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

| Application Summary | |
|---|---|
| **GeCIP domain name** | **Population Genomics** |
| **Project title**<br>*(max 150 characters)* | **Population Genomics GeCIP** |

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

We propose a Population Genomics GeCIP domain with four subdomains:

1) **Genetic population resources**, phasing the 100k Genomes sequences and using them for imputation, analysis of demographic population structure, and reference-free variation analysis.
2) **Linkage disequilibrium maps**, using the 100k Genomes data to develop high resolution next generation linkage disequilibrium maps.
3) **Germline mutation**, using trio data in the 100k Genomes rare disease sequences to explore factors influencing timing, rate and spectrum of germ line mutation.
4) **Population diversity**, addressing the impact of genetic, geographical and environmental variation on incidence, progression and response to treatment, and associated ethical considerations.

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

In addition to their direct value for the clinical care of the participants, the genome sequences from the 100,000 Genomes Project will provide an unprecedented genetic resource for the British population. The Population Genomics Clinical Interpretation Partnership draws together leading UK researchers and foreign collaborators to use the data from the project in order to understand better the processes that have given rise to genetic variation in England, and to generate derived information that will facilitate other genetic analyses, both within the 100k Genomes Project and elsewhere. The first three aims will use the whole genome sequences with very limited (non-disease) additional data to inform us about the history of mutation, recombination and genetic drift that has given rise to the genomes in the current population. The results will inform us about the ancestry of samples and mutations, help fill in missing data for disease-oriented genetic studies, and improve our understanding of the recurrence risk for genetic disease caused by new mutations. The fourth aim combines geographic, epidemiological and genetic data to investigate the relative weight of factors contributing to disease.

**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)*

1) **Genetic population resources**

We will phase the GEL whole genome sequences using enhanced state-of-the-art phasing software that takes advantage of very rare shared variants and scales to 100k genomes. Genetic fine structure will be analysed using further developments of recent haplotype-based methods, using phased full sequence data to provide additional population history information. Genetic structure will be combined with geographic information to support the analysis of the timing and spread of disease-causing mutations. We will call HLA haplotypes using graph reference structures, and also investigate blood and germline virus load. The results will be available to other GeCIPs and we will also use them to build servers that support high resolution phasing, imputation and ancestry analysis of external data sets in a secure setting.

**2) Next generation linkage disequilibrium maps**

We will use the 100k Genomes whole genome sequences to build high resolution linkage disequilibrium (LD) maps for up to 20 different ethnicities. These maps can be used in the context of association fine mapping and homozygosity mapping, in addition to the exploration of linkage disequilibrium structure in the human genome. LD maps also can be used in the study of homozygous regions in the genome, and we will use them in the analysis of recurrent uniparental disomy in cancer samples.

**3) Germline mutation**

We will make use of the whole genome sequences from rare disease parent-child trios to study factors affecting germline *de novo* mutation rates for a variety of different types of mutation, including single nucleotide variants, indels and structural variants. We expect to improve understanding of the sequence context requirements for different mutation processes, the tendency for mutation events to cluster in the genome, and dependency on parental age and ethnicity. We will also explore mosaicism, where a mix of mutant and non-mutant sequencing reads are seen on the same haplotype background within an individual, using data from this to fit models of mutation processes at different stages during the life history of the germline, and hence refining estimates of recurrence risk of apparent *de-novo* mutations.

**4) Population diversity**

We will combine ethnicity and genetic structure data from subdomain 1 with environmental exposure data obtained from spatial location, socio-economic data and electronic health records to investigate the factors affecting the incidence of both rare disease and cancer. We will transform broad data into suitable covariates for inclusion into classic statistical genetics tools, and in parallel use machine learning methods to model the integration of geographic and other exposure-related factors into environmental correlates for use in predictive modelling.

| | |
|---|---|
| **Expected start date** | April 2016 |
| **Expected end date** | March 2019 (provisional) |

| Lead Applicant(s) | |
|---|---|
| **Name** | Richard Durbin |
| **Post** | Senior Group Leader |
| **Department** | |
| **Institution** | Wellcome Trust Sanger Institute |
| **Current commercial links** | Founder and non-executive director Congenica Ltd., consultant Dovetail Genomics, shareholder due to former consultancy in Illumina Inc. |

| Administrative Support | |
|---|---|
| **Name** | Jenny Mansfield |
| **Email** | jm37@sanger.ac.uk |
| **Telephone** | 01223 496848 |

| Subdomain leads | | |
|---|---|---|
| **Name** | **Subdomain** | **Institution** |
| Jonathan Marchini (joint) | Genetic population resources | Oxford University and Wellcome Trust Centre for Human Genetics |
| Richard Durbin (joint) | Genetic population resources | Wellcome Trust Sanger Institute |
| Sarah Ennis | Next generation linkage disequilibrium maps | Southampton University |
| Matthew Hurles (joint) | Germline mutation | Wellcome Trust Sanger Institute |
| Aylwyn Scally (joint) | Germline mutation | Cambridge University |
| Jean-Baptiste Cazier | Population diversity | Birmingham University |

## Detailed research plan – Subdomain 1: Genetic population resources

| Full proposal (total max 1500 words per subdomain) | |
|---|---|
| **Title**<br>*(max 150 characters)* | Building a platform for genetic inference from the Genomics England data |

**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).*

The GEL dataset will constitute the largest human genetic variation resource ever collected in the UK, and maybe the world. Over the last 10 years resources such as these (HapMap[1], 1000 Genomes[2], UK10K[3]) have been widely used by the whole human genetics community for the purposes of genotype imputation into genome-wide association studies (GWAS), characterization of population structure and population genetics. We will bring together experts in genome sequence informatics and statistical genetics to generate a set of derived data sets and analysis tools from the 100,000 Genomics England (GEL) genome sequences. The outputs will have high value for studies of human genetics and human disease using both the GEL subjects themselves and third party data. The top level aims are:

1. Phase the genome sequences to generate the world's largest haplotype reference panel, empowering very low frequency imputation for future genome wide association studies, via imputation server resources.

