**Study Title: Viking Health Study – Shetland**

**An isolated population resource for identifying rare genetic variants influencing complex disease-related traits**

**Ethics Ref: 12/SS/0151 (SESREC2)**

**Date and Version No:** 7th April 2021, version 4.0

| | |
|---|---|
| **Chief Investigator:** | Prof James F Wilson |
| **Investigators:** | Prof Harry Campbell, Prof Sarah Wild, Prof Chris Haley, Prof Caroline Hayward, Dr Veronique Vitart, |
| **Sponsor:** | University of Edinburgh |
| **Funder (if applicable):** | Medical Research Council |
| **Signatures:** | [Redacted for website publication] |
| **Authors:** | Prof James F Wilson and Dr Shona M. Kerr |

All investigators are affiliated to the University of Edinburgh

## TABLE OF CONTENTS

## 1. AMENDMENT HISTORY

| Amendment No. | Protocol Version No. | Date issued | Author(s) of changes | Details of Changes made |
|---|---|---|---|---|
| Substantial amendment 03 (REC Ref SA4) | 3.0 | 28/2/2019 | Jim Wilson and Shona Kerr | 6.3.2 Inclusion criteria<br>For the genetic study, adults having at least two Shetlandic grandparents<br>change to:<br>For the genetic study, adults having at least two Shetlandic or Orcadian grandparents |
| Substantial amendment SA5 | 4.0 | 07/04//21 | Jim Wilson and Shona Kerr | 6.4.2 Study Assessments<br>Participants will be re-contacted and asked to provide a new sample of saliva, for DNA extraction, following a simple non-invasive method and a postal kit. |

## 2. SYNOPSIS

| | |
|---|---|
| **Study Title** | Viking Health Study - Shetland |
| **Study Design** | Quantitative Trait Locus Mapping Study using association, regional heritability, variant collapsing and other methods.<br>A questionnaire-based health survey is also included. |
| **Study Participants** | Volunteers from Shetland, with at least two grandparents from Shetland for the genetic study |
| **Number of Participants** | At least 2000 for the genetic study, plus as many others as respond for the health survey |

| | |
|---|---|
| **Planned Study Period** | 2 years of data collection |
| **Primary Objective** | Identify regions of the genome and variants in them that are associated with study traits, such as arterial stiffness, cognitive function, eye length, C-reactive protein, etc |
| **Secondary Objectives** | Establish a platform resource for health and population genetic research in Scotland with rich phenotyping, genome-wide genotyping and a biobank of tissue;<br>Carry out a health survey in Shetland. |
| **Primary Endpoint** | Genotype – phenotype correlations |
| **Secondary Endpoints** | Detailed information on health status and behaviours in Shetland |
| **Intervention (s)** | None |

## 3. ABBREVIATIONS

| | |
|---|---|
| CI | Chief Investigator |
| PI | Principal Investigator |
| PIL | Participant Information Leaflet |
| R&D | NHS Trust R&D Department |
| REC | Research Ethics Committee |
| SOP | Standard Operating Procedure |

## 4. BACKGROUND AND RATIONALE

### Overview

The aim of this project is to identify genes which influence risk factors for common diseases such as heart disease, diabetes, glaucoma and stroke. Finding the genes and variants in them which predispose to these diseases is the first step on the road to new treatments and methods of diagnosis. In recent years, genetic studies have had considerable success in identifying genes influencing the risk of these diseases. However, these studies can only investigate the effects of common genetic variants – which are carried by more than 10% of the population. It is becoming clear that rarer genetic variants are also important. Special populations with high kinship offer the opportunity to analyse rare variants in a very cost effective manner by taking advantage of the sharing of DNA among relatives. We therefore propose to carry out a genetic study in Shetland in order to identify genetic variants, both common and rare, predisposing to a variety of these diseases. The population of Shetland has a number of characteristics, including critically the very large number of relatives, which are favourable for the identification of these genes. An epidemiological survey will be carried out in Lerwick on 2000 people and quantitative traits underlying susceptibility to a number of common diseases measured. These traits will include measures of plasma lipids and inflammatory factors, blood pressure, eye traits, cognitive traits and body fat. At the same time this offers an opportunity to carry out a detailed health survey in the Shetland population including questions on physical activity, diet, smoking, etc. A dense genome-wide scan (using “gene chip” technology) will be carried out on participants with ancestry from Shetland and genome-wide association and other analyses performed. The data will be analysed together with similar data from other studies run by the investigators in Orkney, Croatia and Mainland Scotland – in order to provide sufficient power to identify genetic variants of interest.

