GeCIP Detailed Research Plan Form

August 2015

Background

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

The aims of the partnerships are:

1. To optimise:

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

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

3. To provide a sustainable thriving training environment.

The initial wave of GeCIP domains was announced in June 2015 following a first round of applications in January 2015. On the 18th June 2015 we invited the inaugurated GeCIP domains to develop more detailed research plans working closely with Genomics England. These will be used to ensure that the plans are complimentary and add real value across the GeCIP portfolio and address the aims and objectives of the 100,000 Genomes Project. They will be shared with the MRC, Wellcome Trust, NIHR and Cancer Research UK as existing members of the GeCIP Board to give advance warning and manage funding requests to maximise the funds available to each domain. However, formal applications will then be needed to individual funders. They will allow Genomics England to plan shared core analyses and the required research and computing infrastructure to support the proposed research. They will also form the basis of assessment by the Project's Access Review Committee, to permit access to data. Some of you have requested a template for the research plan which we now provide herewith.

We are only expecting one research plan per domain and have designed this form to contain common features with funder application systems to minimise duplication of effort. Please do not hesitate to contact us if you need help or advice.

Domain leads are asked to complete all relevant sections of the GeCIP Detailed Research Plan Form, ensuring that you provide names of domain members involved in each aspect so we or funders can see who to approach if there are specific questions or feedback and that you provide details if your plan relies on a third party or commercial entity. You may also attach additional supporting documents including:

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

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

Application Summary	
GeCIP domain name	**Renal**
Project title *(max 150 characters)*	**Genetic causes of kidney disease**

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

1. **To characterise the spectrum of variation in known kidney disease-associated genes that cause disease and correlate this with severity and phenotype**
2. **To identify novel genes associated with (ie causing) kidney disease**
3. **To use these genetic insights to allow clinical feedback in order to improve diagnosis and prognostication and to inform transplantation and reproductive decisions in patients**
4. **To use novel genetic insights to improve understanding of the pathophysiology of kidney disease and stimulate novel approaches to treatment**
5. **To identify common genetic variants that contribute to disease risk or progression in genetic kidney diseases.**

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)*

Kidney disease is an important cause of death and illness in the UK (costing up to 3% of the NHS budget). It is known that in many cases the disease runs in families and is caused by a change in a gene. Many genes for kidney disease are already known but in some cases a gene has not been identified. We will use data from the 100,000 genomes project firstly to define the range of changes in established kidney disease genes causing disease, and secondly to find new genes that can cause kidney disease. We will use this knowledge to offer new and improved genetic tests to people with kidney disease – explaining why they developed disease and also allowing them and their relatives to make better-informed decisions, especially around having children and donating a kidney to a relative in need of a transplant.
In addition, we will use this new genetic knowledge to better understand the mechanisms by which changes in genes and the proteins they produce can cause kidney disease – an important step in designing new treatments for these disorders.

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. Rare variants in genes known to be associated with kidney disease will be identified by: filtering variants identified in these candidate genes against frequency in the general population; co-segregation analyses (including de-novo variant detection); comparison with databases of genetic variants found in people with the disease and prediction of effect of each variant on protein structure and function. Variants will be classified by likely pathogenicity and either fed back to participants if they can be reliably inferred to be disease-causing or subjected to further analyses, such as comparison of frequency in patients and controls. Investigation of the effects of gene changes on protein function will be performed in laboratories where particular genes/pathways are studied. Expected

outputs include enrichment of disease/variant databases, new genotype-phenotype correlations and functional insights regarding mechanisms linking gene dysfunction to disease.

2. New disease causing genes will be identified by aggregation and burden testing, which is identification of clustering of rare variants at a gene in more subjects with a phenotype than would be expected to occur by chance. Novel candidate variants identified will be prioritised using multidimensional analysis incorporating available pathway and transcriptome data. Expected outputs are new disease-gene associations and better mechanistic understanding of disease mechanisms. Statistical approaches that will be used include phenotype similarity regression based on HPO terms that allows analysis taking into account phenotypic heterogeneity.
3. Prospective audit will be undertaken to record the effect of delivering genetic diagnosis to participants. In particular we will record the kidney transplants that are facilitated by a) knowledge of risk of disease recurrence in an affected individual and b) exclusion of the risk allele in related prospective kidney donors. Expected output will be publications reporting changes in treatments offered to patients with kidney failure.
4. Novel genes and pathogenic variants identified above will be studied in vitro by laboratories/investigators with expertise in the phenotype, pathway or protein involved and the nature of the studies will depend on the gene in question. In parallel, mouse, zebrafish or other in vivo models will be generated where the evidence for causation and novelty are strong. This will provide resources to allow manipulations to be performed that might lead to the delivery of effective treatments in the future. Expected outputs will be scientific literature reporting new discoveries.
5. In parallel with the identification of known and novel monogenic drivers of disease, progression rate or age of onset of kidney failure will be used to perform association studies to identify non-Mendelian genetic risk alleles and other modifiers using groups stratified by primary disease gene/mutation. Some such variants have already been identified using conventional genetic studies and it is anticipated that this approach will provide novel insights into the biology of chronic kidney disease, which might be applicable even where the cause of disease is not a monogenic disorder.