2. Characterize the fine-scale genetic structure of the English population at an unprecedented level, providing knowledge for population structure matching and adjustment for disease studies. To succeed, we will pool expertize across the aims to build new methods that can handle the unprecedented scale of the dataset, and for the first time combine phasing and population structure analysis in one single step.

3. Develop and apply reference-free assembly/ alignment methods to the primary read data to characterise more divergent variation and (in the longer term) build a graph-based deep variation reference sequence resource for the English population.

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

**1: Phasing and genotype imputation**

When genomes are sequenced data from the two separate copies of each chromosome, one from the the mother and one from the father, are mixed together. The process of separating the resulting *genotype* into two separate *haplotype* sequences is known as phasing. This is important because many genetic analyses rely explicitly or implicitly on phased haplotype sequences.

**AIM 1.1** We will phase the GEL sequenced samples to generate haplotypes, using a combination of current state-of-the-art[4] and novel methods. New methods are needed to handle the scale of data, and will fully exploit long stretches of sequence shared identical-by-decent (IBD) between samples. The GEL data will contain a very large number of rare, recent mutation variants due to the high coverage sequencing approach used. Sharing a rare variant between a few individuals will often imply a close genealogical relationship and extended shared haplotype. We will use new methods that take advantage of this rare variant structure to rapidly identify shared DNA segments.

**AIM 1.2** We expect that the resulting panel will facilitate imputation of rare variants down to frequencies 1/10,000 for the first time. As an exemplar application we will carry out imputation of the 500,000 UK Biobank samples (www.ukbiobank.ac.uk) and the 50,000 Interval study samples (www.intervalstudy.org.uk). We will also take advantage of novel methods for computationally efficient imputation and data compression.

**AIM 1.3** Many other research groups carrying out GWAS will wish to use the GEL haplotypes as an imputation reference panel, or as a resource for phasing additional samples[5]. The panel will not be publicly available, so we will work to produce a server-based imputation and phasing resource. We will collaborate with the GEL Data Centre to provide a portal that allows properly approved researchers to upload their GWAS data and have imputation performed remotely in a secure way, at cost to the user. We expect synergy in this aim with the Haplotype Reference Consortium (co-led by Marchini and Durbin), which is developing such a service (www.haplotype-reference-consortium.org).

**2: Ancestry imputation and population structure**

An understanding of genetic differences among geographically spread individuals; i.e. population structure, is of central importance in elucidating how mutations (especially rare variants) influence phenotypes[6]. We will use the GEL data to analyse population sub-structure within England and links between individual variants of interest (for example rare disease-causing variants). We will inform mapping studies and sample-matching strategies, particularly for rare variants[7], shed new light on English genetic history, develop a resource for data owners to query the ancestry of their own samples, and perform ancestry-aware imputation of genetic variation in such samples.

**AIM 2.1** Generating a fine-scale map of genetic differences across England. This will address the extent of correlation between genetic and geographic information among individuals in England, and place the phased GEL haplotypes within a context of English population structure. This will build on our previous work in this area[8,9]

**AIM 2.2** Determining the geographic localisation patterns of rare (and common) variants, to help understand how rare pathogenic mutations have arisen and spread within England and to assist imputation of rare variants in future data collections.

**AIM 2.3** Elucidating the genetic history of England, a topic of wide public interest. We will infer past population sizes and fluctuations, and the number and timings of migration events both from outside sources and, for the first time, among genetically distinct groups within England.

**AIM 2.4** Development of an approach to use GEL data to provide a rich annotation of third party samples in terms of ancestry and rare variants carried.

**3. Apply reference-free assembly/alignment methods**

Almost all current human genome sequence analysis, including the variant calling proposed for the GEL data, relies on mapping sequencing reads to a linear reference sequence. The current alternative, de novo assembly, ignores previous knowledge, in particular of the longer-range structure of the genome. Recent developments in the McVean and Durbin groups and elsewhere use a more complex graph representation of genetic variation in a population to capture previously observed assembly structure, allowing more accurate, less-biased mapping and variant calling. We are currently working in both the Global Alliance for Genomics and Health and the Genome Reference Consortium to introduce such methods into standard data analysis.

**AIM 3.1** We will HLA type the sequenced samples by mapping to a reference graph structure for the MHC region representing known HLA types. The HLA type of patients is valuable for clinical research and in some cases for clinical action. We will also aim to identify new rare variation in the MHC present in English samples, which may be missed by standard reference-based methods.

**AIM 3.2** Using similar methods we will search for known integrated virus sequences, providing information about prevalence in English communities. We will also aim to identify possible novel or variant integrated viruses by mapping unmapped reads to models of virus families, and assembling around seed matches.

**AIM 3.3** More ambitiously, we will explore the feasibility to apply graph variation reference methods genome-wide, placing GEL at the forefront of international adoption of a richer mapping and variant detection paradigm.

**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 envisage that the results of our haplotype and ancestry inference will be of use to all other GeCIPs, when carrying out phenotype-based analysis. Furthermore the detailed HLA type may be relevant both to clinical care and other research within GeCIPs. We will make all of these results available within GEL both for GeCIPs and GMCs. We will collaborate with the UK BioBank and Interval projects through shared participation by Marchini, Myers, McVean and Durbin.

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

We will provide training for the postdoctoral researchers who will work on our proposal, who will gain expertise in developing and applying world-leading statistical methodology to large-scale sequencing data. We will explain the use of the new servers in documentation, workshops and courses.

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

**Prof Jonathan Marchini (co-leader)** – Professor of Statistics, Department of Statistics, University of Oxford. Expertise in computational statistical genetics, phasing and imputation.

**Prof Richard Durbin (co-leader)** – Senior Group Leader (currently acting Head of Computational Genomics) at the Wellcome Trust Sanger Institute. Expertise in genome sequence analysis and population genetics.

**Prof Simon Myers** – Associate Professor of Bioinformatics, Department of Statistics, University of Oxford. Expertise in population genetics, especially population structure and demographic inference.

**Dr Chris Tyler-Smith** – Head of Human Evolution group, Wellcome Trust Sanger Institute.

Expertise in human evolution and population genetics, especially Y chromosome analysis.