### Importance

Cardiovascular and metabolic diseases are a major cause of morbidity and mortality globally and are costly for both health systems and national economies. It is estimated that cardiovascular disease (CVD) is responsible for 29% of all deaths globally and that over 180 million people worldwide have type 2 diabetes (T2D), over 200,000 of them in Scotland. CVD and T2D are major public health issues in the United Kingdom, and coronary heart disease (CHD) and stroke kill around 14,000 people each year in Scotland alone. It is well established that both lifestyle and heritable factors contribute to the risk of these diseases, but whilst lifestyle factors are relatively well understood, identifying genetic risk factors influencing complex disease remains a major public

health goal, in order to identify novel pathogenic mechanisms and to develop genetic risk profiling which will be sufficiently predictive to be useful in guiding clinical practice.

This proposal seeks to establish a new genetic epidemiological study in the isolated population of Shetland, taking advantage of new technological and statistical developments to study the role of both common and rare genetic variants in the aetiology of complex diseases of public health importance. The data and biobank will form a strategic platform resource for health and disease studies in Scotland. The health survey of all participants will inform NHS Shetland and assist with planning.

**Justification**

It has been known for decades that a substantial proportion of the risk of complex diseases like diabetes and heart disease is heritable. Despite the enormous progress over the last five years in identifying the genetic factors at play, the complexity in terms of the number of genetic variants involved, the typically rather small effect sizes for common variants and the potential for interactions among genes and with the environment mean that we still cannot predict with any accuracy who will become ill and who will not. This is partly due to the fact that the genes found to date only explain a tiny proportion of the heritability or risk of disease. As technology now allows us to investigate rarer genetic variants as well as those that are common in the population, this proposal seeks to quantify the degree to which rarer genetic variants explain the so called “missing heritability”. It will do this in a very cost efficient manner as we have developed statistical methods which allow the us to predict accurately the rare variants present in most individuals while sequencing only a subset. This is done by chip genotyping all subjects and using the patterns of DNA sharing with the sequenced individuals to predict which rare variants are present in unsequenced individuals. This approach can only be applied in an isolated population with high kinship, and we have experience from previous studies of Multiple Sclerosis that the population in Shetland are willing to take part in research and include many relatives. As the rarer variants tend to have much larger effect sizes they will be more important in determining individual risk and will also increase our knowledge of the mechanisms of disease. This study forms part of a broader group of similar studies focussed on this question, which run from the University of Edinburgh by these investigators and which will all be analysed together in order to have high power to answer the questions posed.

The health survey of all participants will inform NHS Shetland and assist with planning.

**Study population.**

The Northern Isles of Scotland (Orkney and Shetland) have been isolated from the rest of the British Isles by their geographic position at the extreme northern periphery

surrounded by the Atlantic Ocean and North Sea. They have a closely shared settlement history from Viking times to the modern era. The high degree of haplotype sharing is evident in principal components analyses (PCA) of Y chromosome variation: Orkney and Shetland stand side by side, isolated from all other sampling sites in the British Isles, showing shared types with Norway (Capelli 2003). An analysis of Y chromosome diversity in Shetland (Goodacre 2005) estimated the population genetic parameter $\theta_k$, that quantifies diversity, to be 18.2, which is lower than Orkney (23.4) and less than half the diversity seen in mainland Scotland and Ireland (39.1). Our initial survey using X chromosome short tandem repeat data revealed Shetland to have the lowest effective population size ($N_e$) among 10 areas of Scotland, less than one quarter of the $N_e$ of Edinburgh (Vitart 2005). Orkney showed only a slightly greater Ne, hence these island groups represent one of the most isolated populations in the UK, with a shared Scottish and Scandinavian inheritance. This is confirmed by the initial analyses of kinship we have performed – levels of genomic sharing (derived from genome-wide chip data) are very similar in Orkney (36.7 Mb shared on average after exclusion of first and second degree relatives) and Shetland (36.4 Mb). Participants in the genetic study will be volunteers with at least two grandparents from Shetland who will each receive a free health check consisting of the commonly used measures, including blood pressure, blood sugar, cholesterol levels, etc.