Expected start date	**Upon availability of genome data, ie later in 2016**
Expected end date	**Minimum 2 years after anticipated end of 100K genome project, i.e. 2019**

Lead Applicant(s)	
Name	Robert Kleta and Daniel Gale
Post	Head of Centre (RK) and Clinician Scientist (DG)
Department	Centre for Nephrology
Institution	University College London
Current commercial links	N/A

Administrative Support	
Name	Winnie Han Langner
Email	w.han@ucl.ac.uk
Telephone	020 7472 6457

Subdomain leads		
Name	**Subdomain**	**Institution**
aHUS	David Kavanagh	Newcastle University
Amyloidosis	Julian Gillmore	University College London
CAKUT	Helen Stuart	Manchester University
Cystic kidney disease	John Sayer/Albert Ong	Newcastle/Sheffield Universities
Early-onset hypertension	Ben Walsh	University College London
Familial haematuria	Helen Storey	Guys and St Thomas' NHS Trust
Familial tubulointerstitial kidney disease in the young	Christine Gast/Tom Connor	Portsmouth NHS Trust/Royal London Hospital
Proteinuric renal disease	Moin Saleem	Bristol University
Renal tract calcification (or Nephrolithiasis/nephrocalcinosis)	Shabbir Moochhala	University College London
Renal tubular acidosis and other electrolyte disorders	Fiona Karet/Detlef Bockenhauer	University of Cambridge/ University College London
Validation and feedback	Maggie Williams	Bristol University
Bioinformatics	Horia Stanescu	University College London
Education and Training	Paul Winyard	University College London
Public and Patient Involvement	Tess Harris	PKD charity

Detailed research plan

Full proposal (total max 1500 words per subdomain)	
Title *(max 150 characters)*	**Genetic causes of kidney disease**

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).*

1) **Providing a diagnosis to people with kidney disease.** End stage kidney disease (ESKD) is a major cause of morbidity and mortality, accounting for up to 3% of the NHS budget and placing a major burden on those affected, including shortening and impairing their quality of life. While diseases not attributable to rare genetic variants (such as diabetes and atherosclerotic vascular disease) are thought to account for the majority of cases, monogenic disorders are known to be responsible in a substantial proportion of the remainder (for instance 6-10% of ESKD is known to be due to autosomal dominant polycystic kidney disease). Importantly, in up to 20% of UK patients receiving renal replacement therapy, the cause of kidney failure is recorded as "unknown" in the UK Renal Registry (Byrne, Steenkamp et al. 2010), and in a proportion of disease attributed to other causes (eg hypertension or congenital urinary tract anomalies) a monogenic disorder is detectable on genetic testing. Recent advances in molecular genetics have rapidly expanded the set of genes known to be responsible for such disorders, and further genes are likely to be identifiable in this population. A particular strength of the unbiased whole genome sequencing approach in this project will be the ability to identify di- or oligo-genic mechanisms, something that is known to be important in kidney diseases (including Alport Syndrome and nephronophthisis (Hoefele, Wolf et al. 2007; Mencarelli,