**Dr Garrett Hellenthal** – Research Fellow, UCL Genetics Institute. Expertise in population structure and demography.

**Dr Daniel Lawson** – Wellcome Trust Sir Henry Dale Research Fellow, University of Bristol, School of Social and Community Medicine. Expertise in computational statistics and population structure.

**Prof Gil McVean** - Professor of Statistical Genetics, Acting Director of the Oxford Big Data Institute, Wellcome Trust Centre for Human Genetics, University of Oxford. Expertise in genome analysis and population genetics.

**Prof Goncalo Abecasis (foreign collaborator)** – Professor of Biostatistics at the University of Michigan, USA. Expertise in large scale sequencing, genetic association studies and computational methods, and a partner with Marchini and Durbin in the Haplotype Reference Consortium.

**Plans for collaboration** – All these individuals have worked together before in large scale collaborations. Much of this work has been funded by a WT Collaborative Award (PIs Marchini, Myers, Hellenthal), which has substantial funds (~£53,000) to facilitate collaboration via travel between the 5 institutions involved in this research. We plan regular 2 monthly meetings to discuss the research and intervening conference calls.

**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.*

Clinical interpretation of genomes may be aided by having a phased genome sequence for a patient, for example when considering compound heterozygote effects. Primarily our research will feed into analysis of GEL samples by other GeCIPs, but our planned phasing servers also have the potential to provide estimated phased genomes for other UK patients.

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

We envisage that the results of our haplotype and ancestry inference will be of use to all other GeCIPs. Through those GeCIPs there will be of benefit to patients and healthcare institutions. The imputation, phasing and ancestry servers will be of great benefit to the whole human genetics community, especially those within the UK. Our ancestry inference may be reported to the GEL participants via the planned "patient portal".

**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 currently have no commercial plans for our research.

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

1. International HapMap Consortium *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
2. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
3. UK10K Consortium *et al.* The UK10K project identifies rare variants in health and disease. *Nature* (2015). doi:10.1038/nature14962
4. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10,** 5–6 (2013).
5. Marchini, J. & Howie, B. Comparing algorithms for genotype imputation. *Am. J. Hum. Genet.* **83,** 535–9– author reply 539–40 (2008).
6. International Multiple Sclerosis Genetics Consortium *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476,** 214–219 (2011).
7. Mathieson, I. & McVean, G. Demography and the age of rare variants. *PLoS Genet.* **10,** e1004528 (2014).
8. Leslie, S. *et al.* The fine-scale genetic structure of the British population. *Nature* **519,** 309–314 (2015).
9. Hellenthal, G. *et al.* A genetic atlas of human admixture history. *Science* **343,** 747–751 (2014).

## Detailed research plan – Subdomain 2: Linkage disequilibrium maps

### Full proposal (total max 1500 words per subdomain)

| | |
|---|---|
| **Title**<br>*(max 150 characters)* | **Applications of linkage disequilibrium mapping to Genomics England data** |

**Importance.**

Genetic maps are an essential tool for mapping disease genes. These are maps of our genome that capture information on recombination and other factors that must be considered when trying to accurately identify the genomic location of genetic changes causing disease. The first such maps were linkage maps and these were the tools successfully exploited in all linkage studies of families to uncover genes underlying single gene disorders. Following the Human Genome Project, a new age of linkage disequilibrium (LD) maps uses SNP data to massively improve the resolution of genetic maps. LD maps are centrally important for the design and interpretation of genetic associations with disease. In the era of high throughput genomics, it is now possible to create Next Generation LD Maps that are optimally resolved and serve as essential tools for the mapping and interpretation of genomic variation impacting health.

The University of Southampton Genetic Epidemiology & Genomic Informatics Group has a long standing history and expertise in genetic map development and application (Maniatis et al, 2002). The group is already working on one of the largest human whole genome sequence datasets (n = 500) of healthy elderly people (http://www.scripps.org/research__areas-of-research__genome-and-genomic-medicine-research__wellderly-study) to generate these maps. The wealth of data to be generated by the 100,000 Genomes Project will advance LD maps development by virtue of the large sample sizes of multi-ethnic groups across different disease areas. The maps created from the whole genome data of patients contributing to 100,000 Genomes Project can be used for high resolution association mapping, refining previously identified gene regions, detecting novel genomic rearrangements in cancer and understanding relationships between LD and disease.

**Research plans.**

**Development of high resolution LD maps from 100,000 Genomes data.** We have already established that the LD structure of chromosomes derived from array-based genotypes is poorly resolved in some genomic regions through ascertainment and other biases inherent in SNP selection for panels (Pengelly et al, 2015). Deriving genome-wide LD maps from subsets of 100,000 Genomes data samples across a range of ethnic groups will enable the construction of the highest resolution LD maps to date which will underpin disease genome mapping projects outlined below. We envisage creating at least 20 alternative ethnic specific LD maps. A necessary prerequisite for this and other work within the Population Genomics Domain as a whole will be the accurate ethnic characterisation of participant samples.

Immediate application of these maps include:

- Study/confirm impact of marker density
- Refine physical location of ‘hotspots” and relationship with known motifs (e.g. PRDM9); further analyse those negative for known motifs
- Through application of coalescent methods to recover ancestral recombination patterns (e.g. LDHat, LDHelmet) using the same ethnic specific data, identify divergent regions between these two maps thereby isolating regions of LD breakdown not caused by recombination for further study. Correlate with recombination motifs; local sequence characteristics (transposable elements; simple repeats, conservation, GC content).
- Examine relationship between LD and genes/regions harbouring disease loci; functional clustering (strong LD in genes essential for biological function) weak LD in genes promoting diversity (sensory, immune). Correlate mutational profiles with LD.

**100,000 Genomes derived LD maps for homozygosity mapping.** Gibson et al (2006) demonstrated that extended regions of homozygosity are found even in 'outbred' individuals. Through comparison with LDU map structure, these regions were shown to reflect autozygosity arising where long ancestral haplotypes have survived in genomic regions with a low recombination rate. Such regions have been identified by many authors and successfully screened for recessive disease genes (for example, Lencz et al, 2007). Through analysis of LD structure in 100,000 Genomes samples we will develop and apply methodology for identification of recessive disease variants to enhance the utility of clinical reporting and the understanding of disease predisposition.

