,000 Genome data.

7. Functional characterisation of novel germline variants that influence pharmacogenetic phenotypes

Translation of genetic discovery into functional studies in vitro and in vivo, to confirm functional significance and identify variants with potential therapeutic implications will be required.

Contrary to initial expectations, very few causal alleles that confer susceptibility to common cancers localise to gene exons. Instead most alleles for which sufficient evidence exists to support causality map to regulatory regions of the genome, e.g. transcription factor binding sites or enhancer elements. However, the extent to which this is true for germline variants that influence pharmacogenetic traits is not yet clear.

Orr and others will functionally annotate, using bioinformatics approaches, sets of correlated variants that are identified from association analyses of pan-cancer chemotherapy induced toxicities in order to identify SNPs with putative regulatory potential. These SNPs will then be functionally characterised in appropriate model systems using established methods, thereby confirming whether they are indeed causal and illuminating the underlying mechanisms by which they mediate toxicity.

Training

This is an important part of the GeCIP. Members of the Pan-Cancer Drug Therapy Toxicity GeCIP are involved in the cross-disciplinary oncogenetic training (Copson/Abraham) for clinicians including interpretation and communication of genomic data to patients. Dr Ellen Copson will be the training director. She chairs the medical oncology SAC onco- genetics training working party. In conjunction with the tumour GeCIPs, in particular breast cancer, we will develop training for trainees based on resources from the 100,000 genomes dataset particularly aimed at identifying the oncogenetic training needs. The availability PhD, Clinical Fellows and Post Doctorates to work on the 100,000 Genome Project will be set up in conjunction with the tumour group GeCIPs.

Funding for research

The academic members of the consortium have extensive existing funding, and we will be looking to apply for structured substantial funds from GeCIP funders (MRC, Wellcome, NIHR, CRUK), and/or other funders including European funders, to facilitate the research ideas presented in the outline application.

Industry Partnerships

We will utilise the existing strong working relationships that have already been established among many of the consortium members to develop mechanisms for pre-competitive interaction with partners from industry.

Engagement of patient representation groups

The consortium considers patient involvement an essential part of the GeCIP. Both patient and clinical advisory groups working alongside many members of the consortium and in the projects sited in this expression of interest. For example the model of care proposed through the eSMART programme of work has been developed from patient experience in collaboration with the

European Cancer Patient Coalition (ECPC). We will utilise this type of existing resource within our consortium to ensure that patients are engaged with the aims and development of the Pan-Cancer Drug Therapy Toxicity GeCIP.

Full proposal (total max 1500 words per Subdomain)	
Title *(max 150 characters)*	**Clinical and therapeutic significance of the HOX gene clusters in cancer**

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 HOX genes are a family of transcription factors characterized by highly conserved DNA- and co-factor binding domains. This conservation has been driven by their roles in some of the most fundamental patterning events that underlie early development. Most notable of these is the patterning of the anterior to posterior axis, for which a precise spatial and temporal order in the expression of HOX genes is required. This is achieved in part through a chromosomal arrangement whereby HOX genes are present in closely linked clusters allowing the sharing of common enhancer regions. In mammals there are four such clusters (A-D), containing a total of 39 HOX genes. The relative position of each HOX gene 3' to 5' within the cluster is reflected in a number of key attributes, including the spatial and temporal order of expression, whereby the 3' most genes are expressed earlier than their 5' neighbors. The nomenclature of the HOX genes reflects this precise chromosomal ordering, with members of each cluster being numbered with respect to the 3' end, thus for example, the 3' most member of cluster B is HOXB1.
HOX genes are strongly dys-regulated in cancer, and generally exhibit greatly increased expression. This differential change in expression in cancer may reflect the apparent ability of some HOX genes to function as tumor suppressors and some as oncogenes. Thus for example, HOXA5 acts as a tumor suppressor in breast cancer by stabilizing P53, whilst forced expression of HOXB6 can immortalize fibroblast cells.
The dys-regulation of HOX genes has been demonstrated in a range of cancers, and is associated with a number of important clinical and pathological parameters, including survival and response to treatment. Genetic changes within these clusters are considerably more likely to have functional consequences as the genes are relatively very close and share overlapping enhancers. Despite this, genetic changes in the HOX gene clusters have never been studied in relation to cancer. Therefore, the work proposed by this domain includes the following:

1. The assessment of possible genetic differences in the HOX clusters in patients who develop a range of cancers including those of the breast, ovary, prostate, pancreas, head and neck, bladder, lung, brain and digestive tract, together with melanoma, leukemia, and lymphoma. We will compare the HOX clusters of these patients to those of aged-matched controls who do not develop cancer. Initially the preferred tumor types in this study will be breast and prostate.
2. In addition to identifying risk factors for cancer, we will also use longitudinal disease data, where available, to establish the prognostic significance of this genetic background.
3. Follow up studies will assess whether genetic differences in the HOX clusters actually relate to changes in HOX expression using lab based studies with relevant cell lines.