## 5. OBJECTIVES

### 5.1 Primary Objective

Identify regions of the genome and variants in them that are associated with study traits, such as arterial stiffness, cognitive function, eye length, C-reactive protein, etc. Although analyses will be performed using Shetland data alone, more power will come from meta-analysing these with all our other studies in a 20,000 subject meta-analysis.

### 5.2 Secondary Objectives

Establish a platform resource for future health and population genetic research in Scotland with rich phenotyping, genome-wide genotyping, consent for follow up and a biobank of plasma, serum, cells and urine.

Carry out a detailed survey of health status and behaviours in the population of Shetland.

## 6. STUDY DESIGN

### 6.1 Summary of Study Design

The Viking Health Study is a Quantitative Trait Locus mapping study using association, regional heritability, variant collapsing and other genetic analysis methods. The quantitative traits or intermediate phenotypes we study are risk factors for complex diseases such as heart disease, stroke or diabetes. The study measurements or traits will be compared to the genetic data for each individual using a number of different analyses with different strengths and weaknesses. Some are more appropriate for common genetic variants (e.g. genome-wide association) and some for rarer variants (e.g. regional heritability or variant collapsing methods).

The study is set in an isolated population as this allows us to take advantage of the high levels of genomic sharing which are present. This sharing means that we can accurately predict the whole genome sequences of many subjects using a scaffold of markers from the chip-based genome-wide scan. By whole genome or whole exome sequencing a representative subset of individuals it is possible to predict or impute the genomes of most of the others by reference to the scaffold markers. For instance a pair of first cousins will share about 12% of their DNA, and the scaffold of markers reveals exactly which 12% is shared. If one is sequenced, then 12% of the cousin's whole genome can be predicted. Because individuals will have a very high number of both near and distant relatives in the study in this way it is possible to build up nearly complete genome sequences for almost all individuals in a very cost efficient manner.

There is thus an element of family-based design in the study – many nuclear families and distant cousins will be participants, even if they individually volunteer. The inclusion of relatives allows a number of further genetic analysis which are not possible in studies of unrelated individuals.

Participants will visit two clinics, usually within one month of one another, thus the duration of participation is about one month. This includes recruitment, completing the postal health survey questionnaire, attending the 2 hour measurement clinic and the 20 minute venepuncture clinic.

The data will also be used for population genetic analyses, such as those looking at the genetic history of Shetland, Scotland and Europe more generally or estimating population genetic parameters such as effective population sizes, migration parameters and measures of differentiation from large scales such as across Europe to fine scales such as among different isles or parishes in Shetland. This will include analysis of Y chromosome data against surnames, to explore whether different kinds of surname (patronymic, placename, nickname, occupational) have different patterns of

ancestry, and estimates of the number of founders of the Shetland population using Y chromosome, mtDNA and autosomal data.

Subjects who are not eligible for the genetic study will be invited to complete the health survey questionnaire, thus their duration of participation will be a few days maximum, depending on when they fill in and return the anonymous questionnaire. The health survey aspect is a simple epidemiological survey of the population of Shetland.

## 6.2 Primary and Secondary Endpoints/Outcome Measures

The primary end point of this study is the correlation between the genotypes and phenotypes measured. That is the genetic markers which are associated or linked to the study traits. After the data collection phase of the study is complete (phenotyping), and the genetic data have been generated from the DNA samples (genotyping), the analysis phase will begin. The data will be analysed together in a number of ways to reveal which genetic variants or markers and hence which genes are influencing the various risk factors for disease which are being measured. The analysis will be performed using Shetland data alone and then meta-analysed with our other data and finally with data from groups around the world. Novel findings will be published in the peer reviewed literature.

The secondary endpoint is having detailed information on health status and behaviours in Shetland from the health survey.