- Identify refined loci where evidence for LD 'hotspots' is discordant between ethnic populations; assess evidence for extended homozygosity, positive selection. Distinguish 'population' runs of homozygosity forming plateaus in LDU map and extended runs in individual patients.

**100,000 Genomes derived LD maps for fine mapping disease variants.** Extensive linkage disequilibrium (LD) in the genome presents difficulties wherever a number of candidate disease causal variants are located together in close proximity. Array-based LDU maps have been used successfully to delimit narrow regions containing significant disease association signals. In this model one LD unit is the chromosome region over which LD declines to background levels and this scale is therefore useful to objectively delimit a region in which an association signal lies. Sabatti et al, (2009) used a two LDU window to delimit newly identified loci involved in metabolic traits. Using an LDU map Elding et al (2011) were able to identify independent neighbouring genes involved in Crohn disease for which independent effects were otherwise confounded. Recent applications of African-American LD maps in association mapping for Type-2-Diabetes have yielded exciting results (Direk et al. 2014). We have also adapted the methodology for mapping regulatory hot spots in the genome based on tissue expression profiling. We recently identified new location estimates for T2D on LDU maps that co-located the same refined estimates with independent data on adipose expression (Lau et al, submitted).

We propose to exploit 100,000 Genomes-derived LD maps for fine-mapping in known and novel disease association regions for a range of phenotypes. The program CHROMSCAN (Collins and Lau, 2008) has been developed for fine mapping disease variants using an underlying LDU map and has been shown to have high resolution which we expect to increase further using 100,000 Genomes-derived LD maps.

Immediate application of these maps include:

- An extension of this work will be to use the deep phenotyping of GEL participants to create common disease groups with maximally dense marker data and apply these to refine association mapping signals.

**Identification of genetic targets of acquired uniparental disomy aUPD.** Recurrent aUPD in the same chromosomal region in a particular group of cancer patients suggests that the region harbours a key gene that, when mutated, is driving the malignant process. Indeed, many genes involved in a wide range of cancers have been identified by detecting minimal regions of recurrent aUPD. We have previously shown that up to a third of patients with myeloid malignancies are characterised by aUPD and that recurrent abnormalities of chromosome 7q, 11q and 14q target EZH2, CBL and MEG3-DLK1, respectively (Ernst et al 2010; Grand et al 2009; Chase et al 2015). Remarkably, regions of aUPD have also been identified in apparently healthy individuals, albeit at a much lower frequency. Only 0.5% of people under the age of 50 are affected but this rises to 2-3% of elderly individuals. Importantly, the chromosomal regions affected are identical to those seen in patients with haematological malignancies and involve the same mutant genes. The finding of aUPD in an otherwise haematologically normal individual is associated with a tenfold increased risk of subsequently developing haematological neoplasia (Laurie et al 2012; Jaiswal et al 2014). We aim to interrogate data from 100,000 Genomes to:

- identify minimal recurrent regions of aUPD
- detect novel gene targets
- refine frequency estimates of recurrent aUPD
- investigate the relationship between LD maps and recurrent regions of aUPD

**Collaborations including with other GeCIPs.**

This subdomain will collaborate with the Genetic Resources subdomain to identify coarse level genetic identity at the ethnic group level. We envisage further collaboration with cross cutting the deep phenotyping domain involved in association genetics studies on deep phenotypes, and collaboration with cancer domains with respect to recurrent UPD identification.

**Training.**

We envisage PhD student projects on distinct aspects of the proposed work. Informatically capable students undertaking the University of Southampton MSc in Genomic Medicine will have the opportunity to undertake suitably sized projects fitting within the research goals detailed above. Many members of the sub-domain contribute to various aspects of teaching and supervision on this course.

**People and track record.**

The University of Southampton Genetic Epidemiology & Genomic Informatics Group has a long standing history and expertise in genetic map development and application (see refs). The group is already working on one of the largest human whole genome sequence datasets (n = 500) of healthy elderly people (http://www.scripps.org/research__areas-of-research__genome-and-genomic-medicine-research__wellderly-study) to generate these maps. Sub-domain members at UCL have applied LD mapping data for refinement of disease gene loci. Members of sister sub-domains have extensive experience of coalescence and linkage disequilibrium analysis.

**Beneficiaries.**

The main beneficiaries will be those working on human genetics association and homozygosity mapping studies who are using LD maps for their analysis, who will benefit from the improved resolution of the new maps, plus those working on UPD in cancer.

**Commercial exploitation.**

We currently have no commercial plans for our research.

**References.**

Collins, A. The genomic and functional characteristics of disease genes. *Briefings in Bioinformatics*, 2014: bbt091.

Collins, A, and Lau W. CHROMSCAN: genome-wide association using a linkage disequilibrium map. *Journal of Human Genetics* 53, no. 2, 2008: 121-126.

Direk K, Lau W, Small KS, Maniatis N, Andrew T. ABCC5 transporter is a novel type 2 diabetes susceptibility gene in European and African American populations. *Ann Hum Genet.* 2014 Sep;78(5):333-44.

Elding H, Lau W, Swallow DM, and Maniatis N. Refinement in localization and identification of gene regions associated with Crohn disease. *Am J Hum Genet* 92, no. 1, 2013: 107-113.

Ennis S, Collins A, Tapper W, Murray A, MacPherson JN, Morton NE. Allelic association discriminates draft orders. *Ann Hum Genet.* 2001;65(Pt 5):503-4.

Gibson J, Tapper W, Ennis S, and Collins A. Exome-based linkage disequilibrium maps of individual genes: functional clustering and relationship to disease. *Human Genetics* 132, no. 2

(2013): 233-243.

Gibson J et al. Extended tracts of homozygosity in outbred human populations. *Hum Mol Genet* 2006,15 (5): 789-795.