Additional research
We will use the longitudinal data to evaluate whether specific changes in the HOX clusters relate to specific types of risk. For example, some HOX genes are involved in DNA repair, and might have a more significant role in tumor suppression in certain cell types.

Funding
We currently have funding in place to assess the role of HOX genes in cancer, including grants from the EU, the Breast Cancer Campaign, the British Lung Foundation and a number of commercial sources. Data analysis would initially be handled by statisticians working within this framework. We would seek additional funding from these sources, as well as a number of other funding sources to extend our analysis of patient genomes in order to identify key prognostic and therapeutic markers.

Opportunities for education and learning
Our groups are actively involved in the training of specialist haematology, medical oncology, and clinical oncology specialist registrars, MD, MSc and PhD students, and postdoctoral scientists to allow them to develop the key skills needed for the analysis of large data sets. Access to the 100K genome project data would help to facilitate and drive this. In addition, all future grant proposals will contain a specific training element to fund this aspect of student and postdoc learning. In total the group supervises over 20 PhD students and has seen a total of 35 through to completion.
RP will take the lead on clinical training in the context of the genome project. She has extensive experience, including:

- CRUK Protocol review Committee
- London Cancer New Drugs Group (LCNDG): ICDFR
- Educational Supervisor for 2 ST3, ST4
- Lectures London Oncology Course (3rd year) / GSTT Clinical Radiology MSc Course
- Lymphoma Association Medical Advisory Panel (2000 -)
- Member of the NCRI Aggressive Lymphoma Clinical studies Group (2004-)
- Chair of the INC-EU (Impact of Neutropenia in Chemotherapy European Study Group)

International collaboration
We have a number of international collaborators who work on HOX genes in cancer with whom have produced a number of joint, high profile publications. These collaborators include:
Dr Alessandra Care, Director of Research, Instituto Superiore Sanita, Rome, Italy (academic) Prof Manuel Penichet, Consultant Hematologist, UCLA, USA (academic, clinician)
Dr Hannes Klump, Consultant Hematologist, University of Essen (academic, clinician)

Pre-competitive industrial collaboration
We have close links to a number of companies interested in developing HOX-targeting agents for cancer, and who would be willing to help fund the analysis of genomic research to better understand patient needs and stratification.

Full proposal (total max 1500 words per Subdomain)	
Title *(max 150 characters)*	**Radiogenomics (genomics of normal tissue toxicity from radiotherapy)**

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.*

There are ~2 million cancer survivors in the UK, a figure predicted to rise by >3% per year. Nearly half have undergone radiotherapy (RT). RT is used to treat >100,000 patients with curative intent in the UK each year, including many with four of the solid malignancies covered by the 100K Genomes project: lung, breast, prostate and colorectal. Many patients suffer with moderate or

serious long-term side effects. For example, about a third of prostate cancer patients can experience diarrhoea, urgency, rectal bleeding or urinary incontinence. A smaller proportion suffers severe toxicity. Toxicity is not recorded well in routine clinical practice, so that a simple review of treated patients is not sufficient to define a RT toxicity phenotype.

RT schedules treat population averages, accepting that ≤5% of cancer survivors suffer severe long-term toxicity in order to deliver sufficient RT dose. This 5% minority limits the potentially curative radiation doses given to the majority. Methods to increase RT doses without increasing toxicity will allow more patients to be cured, and identifying the sensitive minority could reduce the toxicity burden in that group [Barnett et al 2009, Kerns et al 2014b]. There is a need to increase understanding of the genetic variation that determines a patient's risk of toxicity to:

1. underpin the development of tests to predict risk of toxicity to enable individualised radiation dose prescription leading to reduced toxicity and increased cures;
2. identify novel potentially druggable targets for preventing or mitigating radiation effects.