## 6.3 Study Participants

### 6.3.1 Overall Description of Study Participants

Study participants will be “healthy volunteers” from the population of Shetland. There is no need for them to be free of disease, except if such disease would prevent them from giving consent or physically being able to be measured or give a blood sample. All participants will be aged 18 years or more. For the health survey, participants will live in Shetland. For the genetic study, participants are required to have at least two grandparents born in Shetland. There is no upper age limit. Males and females are eligible, but there is no need for equal numbers of each sex. From previous experience in Orkney and Shetland we know that many people will come forward directly to us after hearing about the study. At the same time NHS Shetland will invite large numbers to participate on our behalf.

### 6.3.2 Inclusion Criteria

For the health survey

- Participant is willing and able to give informed consent for participation in the study.
- Male or Female, aged 18 years or above.

- Living in Shetland

For the genetic study

- Participant is willing and able to give informed consent for participation in the study.
- Male or Female, aged 18 years or above.
- Has at least two Shetlandic or Orcadian grandparents

#### 6.3.3 Exclusion Criteria

The participant may not enter the genetic study if ANY of the following apply:

- Pregnancy
- Severe illness

### 6.4 Study Procedures

#### 6.4.1 Informed Consent

For participants who only take part in the health survey, consent will be given on the completed screening form which is returned to the Edinburgh office prior to the questionnaire being sent to them. For participants in the genetic study, written informed consent will be taken by trained research staff when the participant first attends a clinic visit. In almost all cases this will be the measurement clinic although in a small number of cases for practical reasons to do with scheduling a fasting blood sample, this might occur at the blood clinic. A participant information sheet will be sent to the participant along with the introductory letter from NHS Shetland. Potential participants will thus always have more than 24 hours to read the information and more usually about two weeks. The nurses will be familiar with all aspects of the study and be able to field any queries. The staff will read the items of the consent to the participant, checking that they understand and consent to each item, the participant will then initial the boxes he or she agrees with and sign and date the form at the bottom. The staff member will then sign and date the form to indicate that he or she has taken this consent in accordance with research protocol.

Only those who are able to give consent will take part in the study. All volunteers will speak English. No vulnerable people will take part in the study. Staff taking consent will be trained in this and will understand the ethical principles underpinning informed consent. The participant information sheets will explain the study clearly. The consent process will take about ten minutes.

### 6.4.2 Study Assessments

Participants in the genetic study will attend two clinic visits. In general the first will be the measurement clinic and the second will be the venepuncture clinic.

**Measurement clinic**

Participants will be asked to fast for one hour prior to attending the measurement clinic so as not to influence haemodynamic measures such as pulse wave analysis (which will occur about an hour into the clinic, meaning two hours have elapsed since food was ingested).

After taking informed consent the research staff will go through the health survey questionnaire with participants to check for completeness and assist with any queries. The questionnaire includes questions about the participants' health (e.g. Rose Angina questionnaire, EU respiratory health questionnaire, medical and surgical history), family health, lifestyle (including sun exposure), pigmentation and tanning ability, smoking, alcohol, diet, food preferences, physical activity (International Physical Activity Questionnaire), women's health, sleep (Munich Chronotype questionnaire), wellbeing and socioeconomic status. The identical questionnaire will be used for the health survey only participants and should take one hour to complete at home and less than ten minutes to check.

Research staff will then take demographic details, of first degree relatives and grandparents in order to allow reconstruction of participants' pedigrees. This will take ten minutes or less.

The trained research nurses will then take the participant through ten measurement modules:

1. Recording medications taken by participant. The participant will be asked to bring any medications they usually take with them, or the prescription to allow accurate recording of medicine, frequency and dose, including herbal medicines, vitamin pills, cod liver oil etc. 5 minutes.
2. Measurement of height, weight; head, waist and hip circumference and body fat%. Using stadiometer, scales, Holtain tape measures and Tanita bioimpedance scales. 10 minutes.
3. Measurement of lung function. Using a Microloop spirometer and Spida software, forced expiratory volume in one second, forced vital capacity and other measures will be recorded from two "good puffs". 5 minutes.
4. Measurement of heart rhythm using electrocardiogram. Subject will be kept supine on examination couch from now on. An electronic 30 second ECG will be taken according to standard procedures using Cardioview software. 10 minutes.