Jeffares, D. et al. The Genomic and Phenotypic Diversity of Schizosaccharomyces pombe. *Nat Genet*, 2015, in press.

Jeffreys A. and Neumann R. The rise and fall of a human recombination hot spot. *Nat Genet* 2009, 41 (5): 625-629.

Kadalayil L., Rafiq S, Rose-Zerilli MJJ, Pengelly RJ, Parker H, Oscier D, Strefford JC et al. "Exome sequence read depth methods for identifying copy number changes." *Briefings in Bioinformatics* 2014: bbu027.

Lau W et al. Exploiting large scale computing to construct high resolution linkage disequilibrium maps of the human genome. *Bioinformatics* 2007, 23 (4): 517-519.

Lencz T et al. Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc Natl Acad Sci* USA 2007,104 (50): 19942-19947.

Maniatis, N., A. Collins, C-F. Xu, L. C. McCarthy, D. R. Hewett, W. Tapper, S. Ennis, X. Ke, and N. E. Morton. "The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis." *Proc Natl Acad Sci USA* 99, no. 4, 2002: 2228-2233.

Morton NE et al. The optimal measure of allelic association. *Proc Natl Acad Sci USA* 2001, 98 (9): 5217-5221.

Pengelly et al. Whole genome sequences are required to fully resolve the linkage disequilibrium structure of human populations. *BMC Genomics.* 2015 Sep 3;16:666.

Sabatti C., Service SK, Hartikainen A-L, Pouta A, Ripatti S, Brodsky J, Jones CG et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet* 41, no. 1, 2008: 35-46.

Service S et al. Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nat Genet* 2006, 38 (5): 556-560.

## Detailed research plan – Subdomain 3: Germline mutation

### Full proposal (total max 1500 words per subdomain)

| | |
|---|---|
| **Title**<br>*(max 150 characters)* | **Germline *de novo* mutations and mutation processes** |

**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).*

New (de novo) mutations observed in children are a major cause of rare genetic diseases. The focus of this subdomain is on characterizing the rate, spectra and timing of germline de novo mutations (DNMs) observed in the rare disease families sequenced by GEL, and using this information to: (i) better understand the underlying mutation processes (ii) develop better methods for identifying DNMs, and (iii) use improved understanding of parental mosaicism and germline cellular genealogies to derive better recurrent risk estimates for genetic disease.

It has been shown that at the population level, the main factor increasing the number of DNMs seen in a child is paternal age (Kong et al 2012). However, this is only true for some classes of DNMs (e.g. base substitutions) and not others (e.g. deletions caused by non-allelic homologous recombination) (MacArthur et al 2014). It is likely that other factors, including both genetic variation and environmental exposures, also influence locus-specific and genome-wide mutation rates (as has been observed for somatic mutation processes). The generation by GEL of deep whole genome sequence data on thousands of parent-offspring trios presents a major opportunity to characterize these factors with unprecedented power.

**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.*

This subdomain will generate high confidence de novo mutation (DNM) datasets and use them to investigate the cellular processes and genomic factors influencing the number and distribution of new mutations observed in children. A particular focus will be on improving our understanding of recurrence risk and developing new tools for the detection of DNMs. The research will comprise three main strands:

Strand 1: High-confidence DNM datasets and characterisation of DNM rate, spectra and distribution

- Generation of high confidence datasets comprising DNMs of different types to drive a range of downstream analyses, including all classes of variation, i.e. SNVs, indels and structural variants. This will involve careful handling of false positive and false negative erroneous variants, particularly in cases where parents or offspring are mosaic, which are not well handled by standard variant calling methods. Experimental validation of these variant calls in the DNA of parents and offspring will be required to quantify the accuracy of the datasets. This will require access to the DNA held in the biorepository, for a subset of individuals.

- Statistical description and characterisation of DNMs
  - DNM rate and spectra (e.g. the proportion of base substitutions that arise at

different triplet or larger sequence contexts, or the proportion of large deletions that are caused by non-allelic homologous recombination)
    - Clustering of DNMs within the genome and the degree to which localized hyper-mutation is observed and can be attributed to particular molecular mechanisms.
    - Variation of these and other characteristics with parent-of-origin, parental age and ancestry information from other analyses within the Population Genomics domain.

Strand 2: Validation and application of new methods and tools

- Improved methods to estimate the recurrence risk for future siblings sharing a pathogenic mutation, incorporating information about the timing of DNMs within the germline based on evidence for mosaicism in either parent (in particular the proportion of DNMs arising prior to germ cell specification) and insights into the cellular geneaology of the germline developed in Strand 3.
- Improved tools for detecting DNMs, again incorporating additional understanding developed in Strand 3 and information about the timing and distribution of DNMs. We anticipate that these new methods will be of particular value to increase power to detect certain classes of mutations that are not well captured by current algorithms, for example those that are mosaic in either parent or child.

Strand 3: Germline cellular processes and genomic factors affecting germline mutation rates

- Modelling mutation events, including complex and multiple events, and DNM sequence context to explain the genomic distribution of DNMs.
- Identifying genomic loci affecting the germline mutation rate, using genome-wide association with the 100,000 Genomes Project samples in combination with analyses of existing population genetic data (which carry information via the distribution of derived alleles on different haplotypes).
- Modelling cellular genealogies and mutation processes within human individuals, with particular focus on the structure of spermatogenesis (the cell state transitions, cycle times and death rates involved) and the number of divisions at different stages of the germline. Important inputs will come from details of the paternal age effect and the joint distribution of DNMs in parents and offspring, and analyses will involve mathematical and computational modelling using finite-state/birth-death process approaches.

There will be a considerable overlap in the timelines for each of these Strands, given a degree of interdependence between them:

2016 onwards: Strand 1, generation of primary DNM dataset and statistical analyses
2016 onwards: Strand 2, development and implementation of novel methods
2017 onwards: Strand 3, improved understanding of germline processes and genomic factors

**Collaborations including with other GeCIPs.** *Outline your major planned academic, healthcare, patient and industrial collaborations. This should include collaborations and data sharing with other GeCIPs. Please attach letters of support.*

The planned work is relatively self-contained and is not inherently dependent on external collaborations. However, we may decide to collaborate with other groups with similar datasets to maximise sample size. Where we apply new tools or datasets which may be of use for identifying pathogenic mutations in rare disease patients that have been missed by the standard variant calling pipeline, we will work with the GEL Bioinformatics group and/or the rare disease GeCIPs to enable these additional diagnoses to be made.