We will use the genome data from the 100,000 Genomes patients with four cancer types: breast, lung, prostate and colorectal. Many of the patients will have received RT as part of their treatment. The exact number will depend upon the tumour stages and patient characteristics of those recruited into the study. The 'Minimum Data Sets' for the 100K genome project includes External Beam Radiotherapy data covering radiotherapy site, dose, type and fractionation. However, it does not include toxicity data or data on other potential modifiers or confounders of the relation between genotype and phenotype (RT toxicity).

Using our expertise from UK RAPPER (Radiogenomics: Assessment of Polymorphisms for Predicting the Effects of Radiotherapy) and REQUITE (see http://www.requite.eu/ for data collection forms), we will establish the collection of toxicity data, to add to the Minimum Data Set, and will ensure that relevant data on modifiers and co-morbidities which can influence genotype-phenotype associations are collected. We will work closely with the domains for each cancer type and the Genome Medicine Centres (GMCs) to avoid repetition but ensure maximum value is gained from the available data. The most rigorous way of collecting these data is for a trained clinician/healthcare professional to score the patients before and after treatment using a standard scoring system such as the Common Terminology Criteria for Adverse Events (CTCAE) v4.0 [Trotti et al 2003]. Retrospective collection of CTCAE data has also been used successfully in Radiogenomics studies but prospective collection is preferred. The importance of patient reported toxicity is also recognised. We are, therefore, in a position to provide expertise and resource to support the collection of additional data to add to the richness and value of the overall GeCIP data set.

RAPPER currently recruits patients from 56 clinical centres in the UK, including the majority of the GMCs (6,650 samples to date, target 10,000). In brief, cancer patients within the GeCIP will be identified and invited to participate through their GMC. Clinical assessment will be made either at time of radiotherapy planning or at routine follow-up by a clinical oncologist, or another trained member of the medical team. Lung and colo-rectal cancer patients will be assessed at one year post-radiotherapy, breast and prostate patients after a minimum of two years. Where the opportunity exists (for example for prostate cancer patients also entering the VoxTox study) data will also be obtained at intermediate time points. Our experience with toxicity scoring (Davidson, Barnett) and electronic data collection systems (Davidson, Burnet, West, Brookes) will underpin this activity.

- Primary end-points will be: breast (induration), lung (dyspnoea), prostate (rectal bleeding), and colorectal (diarrhoea).
- Secondary end-points will include: breast (overall STAT score, cutaneous telangiectasia, atrophy, pigmentation), lung (fibrosis, rib fracture), prostate (diarrhoea, urgency, stricture,

erectile dysfunction, colorectal cancers (erectile dysfunction in men, urinary incontinence, and bowel problems).

Current practice in Radiogenomics is to use as phenotypes both individual toxicity end-points and aggregated measures of organ-specific or overall radiosensitivity, using methods such as STAT score, developed by us [Barnett et al IJROBP 2012]. This is necessary because it is expected that some genetic variants will affect the response of multiple tissues to radiation e.g. ATM [Tanteles 2012] and TNF [Talbot 2012], while others will affect single endpoints [Barnett 2014]. This is also true across cancer sites, which emphasises the importance of studying RT toxicity in a unified fashion as part of a single cross-cutting GeCIP domain.

The genetic trait of interest is radiosensitivity, reported as RT toxicity, for which there is evidence of a heritable component. It is a complex and polygenic trait; the allelic architecture will include very rare mutations with large effects (e.g. homozygous mutations in ATM and other DNA damage response genes), less rare mutations with moderate effects and common variants – single nucleotide polymorphisms (SNPs) - with smaller effects. As with studies of genetic variation in other complex traits, very large studies are required for replication of potential associations [Barnett et al Radiother Oncol 2012, Barnett et al 2012c]. The co-applicants have carried out the largest Genome Wide Association Study (GWAS) in the field involving 1,853 samples. They are also directing a meta- analysis of internationally available GWAS datasets to search for further tissue-independent variants. Included studies have collected pre-treatment toxicity data, and potential confounding risk factors that influence toxicity (e.g., smoking, diabetes, age). Given the currently unclear genetic architecture of radiotherapy toxicity, we will study the full range of germline polymorphisms including common variants, found via GWAS and rarer variants that will be discovered via sequencing studies such as the 100,000 Genomes Project

Known predictors of radiotherapy toxicity will be incorporated into a regression analysis and the residual radiosensitivity used as phenotype in the genetic analysis. These predictors include radiotherapy dose and fractionation, co-morbidities (diabetes, collagen vascular disorders), chemotherapy, age, smoking and anatomy (e.g. breast size). Where these data are not already available, collection will be included in future funding applications.