Heidet et al. 2015) but is hampered in clinical practice by the cost of step-wise gene sequencing, which usually results in stopping testing when the first likely pathological variant is identified. Current research shows that a monogenic defect can be identified in 20% of an unselected cohort of prevalent patients with end stage renal failure aged under 30 by sequencing a panel of 400 genes (van Eerde, van der Zwaag et al. 2015), and a clear molecular diagnosis can be established in ~30% of patients with a family history of unexplained kidney disease using whole exome sequencing (Gale, Connor et al. 2014). Providing a molecular diagnosis to patients will directly contribute to clinical care (*acceptable use 1*) by explaining why individuals developed disease, and may directly inform treatment decisions (eg with-holding immunosuppression where nephrotic syndrome is shown to be due to a process that is not steroid responsive). It will also identify cohorts of patients with a shared disease aetiology that could be recruited to clinical trials in the future (*acceptable use 2*). Feedback to clinicians and patients about the gene responsible for their disease will educate them, and the opportunity to use genetic information (for instance offering predictive testing to at-risk relatives) to inform decisions around living kidney donation will raise the profile of the utility and importance of genetic testing in the renal community (*acceptable use 4).* As well as being associated with improved quality and quantity of life, kidney transplantation is less costly than dialysis. Because some diseases can recur following transplantation, and due to risk of disease occurring in at-risk but ostensibly healthy relatives, lack of a secure diagnosis can sometimes prevent living related kidney donation. Securing a molecular diagnosis in some families can allow transplants to be performed that would otherwise be too high risk to perform (such as in unexplained atypical haemolytic uraemic syndrome), resulting in both reduced costs and improved quality and quantity of life for patients. We will document any transplants facilitated by the 100,000 Genomes project data (*acceptable uses 10 and 11*).

2) **Revealing genotype-phenotype correlation.** Studying molecular defects in larger cohorts of patients with a monogenic disease (*acceptable use 5*) can increase the information gained from a genetic test – an example is the knowledge that in polycystic kidney disease and Alport syndrome nonsense mutations are associated with worse prognosis (ie earlier age at onset of kidney failure) than are missense mutations in the same gene (Tsiakkis, Pieri et al. 2012; Cornec-Le Gall, Audrezet et al. 2013). In addition to helping to understand the biology of kidney disease, understanding genotype-phenotype correlation might improve prognostic information given to patients, will contribute to the Genomics England Knowledge Base (*acceptable use 7*) and may also help stratify patients for future interventional trials (*acceptable uses 2 and 6*).

3) **Revealing disease mechanisms.** As the genetic basis of diseases are identified, this enriches the knowledge base of gene (and hence protein) function (*acceptable use 7*). In particular, correlation of mutation with phenotype may provide insights into the role of particular proteins or protein domains. This information can, in turn, provide insight into disease mechanisms that may inform strategies to develop new treatments.

4) **Unravelling novel physiological processes.** Identification of novel proteins or protein domains that cause disease when disrupted can provide insights into the normal function of a gene, and hence protein, revealing new information about human biology (*acceptable use 8*).

5) **Identification of genetic modifiers.** Because kidney failure *in utero* is fatal, patients living with genetic kidney diseases have or had functioning kidney(s) for some of their life, with subsequent deterioration in those who go on to require renal replacement therapy to stay alive. In many genetic kidney diseases, rate of progression is highly variable, even where genetic background and primary mutation are shared (ie within families and among unrelated individuals with the same or similar mutations), indicating that modifiers (which

could be genetic or environmental) have an important impact on patients. An example of this is the appreciation of the importance of vasopressin receptor activation in the progression of polycystic kidney disease which has recently led to licensing of the first treatment to delay progression of this disorder (Torres, Chapman et al. 2012). In addition, it has also been shown that variation in *NPHS2* can influence the progression of Type IV collagen-associated diseases (Tonna, Wang et al. 2008). It is hoped that if further genetic modifiers can be identified, and the related mechanisms understood, this might provide new avenues for treatment even where the underlying genetic lesion cannot be corrected (*acceptable uses 5 and 8*). Such approaches may also provide benefit across a range of kidney diseases.