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

This research domain is less relevant for clinical training, but we will be training PhD student and post-doctoral trainees as part of this sub-domain, within our respective groups.

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

The co-leaders of this sub-domain (Scally and Hurles) have long-standing research expertise in investigating germline mutation processes of various different classes of mutation, especially with regards to analysing mutation processes using massively parallel short-read sequencing, as detailed in the CVs.

The additional members of this sub-domain have long-standing complementary expertise in the analysis of mutation processes in highly duplicated regions of the genome (Hollox and Armour) and in evolutionary analyses of germline mutation processes (Eyre-Walker), as detailed in the CVs.

The investigators will coordinate their work through regular conference calls, and less regular face-to-face meetings, primarily in the context of the wider GECIP.

**Clinical interpretation.** *(Where relevant to your GeCIP) Describe your plans to ensure patient benefit through clinical interpretation relevant to your domain. This should specifically address variant interpretation and feedback and your interaction with the cross-cutting Validation and Feedback domain.*

If we apply tools and datasets that have the potential to identify diagnostic mutations missed by the standard variant calling pipeline then we will work with the Validation and Feedback domain to integrate these additional mutations, for patient benefit.

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

The planned research will validate tools and datasets that will likely enable the identification of diagnostic mutations missed by the standard variant calling pipeline. Enabling these diagnostic mutations to be fed back to the patients will benefit the patients, their clinicians and the NHS.

The planned research will also enable improved counselling of parents of recurrence risks in subsequent siblings of carrying the same pathogenic mutation.

**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 do not anticipate that the planned research will generate commercially exploitable results. Therefore there are no commercial partners in place currently. If this changes, we will work with Genomics England to manage any IP that is generated in this sub-domain.

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

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, et al. (2012) 'Rate of de novo mutations and the importance of fathers age to disease risk'. *Nature* 488: 471-475.

MacArthur JA, Spector TD, Lindsay SJ, Mangino M, Gill R, et al. (2014) 'The rate of nonallelic homologous recombination in males is highly variable, correlated between monozygotic twins and independent of age'. *PLoS Genetics* 10: e1004195.

## Detailed research plan – Subdomain 4: Population Diversity

### Full proposal (total max 1500 words per subdomain)

| | |
|---|---|
| **Title**<br>*(max 150 characters)* | **Population Diversity** |

**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).*

The 100K initiative that collects both sequence and clinical data from across the country provides a unique opportunity to address several classic genetic epidemiological issues. Such phenotypically complete, and therefore complex, data will allow us to disentangle genetic background from other geographic, socioeconomic or cultural factors to properly address questions related to Population Diversity. Whether we are looking at Cancer or Rare Diseases there is some impact from the genetic background; it can lie at the Caucasian/non Caucasian boundary (Amundadottir, 2006; SenGupta, 2013), or within the diversity of the white British population as demonstrated by the People of the British Isles study (Winney, 2012; Leslie, S., Winney & Hellenthal, 2015). While the West Midlands Genomic Medicine Centre is ideally placed for access to a region of a strong ethnic diversity, we are also interested in the subtler differences that exist within the "indigenous" British population (Davies & Cazier, 2012). In the case of cancer, genetic background can lead to a diversity of incidence, as well as progression and response to treatment. However the same can also be said of "environmental" factors and having access to such longitudinal clinical data from across the country should prove useful in disentangling these interrelated effects. We would also aim to associate somatic changes to such "environmental" factors. Interestingly similar effects could also be true for rare diseases where a specific "isolated" ethnic group might lead to increased consanguinity and therefore risk. Rare dominantly acting variants are often ‘founders’ which can also lead to marked local differences in frequencies. For recessive diseases associated socioeconomic and cultural factors can play a very important role in the incidence, report, prevention, monitoring and progression of the disease, independently or not, of the ethnicity itself. Spatial data (locational metadata associated with disease and genetics data) requires unique data handling techniques to preserve the anonymity of individuals and groups (such as ethnic communities), as location can be indicative of identity, in breach of ethics.

We therefore intend to work on 3 main branches, across all the GeL proposed diseases:

- Genetics: Study of the genetic impact on incidence, progression, response to treatment.
- Environmental impact on incidence, report, progression and response to treatment. In "Environment" we would include diet, smoking, culture and broader socioeconomic and geographic factors such as deprivation, education level, spatial attributes including access to healthcare facilities, rural/urban location, the north-south divide, the toxicity of the area, etc.
- Ethical impact: how ethnicity impacts the geographical reporting of disease; ethics of spatial data handling and locational privacy; spatial clustering and identity/anonymity.

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

Our proposal relies on the identification of structure through the vast number of disorder, genomic, phenotypic and clinical data. We therefore propose to perform the analysis at two parallel tracks of disorder and global levels:

At the disease level we shall look for effect of the various parameters collected such as ethnic background, deprivation index or access to care, in relation to the Genomic association into clinical covariates as well as the onset, progression and response to treatment. Our experts in the relevant Clinical, Geography or Toxicity fields will ensure the broad data collected are appropriately transformed into suitable covariates for inclusion as covariates in classic statistical genetics tools. In a further step we shall coordinate with the relevant disease facing GeCIP domain to gather the maximum knowledge of their specificity, especially in relation to ethnic background or environmental exposure, and better understand potential ethical impact.

In parallel, we will mine the data from a geographical perspective and look to further improve our understanding of spatial diversity, such as genetic, cultural, disease distribution. Beyond the use of classic epidemiological and geographical tools we plan to use Machine Learning technique to better characterise the underlying structure between these various "environmental" parameters, and their impact on the same Clinical phenotype including onset, progression and response to treatment. Both approaches will almost use the same data, but from very different angles, providing a better understanding on the impact of "environment" and genomic background on disease.