Given the number of patients for each cancer type in the project there will be insufficient power to conduct a genome-wide association study without the addition of further cohorts for replication. The participants have access to the largest Radiogenomics cohorts globally through the RAPPER and REQUITE studies, and via the RGC. Collaboration will provide substantial added value to the GeCIP.

Proposed additional research activity

The radiotherapy toxicity sub-domain will collect longitudinal toxicity data and will include Quality of Life questionnaires which could be of use to other domains. The data will include that from the radiotherapy planning systems such as organ-specific dose-volume histograms.

Other biomarkers have been evaluated for predicting radiotherapy toxicity including proteomics using blood plasma, transcriptomics and epigenetics in lymphocytes. Other studies on tumour response have used epigenetics or transcriptomics. The REQUITE project is collecting RNA samples from blood for future transcriptomic analysis. We will seek funding to find potential gene expression signatures that predict adverse reactions, using the 100,000 Genome and REQUITE samples as training and test sets.

Current funding and plans for procurement of funding for proposed research activities

Radiogenomics started later than the large cancer predisposition consortium (3 SNPs from ~1,800 patients *vs* ~70 SNPs from ~50,000 patients for prostate cancer) but has a track record in securing

funding. RAPPER was funded by CR-UK (2006-2010). RAPPER was conceived as a sample collection and candidate gene study (Burnet et al 2013). The work involved prostate and breast cancer patients with a goal of identifying SNPs associated with toxicity independent of the site irradiated. This work revealed that individual SNPs have small effects on a risk of toxicity, a finding consistent with GWAS for other traits/diseases (Barnett et al 2012c). With decreasing genotyping costs, we obtained additional funding to carry out a GWAS (RAPPER-GWAS; 2010-2013; C1094/A11728). We then performed the largest GWAS in the field (Barnett et al 2014) and our data were used in RGC collaborative projects (Fachal et al 2014). RAPPER-3 (CRUK 2014-2018) is funding the continuing sample collection. RAPPER led to the establishment of the RGC, and REQUITE (2013-18) was the first RGC funded project. REQUITE has €6 million under the EU FP7 framework to collect data and samples from 5,300 patients with breast, prostate and lung cancer. Additional funding will be sought in the UK from CR-UK or cancer-specific charities (e.g. Breast Cancer Campaign/Breakthrough Breast Cancer, Prostate Cancer UK, Roy Castle Lung Cancer Foundation) and through international collaborations from the US National Cancer Institute and the EU Horizon 2020 programme.

Opportunities for education and training

The Radiogenomics field is interdisciplinary, comprising Clinical (Radiation) Oncologists, Medical Physicists, Geneticists, Epidemiologists and Radiobiologists to work on prediction of RT response. There will be an increasing need for both Clinical and Medical Oncologists to incorporate genomic data into their clinical decision-making. This will necessitate a new generation of trainees who have a sound knowledge of genetic theory and genomic practice.

The Radiogenomics GeCIP domain will provide research training in genomics to this new generation of clinical trainees. This builds upon an established track record of the participants bringing doctors in training into the field.

Training is already well established in our group. Gillian Barnett was awarded her PhD for work on 'Individual variation in late toxicity from radiotherapy' (Barnett et al 2009, 2012a-c, 2014). Laura Fachal obtained a PhD for her work in radiogenomics (Fachal et al 2014). Tim Rattay, a ST5 breast surgeon carried out an Academic Clinical Fellowship in the Talbot laboratory, subsequently winning an NIHR PhD fellowship with Talbot and Symonds entitled 'Predicting acute radiosensitivity in breast cancer'. Kerstie Johnson, a ST7 Clinical Oncologist is employed to carry out Radiogenomics research on the REQUITE project, and is carrying out an MD.

A training programme will be established to (a) hold short courses on Radiogenomics (b) hold networking events bringing together established Radiogenomics researchers and trainees in England. We will seek funding for both clinical and non-clinical PhD fellowships, and MSc projects, to work within the domain. Evidence of success will be an increased number of clinical oncologists registered for MSc, MD or PhD degrees.