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. **Identification of rare variants in known kidney disease genes.** A rich dataset of gene-disease associations is already available in the published literature and public databases. We will identify the spectrum and frequencies of disease-causing variants in known kidney disease genes in patients with each renal disorder (comparing with individuals without kidney disease). This will be done by mapping sequencing data, annotating and filtering variants to remove those that are known to be common in the healthy population or predicted to have no effect on primary protein structure. Variants in genes known to be associated with the disorder for which patients were recruited will be compared with mutation data held in locus-specific disease databases to determine evidence of pathogenicity. Large databases of pathogenic complement gene and type IV collagen gene variants, among others, are curated by GeCIP members. Each phenotype-specific subdomain will assemble a virtual gene panel of candidate genes and variants within these genes will be filtered and inspected for evidence of pathogenicity. Information that will inform this analysis will include population frequency data, cosegregation analyses (depending on the penetrance and phenotype), evolutionary conservation, predicted effect on protein structure (including using crystal structures for modelling where available) and review by individuals with expertise in the relevant protein or gene. A validation and feed-back subdomain comprising clinical and molecular geneticists will liaise with the V&F GeCIP to provide an opinion regarding pathogenicity of these variants (in the context of detailed phenotype data for the patient(s) in which they have been identified) and assess whether the evidence of pathogenicity is strong enough to report to patients and for confirmation by the relevant Genomic Medicine Centre. Metrics concerning the proportion of patients in whom such variants can be identified will be collected and treating clinicians will be given the chance to feed-back on whether and how the information changed management (for instance changing treatment delivered to a patient or informing a decision by a relative to donate a kidney).
2. **Revealing genotype-phenotype correlation.** Data relating age to severity of disease (ie changes in serum creatinine over time or age at onset of end stage renal disease) are being collected, and this information will be correlated with genetic findings: In patients in whom a pathogenic genetic variant is identified, the predicted effect of mutations on protein structure (eg location of missense mutations or comparing patients with missense and truncating mutations) will be used to subdivide patient groups for comparison of disease severity.
3. **New gene-disease associations.** It is anticipated that not all the participants in the project will have a variant within the coding regions of known genes associated with their phenotype. A range of techniques, including region (or gene) aggregation tests (so called

collapsing methods) of multiple variants (burden test, adaptive burden test, variance component test, combined/omnibus tests, exponential-combination (EC) tests) (Lee, Abecasis et al. 2014), haplotype association analyses (Liu, Zhang et al. 2008), and phenotype similarity regression analysis will be used to identify clustering of rare variants at genomic loci in individuals with a shared phenotype or partially shared phenotype (Westbury, Turro et al. 2015). These methods are implemented in open source software (PLINK/SEQ) as well as in proprietary software (SVS – SNP and Variation Suite). Pathway analysis (ORA – Over Representation Analysis, FCS – Functional Class Scoring, PTB – Pathway Topology Based) (Garcia-Campos, Espinal-Enriquez et al. 2015) using literature mining tools (Mandloi and Chakrabarti 2015) as well as transcriptomic and proteomic datasets (both publicly available data e.g. KEGG, RegulonDB, STRINGDB, Pahter, GO, REACTOME, Ingenuity, Pathway Commons etc and data generated by GeCIP members) will be used to identify genes and mutations that are strong candidates for causing disease. Further studies to elucidate the mechanisms underlying newly identified gene-disease associations will be determined by the genes involved. The Renal GeCIP contains individuals with expertise in renal (glomerular and tubular) cell lines as well mouse and zebrafish genetic modification and phenotyping, and it is anticipated that as new mutations are identified they will be investigated in these well-established model systems.

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 Renal GeCIP will build and maintain close links with related GeCIPs, and agreement in has already been reached to collaborate with the fetal medicine subdomain of Paediatric GeCIP, the Health Economics GeCIP and the Cross-cutting infectious organism GeCIPs.

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

Training will be delivered in two main formats. The first will be a course on Genomics of Kidney Disease, which is a 2-3 day Renal GeCIP-run taught cause aimed at clinicians (including trainees) to provide clinically relevant information as to how advances in Genomics can be used improve and develop clinical practice (particular diagnosis and management of kidney disease). Commercial support has already been obtained for this course so the cost to attendees will be minimal, allowing broad participation.

The second format will be MD and PhD fellowships to allow individuals (especially clinical genetics and renal medicine specialist trainees) to develop expertise in, and contribute to analysis of, genomic data. This will include fellowships in Renal Genomics and additional studentships in functional studies of variants or genes identified across specific aspects of renal disease and biology, including podocyte biology, tubular biology, zebrafish models and mouse models. Fellows will be assigned projects and supervisors from across the Renal GeCIP.

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

The Renal GeCIP is a large group that comprises individuals with strong track records in contributing to understanding of biology relating to and disorders of the kidney. Individuals from the Renal GeCIP have identified the molecular basis of dozens of monogenic kidney diseases, and made significant contributions to the understanding of a great many more. A number lead research groups studying the mechanisms underlying these disorders and related systems (eg podocyte biology, complement, tubular physiology etc) and have appropriate laboratory infrastructure to investigate new genes and disorders. In addition, the Renal GeCIP has strong representation in computational biology, bioinformatics, cell biology and mouse genetics, with extensive experience of generation and phenotyping of numerous genetic models, including in high-throughput facilities such as MRC Mammalian Genetics Unit in Harwell. Individuals in the Renal GeCIP are involved in higher specialist clinical training programmes and supervise and mentor the next generation of renal physicians, some of whom are likely to exploit the opportunities offered by the 100,000 Genomes Project to pursue a higher research degree.