**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.*

This proposal links directly with the clinical impact of existing data with every single disorder and therefore is heavily dependent on collaboration with the Clinical facing GeCIP domains. While this subdomain lead is a member of many such domains, he will make sure to reach out individually to each of them to ensure collaboration.

Furthermore several links have been made with crosscutting domains. A common goal of secure data collection and access is driving the connection with the Electronic records, Ethics and Social Science and Health Economics domains. On the other hand the subdomain lead is a member of the Machine Learning, Quantitative Methods and Functional Genomics domain and aims to share method development.

Finally there is a special link with the Education and Training GeCIP domain as this subdomain lead is closely involved with the delivery of his local MSc in Genomics Medicine program.

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

We aim to use several avenues to incorporate the knowledge from Academia and the Clinic, as well as uses of Bioinformatics into our education and training programme.

The University of Birmingham is hosting one of the MSc in Genomic Medicine. We propose this link to the HEE to be the main vector for translation of the information. This will provide access to the HEE network to disseminate the information, novel findings and approaches. Moreover, we would like to add an online-only module to share across the selected MSc centres addressing the topic of *Population Diversity*. Of note, the subdomain lead is in charge the Bioinformatics and Advanced Bioinformatics modules of the bid, which will ensure a smooth translation of the GeCIP information. Furthermore several members are personally involved in other GeCIP domains both analytical and disease facing. Such cross involvement will help to ensure use and development of the best methods as well as translation of the findings of this GeCIP to applications to the relevant diseases.

Furthermore the direct involvement of the clinical facing GeCIP domains leads to this group of all concerned conditions will ensure the proper information is circulated.

Finally the subdomain lead, Prof Cazier, is starting a novel MSc in Bioinformatics in 2016 which will naturally give a large part to Population Diversity research and the finding of this subdomain.

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

Our proposed domain requires an interdisciplinary team with a very broad diversity of skills including Genetic, Clinic, Bioinformatics, Statistics as well as Human and Economic Geography, Geographic Information Science and Ethics. As such, in drawing together the proposed group of collaborators, we aim to these diverse fields with membership including:

- Clinical application with Prof Tomlinson, Prof Raza, Dr Raghavan with support from Prof Zalloua, as well as the disease related GeCIP domain leads.
- Bioinformatics and Statistics: Prof Bodmer, Prof Holmes, Prof Cazier, Prof Gkoutos and Dr Arribas-Bel as well as our international collaborator Dr Brown.
- Population Genetics: Prof Bodmer and Prof Cazier as well as our international collaborator Prof Zalloua and Dr Carvajal-Carmona.
- Socio economic and geographic analysis using spatial and econometric modelling with Prof. Singleton, Dr Tranos, Dr Cheshire, Dr Arribas-Bel, Dr Mark Green and international collaborations from Prof Schuurman and Prof Mobley.
- Ethics with Dr Cinnamon as well as our collaborators, Prof Schuurman and Dr Leszczynski.

**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.*

The main goal of this domain is to maximise the clinical impact of the 100K genome project by taking into account all the relevant factors that can affect a patient. Working in close relationship the clinical-facing GeCIP domains will ensure the enrolled patients, and their host GMC, have access to the complete knowledge to inform their decision.

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

This work will have three level of beneficiary: the patient, the NHS and the researchers. Integrating all available, and identifying all relevant, parameters aims to provide even more personalised medicine. It will therefore benefit first and foremost the patient who can obtain a diagnosis, treatment and monitoring particularly suited for his/her need. This increased efficiency will in turn provide substantial savings for the NHS by allowing earlier diagnosis in at-risk population, higher efficacy in the treatment and adequate follow up. Finally, this increased understanding will support the development of the integration between Health and Environment.

**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?*

While several industrial partners have expressed interest, no specific agreements have been reached.

West Midlands Academic Health Science Network has committed their support to this work. In addition to advice and expertise from their NHS and academic partners across the region, they will provide access to an extensive number of industry partners. These range from global corporations to small and medium regional enterprises across a diverse range of specialist areas, including biomedical devices, pharmaceuticals, digital entrepreneurs and IT.

WMAHSN and their Industry Reference Group, offer the opportunity for collaborations in areas such as commercial ethics, environmental impact, and the translation of research into commercial settings.

Dr Chris Parker CBE is the WMAHSN Managing Director who has agreed to oversee this partnership, supported by Neil Mortimer, WMAHSN Head of Programmes.

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

- **A Common Variant Associated with Prostate Cancer in European and African Populations.** Amundadottir, *et al.* (2006) *Nature Genetics*, 38 (6), 652-8.
- **Gene-Ancestry Interactions in Genome-Wide Association Data.** Davies JL*, Cazier J-B*, *et al.* (2012) *PLoS One*, *7(12),*

- **Fine scale genetic structure of the British population.**
  Leslie, S., Winney, B., Hellenthal, G. *et al.* (2015) *Nature* **519**: 309–314
- **Analysis of colorectal cancers in British Bangladeshi identifies early onset, frequent mucinous histotype and a high prevalence of RBFOX1 deletion.**
  Sengupta N, *et al.* (2013) *Mol. Cancer*, *12:1*
- **Technical and implementation issues in using next generation sequencing of cancers in clinical practice.**
  Ulahannan D, *et al.* (2013) *British Journal of Cancer*, *109(4):827-35*
- **People of the British Isles: preliminary analysis of genotypes and surnames in a UK-control population.**
- Winney B, *et al.* (2012) *Eur J Hum Genet.*;20(2):203-10
- **Geographic disparities in late-stage cancer diagnosis: Multilevel factors and spatial interactions.**
  Mobley, L, *et al.* (2012) *Health & Place*, v 18: 978-990.
- **Common variants at the MHC locus and at chromosome 16q24.1 predispose to Barrett's esophagus.**
  The Esophageal Adenocarcinoma Genetics Consortium, *et al.* (2012) *Nature Genetics* 9;44(10):1131-1136
- **Genome-Wide Association Study in a Lebanese Cohort Confirms Phactr1 as a Major Determinant of Coronary Artery Stenosis.**
  Hager J, *et al.* (2012) *PloS One*, 7(6):e38663
- **Breast cancer stage at diagnosis: is travel time important?**
  Henry KA, *et al.* (2011) *Journal of Community Health*, 36(9): 933-942
- **Detecting an association between socioeconomic status and late stage breast cancer using spatial analysis and area-based measures.**
  Mackinnon JA, *et al.* (2007) *Cancer Epidemiology Biomarkers Prevention*, 16(4): 756-762.
- **Early- and late-onset breast cancer types among women in the United States and Japan.**
  Matsuno, RK, *et al.* (2007) *Can Epidem Biomarkers & Prev,* 16(7):1437-42.