Participants in the domain are involved in the assessment for the Fellowship of the Royal College of Radiologists (FRCR) examination. We will work with the RCR to update the syllabus for the FRCR to incorporate the increasing role of genomics in radiotherapy.

Potential international collaborators

The international collaborative group for radiogenomics is well established. The RAPPER group initiated the establishment of the international Radiogenomics Consortium, essential for the collaboration required in radiogenomics [West 2010]. The RGC is led by Catharine West and Barry Rosenstein (letter of support). The RGC has 184 members from 106 institutions in 23 countries and has published the results from several collaborative studies [Barnett et al 2012c, Talbot et al 2012, Fachal et al 2014, Kerns et al 2014). The REQUITE project includes researchers and doctors

from ten clinical centres in Europe (Barcelona, Ghent, Heidelberg, Leicester, Leuven, Maastricht, Manchester, Milan, Montpellier, Santiago de Compostela), plus Mount Sinai School of Medicine in New York. The RGC recently completed its first meta-analysis of GWAS (n~1,800) of prostate patients, which identified 3 SNPs associated with RT toxicity (paper drafted). Letters of support attached from Prof Riccardo Valdagni (Milan, REQUITE prostate lead) and Prof Dirk de Ruysscher (Leuven, lung lead).

Mechanisms for pre-competitive interaction with partners from industry

The participants have existing links with industrial partners. For example Source Bioscience, a UK SME involved in cancer diagnostics and genomics, is a commercial partner in the REQUITE project. Discussions are already underway about potential platforms for clinical implementation of a radiogenomic signature.

Full proposal (total max 1500 words per Subdomain)	
Title *(max 150 characters)*	**Cancer symptomology**

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 subdomain proposes to study broadly cancer symptomology with a specific focus on:

1) Fatigue & rehabilitation in patients successfully treated for cancer: Survivorship.
2) Symptom burden on adjuvant (or curative intent) cancer treatment – including fatigue, pain, breathlessness etc.: Supportive care in cancer.
3) Symptom burden in patients who have metastatic cancer and a poor prognosis (less than a year on average): Palliative care in cancer.

In all of these areas the symptom intensity and impact on functioning is only (sometimes) routinely measured as a subjective symptom.
There is however a range of validated symptom specific outcome measurements and widely used quality of life tools available to monitor cancer patients throughout all areas of treatment and beyond. We feel linking quality of life and other subjective data to specific biological markers will be invaluable in designing more targeted therapies for refractory symptoms, prognostication and for screening patients for earlier interventions for complex symptoms.

Our group is a collaboration of academic and clinical palliative care professionals & oncologists. It is a multidisciplinary collaboration and links South West London teaching hospitals (St Georges, Guys & Kings) through the CLAHRC. These are also the same institutions that have been awarded the south London genomics centre and there is an established biomedical centre funded by the NIHR. We also plan to further promote and expand supportive & palliative care research as a previous collaboration through the National cancer research institute (NCRI) initiative which was designed to promote these under researched areas. This aim has been cemented with the formation of the Cicely Saunders Institute at Kings College London. A world class purpose built institute headed by Professor Irene Higginson who is a NIHR senior investigator.
We would propose in time to link with the Marie Curie Palliative care research unit in time if the outline bid was successful. This unit is headed by Professor Paddy Stone whose expertise is in prognostication and would advise on this aspect of the bid. This would provide a world class pan-London range of palliative care expertise and further augment research capacity in basic symptom biology.

As we stated at the start of our proposal we want to focus on supportive and palliative care outcomes in cancer patients. This moves away from a specific focus on survival or disease free progression which would be the usual oncology trial endpoints. Our focus is on symptoms, functioning and quality of life[3]. Our aim is to conduct further mechanistic research in this area to be able to elucidate the biology of complex symptom clusters. These objectives are linked to both the National end of life care strategy & the National cancer survivorship initiative. There is recognition in both that research is a generally disproportionately underfunded area while recognising the importance of the area per se. The major difficulty of focusing on the symptom and functional aspects of a disease is that often the clinical trials and mechanisms (& therefore funding calls) are focused on the disease process. This has been reflected in the strategy taken by CRUK and the research councils. However patient experience and the management of the side and long term effects of treatment have been consistently highlighted by patients themselves & other groups such as Marie Curie care. Organisations such as the Lind Alliance in conjunction with the NIHR have enabled promotion of top priorities which have been generated by patients and professionals collaboratively. We plan to build on this with our mechanistic studies of symptom biology. Our overall bid in South West London also included Macmillan which was a partner organisation in the survivorship strategy and is intricately involved in improving the cancer experience at all stages of treatment. We want to be able to use Patient reported outcome measures (PROMs) already in use to monitor symptom intensity and progression longitudinally and in response to interventions. We plan to link these to multi-omic outputs for specific symptoms such as fatigue, pain and breathlessness and overall quality of life.