In addition to nominated individuals responsible for education and public/patient involvement, the Renal GeCIP comprises sub-domains responsible for data analysis, including a bioinformatics group and a sub-domain for each of the renal eligibility criteria. A validation and feedback sub-domain will coordinate expert interpretation of variants for feeding back to the GMCs.

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 Renal GeCIP has a validation and feedback sub-domain that is composed of molecular geneticists who are already responsible for interpreting genetic variants for currently available genetic tests in various clinically accredited genetics laboratories. By liaison with sub-domains (via respective leads) they will be able to access additional disease-and gene-specific expertise available within the Renal GeCIP in order to assess pathogenicity of variants. This sub-domain will be the point of contact for the Cross-cutting V&F GeCIP.

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

Kidney failure places a major burden on individuals affected and healthcare providers, accounting for approximately 3% of the NHS budget. A significant proportion of kidney failure (at least 20%) is currently unexplained and a further 15-25% is known to be caused by single gene defects. Currently, most renal patients have not received a genetic diagnosis, but determining the genetic cause of disease in an individual can be of benefit for the following reasons:

1. Providing an explanation to a patient or their family of why they are affected by the disease
2. Revealing the mode of transmission in a family which is not always possible to determine from the family history – this can have implications for reproductive decision-making by affected individuals and their relatives
3. Guiding treatment decisions: for example demonstrating that nephrotic syndrome has a genetic cause in an individual might prompt avoidance of harmful steroid or

immunosuppressive therapies.

4. Disclosing risk of relapse following cessation of therapy, or risk of relapse following kidney transplantation
5. Allowing living kidney donation by at-risk family members
6. Allowing prenatal or preimplantation genetic testing
7. Allowing enrolment into clinical trials of new therapies directed at defined monogenic disorders (for instance Alport Syndrome)

Researchers and the wider scientific and healthcare community will also benefit from identification of genetic basis of disease in renal patients for the following reasons:

1. Revealing the frequency of genetic diseases will allow better planning and provision of healthcare resources to cater for needs of the population, and may stimulate the development of patient support groups that can provide education and other support to patients with rare diseases
2. Identification of cohorts of patients with a shared disease aetiology and pathogenesis will allow design of, and recruitment to, adequately powered clinical trials to determine the effectiveness of new therapies.
3. Novel insights into outcomes and mechanisms linking gene changes with pathology will allow more accurate prognostication in future patients
4. Identification of novel genes will be instructive as to the biology – monogenic disorders can be viewed as a highly informative experiment of nature that can reveal the importance of hitherto unstudied genes, proteins or pathways and may lead to novel drug discovery.
5. The wealth of data provided to a group working together across different UK institutions will build relationships and collaborations across the scientific community. This might result in methodological developments.

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 commercial entities have expressed an interest in and willingness to collaborate, formal partnership agreements for the analysis and use of Renal GeCIP findings are yet to be agreed.

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

Byrne, C., R. Steenkamp, et al. (2010). "UK Renal Registry 12th Annual Report (December 2009): chapter 4: UK ESRD prevalent rates in 2008: national and centre-specific analyses." Nephron Clin Pract **115 Suppl 1**: c41-67.

Cornec-Le Gall, E., M. P. Audrezet, et al. (2013). "Type of PKD1 mutation influences renal outcome in ADPKD." J Am Soc Nephrol **24**(6): 1006-1013.

Gale, D. P., T. M. Connor, et al. (2014). Whole exome sequencing in familial kidney disease. American Society of Nephrology Congress, Atlanta, GA, USA, American Society of Nephrology.

Garcia-Campos, M. A., J. Espinal-Enriquez, et al. (2015). "Pathway Analysis: State of the Art." Front Physiol **6**: 383.

Hoefele, J., M. T. Wolf, et al. (2007). "Evidence of oligogenic inheritance in nephronophthisis." J Am Soc Nephrol **18**(10): 2789-2795.

Lee, S., G. R. Abecasis, et al. (2014). "Rare-variant association analysis: study designs and statistical tests." Am J Hum Genet **95**(1): 5-23.

Liu, N., K. Zhang, et al. (2008). "Haplotype-association analysis." Adv Genet **60**: 335-405.