5. Measurement of blood pressure using OMRON digital sphygmomanometer. Blood pressure will be measured twice for each participant. 5 minutes.
6. Pulse wave analysis using SphygmoCor and tonometry of radial pulse. A sensitive pressure metre will be used to record the pressure waveform of the pulse and a ten second recording is taken from which various measures of central pressure and arterial stiffness are derived. Two pairs of two recordings are performed five minutes apart. 10 minutes
7. Measurement of carotid intima-media thickness and plaque scores using ultrasound. A portable Sonosite ultrasound is used to capture images of the common carotid artery 2 cm below the carotid bifurcation for the various measurements. 10 minutes.
8. Nonmydriatic photography of the retinal fundus. Subject sits up and is moved to darkened area for retinal photographs. After ten minutes dark adjustment of the pupils, photographs are taken of both eyes using the Canon CR-DGi retinal camera with 10 megapixel digital back. Photos are centred half way between macula and disc. 10 minutes.
9. Measurement of refractive error, eye length, eye pressure and corneal thickness using autorefractor, IOL master, tonopen and pachymeter. The ophthalmology module, although using four different instruments, is very quick as the participant sits in a wheeled chair and turns from one machine to the next sequentially. Spherical, cylindrical and axis refractive errors are first measured with a Canon RF10 autorefractor, for which the subject must fixate on a light. Orbital axial length and anterior chamber depth are then measured by optical coherence interferometry using the Zeiss IOL Master (this does not involve eye contact). After instilling proxymetacaine local anaesthetic drops, intra-ocular pressure is measured by applanation tonometry with a Tonopen – which uses tip covers to minimise the risk of cross-contamination. Finally central corneal thickness is measured with by ultrasound pachymetry (IOPac, Heidelberg engineering). These drops do not influence the ability to drive. 15 minutes.
10. Assessment of personality traits and cognitive function including memory, processing speed and executive function using standard instruments (e.g. from Wechsler Adult Intelligence Scale III). The extraversion and neuroticism scales of the Eysenck Personality Questionnaire are first assessed using the short form of the EPQ, followed by immediate paragraph recall (logical memory), then digit-symbol coding (processing speed), verbal fluency (executive function), Mill-Hill vocabulary and finally delayed paragraph recall. This is identical to the cognitive

battery used in ORCADES and the Scottish Family Health Study (Generation Scotland). 30 minutes.

In total the measurement clinic will last about two hours. Measurements will follow the same SOPs used in the ORCADES study.

Minims of 0.5% proxymetacaine will be prescribed by Dr Brian Fleck BSc (hons) MBChB MD FRCOph FRCSEd, Consultant Ophthalmologist, NHS Lothian. One drop will be used in each eye and a log kept of their administration. Weekly lists of those for whom proxymetacaine can be administered will be sent from Edinburgh to Shetland, these will be signed by Dr Fleck and will consist of the clinic list for measurement clinics.

**Venepuncture clinic**

*Consent.*

In the rare situations where the venepuncture clinic is the participants' first visit, consent will be taken as described above. The data sheets used by nurses to record the clinic visit, the number of tubes of blood successfully taken, etc will have a mail merged field to confirm that consent has been taken, which will be verified by asking the participant if they have been to a measurement clinic already.

*Venepuncture*

Venepuncture clinics will occur between 8.30 and 10 am with the participant having fasted since 10 pm the night before. Participants will sit in a chair suitable for venepuncture or lie on an examination couch if preferred and the procedure will be performed by a trained research nurse. Participants will provide a blood sample of 53 ml or six tablespoonsful using the Sarstedt monovette system, thus there is only one needle which has an adapter to allow multiple tubes to be drawn. In total 9 tubes will be used: two 9 ml EDTA, one 8.2 ml citrate, one 9 ml and two 4.9 ml serum gel and three 2.6 ml EDTA. The various tubes are for DNA, RNA, cells, biochemistry, haematology, plasma and serum. Venepuncture will take fifteen to twenty minutes.

*Urine*

A urine sample will be provided into a small beaker and decanted into a 12 ml tube. This urine sample is of course not first morning urine This will take five minutes.