Patients undergoing treatment for cancer are warned to expect a significant number of symptoms related to the treatment including fatigue which is one of the most troublesome. Surviving cancer, however, is not synonymous with a life free of problems related to the disease and its treatment. In addition at a time of relapsed disease or metastatic spread the intensity of these problems will worsen and functional ability decrease as the end of life approaches.

Symptoms should not be viewed in isolation, but rather as a cluster of interrelated symptoms with a large overlap in their aetiology. This seems to be from prolonged inflammatory response to the treatment and to the cancer itself. Our group identified a number of potential biomarkers. Many of the significant analytes were chemokine ligands found to be linked through an inflammatory pathway promoting T cell and granulocyte production and activation. However there is conflicting data dependant on which markers have been pre-selected – one of the major limitations of the research to date. Our group overcame this using in silico pathway analysis to provide plausible biological linkage and combined this with unselected mass spectrometry analysis of depleted plasma. We found both specific interleukins and chemokines such as Granulocyte colony stimulating factor (GCSF), vascular endothelial growth factor (VEGF) and brain derived neurotrophic growth factor (BDNF). We also found nonspecific inflammatory plasma proteins on mass spectrometry - serum amyloid A, collectin, and subunits of immunoglobulin G and complement C1Q. We linked this to clinical data demonstrating poly-symptomology in a group that met a specific clinical phenotype on a diagnostic interview. Our group and others has consistently demonstrated links between subjective symptom and biological markers.
One of the major aims of our proposal would be to confirm this proteomic signature and to further link to clinical data longitudinally. We would also want to cross reference with the presence of other symptom clusters and measure biomarkers in response to a broad range of intervention trials. These would include pharmacological & complex non drug interventions.

Our group has previously suggested that relatively low cost interventions with self-management and e-Health elements may be appropriate for the majority of survivors, with resource intensive

interventions being reserved for those most in need – currently those on treatment or in patients with advanced cancer. The patients with active disease may be more suited to pharmacological interventions provided a targeted mechanism is discovered. However this could also be rehabilitation and exercise based along with cognitive behavioural component. An objective marker collected alongside PROMs would augment our current understanding of these complex symptoms.

In summary the use of self-reported & objective activity together with symptom prevalence ought to be increasingly straightforward with the increased use of technology and using of monitoring on a phone application. Harnessing this data even in the absence of any intervention would provide an objective study of activity during treatment that could be correlated with intermittent patient reported & recorded measures of quality of life and other symptoms. This will be especially important for patients undergoing complex or prolonged treatment regimens (most notably haematological cancers). We are still not readily monitoring fatigue, pain or breathlessness or its relationship with activity during treatment. The link to translational research is still missing. We need to be routinely collecting samples in order to monitor any intervention "dose response" and to identify mediators and moderators of fatigue and other symptoms. These measures should also be included in clinical trials as a far as possible within a core set of patient reported outcome domains. Routine collection of multi-omic data ought to be collected in parallel. The moderate effect size of exercise, complex psycho-social or drug based interventions has been disappointing up to this point. If we can identify the biological mediators and moderators of these symptoms we will be to screen those at risk, intervene earlier and design targeted interventions. This will affect not only quality of life but social, economic & occupational impact.

Proposed additional research activity including research activity focused on multi-omics and/or longitudinal data (as appropriate to domain).