## 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)*

As cross-cutting domain we require access to data on all participants.

In particular, we will make extensive use of whole genome VCF files, including measures of variant call confidence such as DP4. We require access to BAM files for subdomain 1 aim 3 for reference-free methods, for subdomain 3 to evaluate candidate de novo mutations (spot access) and perhaps for a structural-variant mutation scan, and potentially for subdomain 4 for whole genome re-analysis.

With respect to additional data, subdomain 1 aim 2 requires age plus geographic information: ideally the first component of the postcode for the participant's birthplace, and if possible that of their parents and grandparents, but even the current residence of the participant would be helpful. Subdomain 3 requires the parental age at conception for parent-offspring trios (derivable from the age of all members of the trio).

Subdomain 4 requests in addition widespread clinical, EHR and personal data (see below in Key Phenotype Data).

Subdomain 3 will require access to DNA samples for validation purposes. Validation will focus on an unbiased subset of called mutations rather than potentially diagnostic mutations, so will not be clinical in nature, and perhaps not appropriate to be conducted by a GMC.

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

This domain predominantly uses computational genetics software applied to whole genome data sets across large numbers of samples, up to all of the 100k genomes samples. Although in some cases we can use incremental analyses (e.g. trio by trio for some mutation analyses) in most cases the analyses will be run in batch, and so we will apply them at a series of timepoints through the lifetime of the project on agreed data set "freezes".

Scientifically, the analyses are described in fuller detail in the subdomain detailed research plans. Computationally, for most analyses we will create merged VCF (or equivalent) files of all or identified subsets of the samples by chromosome, then split them into convenient-sized chunks (e.g. 5Mbp for many analyses) for distribution across a compute farm, then merge the results back into chromosomal results files. All compute-intensive analyses will be piloted on subsets of the data, either small numbers of samples or small regions of the genome. Mostly the heavy computation will be carried out using C or C++ software that operates in reasonable memory compute nodes (e.g. 24GB or 32GB). Following the large scale compute operations there will be further statistical analyses prior to preparation of results for publication or other applications.

**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)*

Age and geographic data on the participants, and if possible their parents and grandparents, are necessary as described above in the Data Scope section.

Subdomain 4 requires more precise geographic information, i.e. more digits from the postcode, as well as clinical data for all disorders to allow them to inform the genetic association as well as investigate its relationship with personal data. It also requires personal data so as to link to the history of environmental exposure of the patients. Ideally this would include the history of the precise address, socio-economic condition, diet, etc. However we realise this is not practically convenient, and ethically difficult. We are therefore considering using the full postcode and uploaded related information such as Zone Code, MSOA, Index of Multiple Deprivation (overall measure and subdomains), ONS Output Area Classification or London Output Area Classification, to extract the relevant information within the Embassy.

**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)*

Subdomain 3 on new mutations will attempt to select candidates from existing calls then investigate further using the BAM files. For structural variant mutations this may perform very poorly, making a strong case to recall these from scratch from the BAMs, which would require a pass through the trio BAMs.

Subdomain 1 aim 3 would require effectively realignment/assembly of at least a subset of the reads. We would pilot this carefully in collaboration with GEL Bioinformatics to evaluate feasibility and value for resource use.

**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)*

This GeCIP is computation heavy, and we will bring in a variety of computational genetics tools including for phasing (e.g. SHAPEIT), imputation (e.g. IMPUTE, MINIMAC, PBWT), LD map creation, de novo mutation detection (DE NOVO GEAR) and a variety of statistical genetics and machine learning tools for subdomain 4. We are happy to work closely with GEL Bioinformatics, with whom we have already been in contact. In particular, for new tools, which will be developed outside the GeCIP and brought into the data centre, we may wish to bring them in as source code and compile them within the data centre alongside standard libraries to ensure binary compatibility and optimisation.

**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)*

We will want to bring in some existing genome variation datasets (at VCF level) including 1000 Genomes Project and POBI for comparison, plus annotation and other reference data sets to be discussed with GEL Bioinformatics.

Subdomain 4 wants to access Public Health and similar records to get full environmental information about each patient. Given the confidentiality of the records they are seeking to work closely with Genomics England to get a secure solution to link the patient clinical data beyond the set collected by the 100k Genome Project.

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

Phasing and fine structure analysis are computationally demanding (e.g. hundreds of cores for weeks) but can be well modelled and planned. The reference-free analysis will require careful piloting and planning, and may require large memory machine access (512GB or more if available).

Provision of servers as envisioned in subdomain 1 for imputation, phasing and ancestry analysis will require an additional establishment of compute infrastructure and external access, concerning which we have had initial discussions with GEL.

## 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)*

Subdomain 4 requests metabolomics and transcriptomics data when they are available for the integration with environmental factors, especially toxicology.

## Data access and security

| | |
|---|---|
| **GeCIP domain name** | Population Genomics |
| **Project title**<br>*(max 150 characters)* | Population Genomics |

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

*4 Clinical care*

*4 Clinical trials feasibility*

*4 Deeper phenotyping*

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

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

□ *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*

*4 Other health use - clinical audit*

*4 Public health purposes*

*4 Subject access request*

*X Tool evaluation and improvement*

**Uses marked with an *X* are relevant to the whole domain. Uses marked with *4* are relevant only to subdomain 4 (Population Diversity).**

***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.