*Saliva*

Participants with no or low quality DNA from their blood sample will be invited to provide a saliva sample of 2ml (one teaspoon), using a DNA Genotek Oragene saliva collection kit. This will be posted by the participant to the Edinburgh CRF laboratory, for extraction of DNA.

### 6.5 Definition of End of Study

The end of study is the date of the last clinic being run for the last participant.

## 7. INTERVENTIONS

There are no interventions.

## 8. SAFETY

None of the measurement procedures in this study can give rise to serious adverse events.

## 9. STATISTICS AND ANALYSIS

### 9.1 Number of Participants

The health survey will make use of as many subjects as return the questionnaire over and above the 2000 participants in the genetic study (whose questionnaires will also be part of the health survey). The Viking Health Study – Shetland is part of a larger programme of work to understand the genetic risk factors for complex disease and which will include a total of 20,000 research participants with genome-wide genotyping and similar phenotypes. These include 10,000 from the Scottish Family Health Study (Generation Scotland), 2000 from the Orkney Complex Disease Study (ORCADES), and 6000 from the Croatian studies of the islands of Vis (n=1000), Korčula (n=4000) and the city of Split (n=1000). The total of 2000 in Shetland has been chosen as a balance between the practicalities of the number we predict it will be possible to collect in two years and would be willing to take part, based on our experience in Orkney, and the desire to recruit the maximum number possible. In terms of statistical power the main analyses will involve all 20,000 participants. This provides very good power for association analysis such that there is 80% power to detect genetic variants at a 0.5% minor allele frequency which explain 0.2% of the phenotypic variance in the study trait. A number of other approaches will be used, including a novel “regional heritability” analysis based on linkage and for which a sample of 20,000 will also provide power to detect loci explaining less than 1% of the trait variance – this approach is more powerful than association when multiple variants at one locus influence the trait.

In terms of sequence imputation our analyses in Orkney show a predicted yield of 89% of variants either sequenced or imputed if 20% are actually sequenced. This is because of the high number of copies of any given haplotype in the population – on

average 8 copies per 1000 sampled (this same figure is 0.02 copies for 1000 Mainland Scots). Initial analysis of the genome-wide data we have for Shetland show that the levels of genomic sharing (mean of 36.4 Mb when all first and second degree kin are excluded) are very similar to that in Orkney (36.7 Mb) and hence similar efficiencies will be possible. We will genotype more than 10% of the indigenous population and sequence about 2% of them, both figures which are higher than the minima often quoted, e.g. by Kong et al (2008).

### 9.2 Analysis of Endpoints

Association analyses will be carried out in a mixed model to account for pedigree structure, using the mmscore algorithm in the GenABEL R library. This uses genomic estimates of realised kinship which are more accurate than traditional pedigrees. It is also an improvement over corrections using genomic inflation factors, as the correction is carried out at the individual level. Population stratification will be checked using nonparametric multidimensional scaling in R. Residuals will then be modelled with sex and age as fixed effects and SNP effects and other covariates (such as BMI, etc) tested in the regression. Traditional imputation and inverse-variance weighted meta-analyses across the different studies in the programme will be carried out using standard methods implemented in ProbABEL and MetABEL (Aulchenko 2010).

Exome or whole genome sequencing will be performed on a subset of 400 participants. These individuals will be chosen using an algorithm implemented in our ANCHAP software (Glodzik et al submitted) which allows ranking according to their length of useful sequence. The remaining subjects will be imputed using a combination of our long-range phasing/identity-by-descent-based methods (Glodzik et al submitted, Palin 2011) and traditional phasing and imputation with SHAPEIT2 and IMPUTE2 (Delaneau et al, in press, Howie 2012).