Overall Aim: Establishing a supportive, palliative and end of life care collaborative across South London. The palliative care stream of the CLAHRC aims to establish a science-based, inclusive, focused, well managed collaborative in Palliative and End of Life Care (PEoLC) that leads and conducts applied health research relevant across the NHS, and that translates research findings to improve outcomes for patients and their families. This incorporates patient, family and public engagement**.** This work primarily focuses on the identification of outcome measures and is directly relevant to the on-going identification of symptoms cancer patients experience. The link of this clinically and population based research would significantly augment any multi-omic mechanisms identified. These PROMs & biomarkers would also help to identify suitable patients for interventional clinical trials. The other link is to Macmillan cancer charity and their ongoing investment & evaluation of physical activity on treatment and in both the survivorship and end of life periods. Again identification of suitable patients both for interventions and exercise programs would link in to current research activity in this area. Both of these longitudinal research programs are funded for the next 5 years before review and would be augmented by the multi-omic outputs.

Current funding and plans for procurement of funding for proposed research activities.

There are at present no current funding for multi-omics. However related applications have been to the NIHR, CRUK, Wellcome and Marie Curie cancer care have been made. These organisations would be approached for more funding & we expect there may be more focused calls following this initiative.

Opportunities for education and training.

We are a member of the biomedical cluster together with Guys & Kings. There is an established training program in a range of multi-omic techniques but also related areas of bioinformatics, statistics and public and patient engagement and drug discovery. This program also includes invited expert lectures and encourages collaboration at the doctoral & post-doctoral level.
At St Georges we have the biomics centre which provides bespoke training in the use of all the relevant technologies for scientists and clinicians alike. They also provide training on the relevant hardware and software and links to industry partners to ensure access to emerging technologies and relevant training. At the Cicely Saunders institute there is again a relevant rolling program for training on PPI, statistics, questionnaire development and epidemiology research methods. This would complement the multi-omic training. One of the major advantages of the combined medical school and trust site is the ability for all concerned to be able to access training on site.

Potential international collaborators

We have not made any approach to international collaborators at the outline stage. However our group has links to the survivorship group within the EORTC (European Organisation for the research and treatment of cancer) and the EAPC - European association of palliative care. Outside of Europe our group has collaborated with the ACS – American Cancer Society and again would explore future joint working if the bid was successful.

Mechanisms for pre-competitive interaction with partners from industry. Enabling Invention, Innovation and Impact.

The Enterprise Team at St George's supports researchers in bringing their innovations to market. They can assist with the relevant areas which potentially include: Intellectual Property; commercialisation of new ideas and proof of concept funding. We would hope that any identified symptom mechanisms would lead to more targeted drug therapies and liaison with industry and the relevant drug company's development units.

Full proposal (total max 1500 words per Subdomain)	
Title *(max 150 characters)*	**Cancer Prevention**

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.*

Background

Globally cancer incidence is increasing and given the paucity of effective and affordable treatments for advanced disease research efforts must focus on cancer prevention. We are now in a unique position to attempt to bridge the gap between the vast cancer genomic susceptibility data and to translate this to clinical benefit. The 100K Genome project provides an ideal opportunity to harness integrated genomic and environmental exposure data and identify and develop novel targets and prevention strategies for cancer. As we identify increasing numbers of patients of enhanced risk of developing cancer through these programmes the need for interventional preventative strategies becomes ever more pressing. By comparison with the treatment of advanced cancer, cancer prevention therapies are substantially under developed. The need for these interventions can only increase as the population is stratified by the detection of pre-malignant precursors in a screened population.

Summary

This GeCIP will undertake multidisciplinary integrated cancer-prevention research and will work collaboratively with cancer-specific domains to ensure prevention is central to research

endeavour. Further added value is envisaged through interaction with cardiovascular and metabolism GeCIPs to interrogate shared risk factors and develop novel stratified prevention strategies which will have benefits beyond cancer prevention. This presents an exceptional opportunity to interrogate real-time genomic, transcriptomic, metabolomic and phenotypic data and dissect cancer susceptibility and therapeutic interventions in translational research from in-vivo organoid modelling to human intervention trials.

This subdomain will have key focus areas

1. To ensure that early stage disease is appropriately targeted within each cancer domain, and where appropriate, prompt for inclusion of precursor lesions.
2. To explore, with the different cancer domains, early cancer drivers that may be common between the cancer types.
3. To integrate genomic and environmental data to generate a cancer risk profile that will be used to stratify cancer-preventive interventions.