Mandloi, S. and S. Chakrabarti (2015). "PALM-IST: Pathway Assembly from Literature Mining--an Information Search Tool." Sci Rep **5**: 10021.

Mencarelli, M. A., L. Heidet, et al. (2015). "Evidence of digenic inheritance in Alport syndrome." J Med Genet **52**(3): 163-174.

Tonna, S., Y. Y. Wang, et al. (2008). "The R229Q mutation in NPHS2 may predispose to proteinuria in thin-basement-membrane nephropathy." Pediatr Nephrol **23**(12): 2201-2207.

Torres, V. E., A. B. Chapman, et al. (2012). "Tolvaptan in patients with autosomal dominant polycystic kidney disease." N Engl J Med **367**(25): 2407-2418.

Tsiakkis, D., M. Pieri, et al. (2012). "Genotype-phenotype correlation in X-linked Alport syndrome patients carrying missense mutations in the collagenous domain of COL4A5." Clin Genet **82**(3): 297-299.

van Eerde, A. M., A. van der Zwaag, et al. (2015). "Targeted sequencing of 399 renal genes reclassifies primary disease diagnoses in young ESRD patients." Nephrology Dialysis Transplantation **30**(suppl 3): iii381.

Westbury, S. K., E. Turro, et al. (2015). "Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders." Genome Med **7**(1): 36.

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)*

Analyses will use phenotype data, filtered variant lists, VCF and BAM files for all individuals recruited under the renal eligibility criteria.

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

Successive filtering steps.

1. single variant analysis: which will imply the Identification of obvious disease causing variants using Ingenuity and an in-house High Throughput Sequencing Pipeline;
2. haplotype association analysis: for which we will use SVS and PLINK/SEQ;
3. collapsing methods for the identification of rare variants SVS, PLINK/SEQ, in house methods to be developed;
4. data driven (after the identification of possible candidate genes) pathway analysis: starting with the literature generation of pathways (PALM-IST, Pathway Assembly Literature Mining) and continuing with an array of first (ORA), second (FCS) and third generation (PT) pathway analysis tools, as appropriate for the task.

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)*

Key phenotype data will vary according to the phenotype for which participants are recruited. Serum creatinine/date of onset of ESRD will be needed for all patients. Key data points by phenotype are as follows:
Early onset extreme hypertension: electrolytes and blood pressure
Proteinuric renal disease: Kidney biopsy report, responsiveness to steroids
Familial haematuria: Kidney biopsy report; urinary protein:creatinine ratio
Atypical haemolytic uraemic syndrome: Kidney biopsy report; urinary protein:creatinine ratio
Cystic kidney disease: Imaging reports
CAKUT: Imaging reports
Renal tubular acidosis: Serum and urinary electrolytes; imaging reports
Renal tract calcification: Serum and urinary electrolytes; imaging reports

Alignment and calling requirements. *Please refer to the attached file (Bioinformatics for 100,000 genomes.pptx) for the existing Genomics England analysis pipeline and indicate whether your requirements differ providing explanation. (max 300 words)*

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

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)*

Most of the tools in the Embassy apps list (except the ones dedicated to alignment and mapping).

Ingenuity

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)*

BAM files and VCF files from previously investigated nephrology patients (WES or WGS).

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

It is anticipated that 20-40 cores will be needed for initial .vcf file-based variant and phenotype data analyses planned. Resources needed will depend on the numbers of patients recruited and whether additional analysis of raw data (.bam) files is needed.

Omics samples

Analysis of omics samples. *Summarise any analyses that you are planning using omics samples taken as part of the Project. (max 300 words)*

The initial emphasis will be focussed on availability of genome data. Utilizing omics samples will be driven by gene discoveries, sufficient recruits for various diseases, and therefore likely initiated after a significant number of samples have been accrued. An example would be proteomic analysis of a biopsy samples in patients with unexplained amyloid deposition, where peptides deposited in tissue staining for amyloid will be correlated with missense mutations to identify candidate amyloidogenic variants.

Data access and security	
GeCIP domain name	**Renal**
Project title *(max 150 characters)*	**Genetic causes of kidney disease**

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

☒ *Clinical care*

☒ *Clinical trials feasibility*

☒ *Deeper phenotyping*

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

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

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

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

☒ *Non hypothesis driven R&D - health*

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

☒ *Other health use - clinical audit*

☒ *Public health purposes*

☐ *Subject access request*

☐ *Tool evaluation and improvement*

Information Governance

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

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

Other attachments

Attach other documents in support of your application here including:

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