Analysis of rare variants. We will use a variety of approaches to predict the functional consequences of each variant. For instance, SIFT (http://sift.jcvi.org/) is a sequence homology-based tool that distinguishes damaging from tolerant amino acid substitutions and predicts their phenotypic effect on the basis of evolutionary conservation (Kumar 2009). POLYPHEN2 predicts whether a SNP is ‘benign’, ‘possibly damaging’ or ‘probably damaging’ on the basis of evolutionary conservation and protein structural data (Adzhubei 2010); trait values will be regressed on these categories. We will apply the rejected substitution score, which does not require functional information and so is also applicable to synonymous sites, and which has recently been shown to predict functionality for two Mendelian diseases in exome data (Cooper 2010). We will contrast the tails of the distribution of each quantitative trait, including analyses with a sample from the middle of the trait distribution to allow

differentiation of gain- and loss-of-function variants. We will deploy a number of methods which aggregate variation at one locus, for example using QuTie or GRANVIL (Morris 2010). Individuals and their respective QT values are split into two groups based on carriage (or not) of rare variants within the defined gene region, and the difference between the means of QT values from individuals is tested using regression and Student's t test. Chromosome- and genome-wide summaries are produced with lists of SNPs within potentially "significant" regions. This test assumes directionality, i.e. that the rare variant is the risk variant, and so some power is lost as non-functional alleles are included, but it would be possible to include bioinformatic information to filter variants and improve the test: most non-synonymous variants will shift phenotype in one direction. A simulation study showed the method has much higher power for resequencing studies such as proposed here, compared to genome scan data. Another collapsing method (Kernel Based Adaptive Cluster) combines variant classification and testing by applying adaptive weights to markers before testing and performs well in simulations (Li 2008). We shall also apply approaches such as C-alpha looking at the heterogeneity of variance across rare variant carriers and non-carriers (Neale 2011). 1000 Genomes data will be used to assemble a catalogue of genes which can tolerate loss-of-function mutations, such that novel loss-of-function mutations in these regions are down-weighted in ranking for replication; dbSNP will also be used to help filter out non-functional variants. Any variants cosegregating with the phenotype of interest will be up-weighted in the ranking.

## 10. ETHICS

### 10.1 Physical Risks

Venepuncture will be performed by experienced nursing staff but can cause a small level of discomfort in some people. Measurement of intra-ocular pressure and central corneal thickness will require corneal contact, which carries a very small risk of cross-infection. Local anaesthetic drops will be used, which may be mildly unpleasant to some people. These tests, which are performed routinely in ophthalmology departments, will be undertaken by trained staff using Standard Operating Procedures developed by Dr Brian Fleck, Queen Alexandra Eye Pavilion, Edinburgh and already successfully carried out on over 1200 recruits in Orkney. All other measures are non-invasive.

### 10.2 Test results

It is expected that some participants will be found to have abnormal biochemical test results or physiological measurements (e.g. blood pressure). Subjects will be asked at the beginning of the study for permission to give their GP these results for appropriate follow-up. In the case of genetic test results, it will be made clear in the project information sheet that these will not be returned. The study is principally dealing with genetically complex risk factors for disease, in which there are no simple predictive relationships between test result and disease. Any serious health issues identified will be immediately referred to the participant's GP (with participant consent). A recorded delivery letter must be used to inform the GP of any such issues.

### 10.3 Participant Confidentiality

Passwords will restrict access to computer data to a few project staff and paper-based data will be stored in locked cabinets in University premises. Personal data and biological samples will be identifiable only by a study number. The key linking the study number to the identity of participants will be held securely and will be accessible only to a few senior project staff. Only anonymised data and specimens will be shared with collaborating research groups. The study will comply with the Data Protection Act which requires data to be anonymised as soon as it is practical to do so. All staff will sign a confidentiality agreement.

### 10.4 Other Ethical Considerations

The setting of this study in an isolated population with high kinship brings special ethical responsibilities. These include the avoidance of stigmatisation of the community or certain parts of it and particular attention to pedigree data which by their nature can be difficult to anonymise. However, we have over ten years' experience in accounting for these issues and will take great care if presenting results which could lead to stigmatisation. For example we successfully argued that it was not important which particular Scottish populations we were studying in one paper on fine scale population structure (O'Dushlaine et al 2010). Only pedigrees which can effectively be anonymised, for example by being small and of a common structure, will be shared with collaborators and where possible sex will not be revealed.

Whole genome or exome sequencing is becoming a common part of clinical studies, and is described in a paragraph in the PIS where it is made clear that results will not be returned as they are mostly of unknown value and the testing is not performed in a laboratory accredited for clinical diagnostic testing.