Major beneficial therapeutic interventions have already been identified demonstrating the importance and potential of therapeutic prevention in cancer. Two important examples being Aspirin and Tamoxifen. As yet the targeting of these interventions to the most appropriate populations has been less explored and this strategy will become more important as new preventative interventions are identified. The current focus of precision medicine in the treatment of established disease will inevitably shift towards disease prevention in forthcoming years. The 100k genome project should be in a position to exploit this. There will be an appropriate and inevitable focus of the cancer GeCIPs on advanced disease and inherited syndromes. A Cancer prevention GeCIP can act to ensure appropriate focus on primary, secondary and tertiary prevention.

Our overarching aim is to utilise whole genome sequencing data as a discovery tool for development of novel targeted therapy strategies, to identify novel potential targets for cancer prevention and to stratify patients for appropriate prevention treatment.

Specific questions that can be addressed across the cancer types within the 100K Genome project include:

1. The identification of novel cancer preventative targets that would be expressed across different cancer sub-types.
2. The detection of molecular susceptibility genes and molecular resistance to cancer preventative therapy (such as aspirin, metformin, tamoxifen).
3. Combining environmental and genomics data to enhance our understanding of cancer initiation and provide novel risk stratification models for primary cancer prevention.
4. Development of secondary/tertiary prevention of cancer:
5. Correlating clinical/histological information with genetic profiles for better risk stratification in the adjuvant setting. Combining this with metachronous metastases data will help provide a genetic basis to underpin the investigation of agents that could provide effective secondary/tertiary prevention that may be applicable across different cancer types.

Management Concepts

1. The cancer prevention GeCIP will comprise multi-disciplinary members providing the necessary spectrum of expertise in terms of clinical expertise, laboratory science, animal models and genomic analysis.
2. The cancer prevention GeCIP will have the capacity to undertake further analysis in separate domain cohorts in the 100k genome project and interrogate candidates from established

cohorts across the UK and beyond.

3. Clinical reporting of data. This will not be a focus of the cancer prevention GeCIP. Reporting will be coordinated through the disease specific domains. Nonetheless the cancer prevention GeCIP will be available on a consultative basis if there are specific issues.
4. The cancer prevention GeCIP will work closely with the National screening committee to coordinate development and evaluation of new prevention strategies.
5. The cancer prevention GeCIP will work closely with the UKTCPN (United Kingdom Therapeutic Cancer Prevention Network). This national organisation is directly linked into the ECMC network (thereby with the NIHR and CRUK) and will enable translation of findings into clinical studies.

Operational plan

1. Sample processing. It is currently not decided whether FFPE samples will be utilised for sequencing. Validation in secondary cohorts will however focus on FFPE. Optimising results from FFPE should be an important part of this programme.
2. Driver gene discovery. Driver gene identification of early stage disease will require complex bio-statistical analysis across different tumour types.
3. The genome data will also allow investigation of germline variants that affect different cancer susceptibility and response to therapy. We aim to translate findings into novel prevention targets focused on prevention.
4. Complementary studies. Further analyses of RNA and protein, and the use of model systems have several potential uses downstream of the genome sequencing. For example, it may inform on function, provide actionable targets, and allow more generalisable driver identification.
5. Cross-cutting analysis. This will be an important focus of the group, who will work across the disease specific domains to identify common genetic factors that influence cancer risk and instigate secondary studies to identify novel therapeutic interventions, particularly those that might benefit a number of different cancer types.
6. Training and education. To ensure long-term durability to the 100k genome project, we will engage with the new MSc genomics programmes and ensure cancer prevention is positioned within the educational programmes.
7. Maximising Impact. We will prospectively undertake discuss strategies to maximise impact from the outset. All GeCIP members will undertake data dissemination to the academic community by presenting research at cancer-specific and cross-disciplinary cancer conferences to maximise exposure to the wider scientific community.

Funding

The UKTCPN is supported by the ECMC. This infrastructure funding will assist the initial development of the GeCIP. Additional support will be sought from the NCRI through SPED and through the national cancer screening committee.

Individual study funding will be sought through the recognised national funding bodies as appropriate.

Key Initial Developments

1. Develop working relationship with each cancer domain. Ensure cross representation on steering committees. Ensure that sample collections encompass early stage disease.
2. Approach National Screening Committee and SPED for a linked strategy
3. Publicise through NCRI cancer prevention symposium in November 2015

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	Pan cancer
Project title *(max 150 characters)*	**Pan cancer 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.