## 11. DATA HANDLING AND RECORD KEEPING

Health questionnaires for participants not taking part in the genetic study are anonymous and will be sent directly to Edinburgh. Paper copies of measurement and other data for participants in the genetic study will be sent weekly from Shetland to Edinburgh using recorded mail. Electronic data, such as biochemical results will be transmitted using secure NHS email, via the study physician, Dr Sarah Wild. No participant-identifiable data will be sent by non-secure email. All staff involved in this project will follow University of Edinburgh guidelines for storage of data (e.g. on mobile devices or laptop computers).

All study data will initially be entered into a Microsoft Access database at the Centre for Population Health Sciences, University of Edinburgh. Most of the questionnaire data will be read in using optical mark recognition software and other readings will be manually entered into bespoke Access forms by an experienced data entry clerk. Some double entry will be performed to assess error rates. Various measurement devices allow direct export of the data which will then be read into the study database. The participants will be identified by a study-specific participant ID number in any database other than that used during the data collection process to manage clinic appointments. The name and any other identifying detail will NOT be included in any study data electronic file. Data will be managed by the project manager during the data collection phase but in the longer term by the senior data manager employed on the MRC “QTL in Health and Disease” programme. Genetic data will be managed by the data manager and held on University of Edinburgh servers. Dr J Wilson is the data custodian.

Data will be made available to bona fide researchers for collaboration after approval by the study investigators. In general analyses will be performed in Edinburgh and summary statistics shared for joint or meta-analysis but in some cases access to individual level data will be required. In these instances collaborator IDs will be used to further protect individual identities and collaborators will sign written material transfer agreements governing their ethical responsibilities.

## 12. FINANCING AND INSURANCE

### 12.1 Funding

The study is entirely funded as part of the quinquennial programme grant “QTL in Health and Disease” at the MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh in collaboration with the Centre for Population Health Sciences, University of Edinburgh.

### 12.2 Negligent Harm

The University of Edinburgh has arrangements in place to provide for harm arising from participation in the study for which the University is the Research Sponsor. All staff working on the study will be employees of the University of Edinburgh.

### 12.3 Non-Negligent Harm

The University of Edinburgh has arrangements in place to provide for non-negligent harm arising from participation in the study for which the University is the Research Sponsor. All staff working on the study will be employees of the University of Edinburgh.

## 13. PUBLICATION POLICY

The investigators will be involved in reviewing drafts of manuscripts and other publications arising from the study. The chief investigator will review abstracts for conferences and press releases. The investigators will decide the appropriate authorships based on individual contributions, whether the manuscript is led from Edinburgh or is part of an external collaboration. In most cases these policies will be carried out as part of the executive committee of the MRC quinquennial grant “QTL in Health and Disease”, which includes the investigators. Authors will acknowledge that the study was funded by the Medical Research Council and other contributors will be acknowledged.

## 14. REFERENCES

Adzhubei IA et al (2010) A method and server for predicting damaging missense mutations, Nat Methods 7: 248-9.

Aulchenko YS et al (2010) ProbABEL package for genome-wide association analysis of imputed data, *BMC Bioinformatics* 11: 134.

Capelli C et al (2003) A Y chromosome census of the British Isles, *Curr Biol* 13: 979-983

Cooper, GM (2010) Single-nucleotide evolutionary constraint scores highlight disease-causing mutations, Nat Methods 7: 250-1.

Delaneau O et al (2012) A linear complexity phasing method for thousands of genomes, *Nat Methods* 9: 179-81.

Glodzik D et al, in submission, Inference of shared ancestral haplotypes in human population isolates and their use to optimize sequencing studies, *Genet Epidemiol*.

Goodacre S et al (2005) Genetic evidence for a family-based Scandinavian settlement of Shetland and Orkney during the Viking periods, *Heredity* 95: 129-135.

Howie B et al (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing, *Nat Genet* 44: 955-9.

Kong A et al (2008) Detection of sharing by descent, long-range phasing and haplotype imputation, *Nat Genet* 40: 1068-75.

Kumar P et al (2010) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm, Nat Protoc 4: 1073-81.

Li B & Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data, Am J Hum Genet 83: 311-21.

O'Dushlaine C et al (2010), Genes predict village of origin in rural Europe, *Eur J Hum Genet* 18: 1269-70.

Morris AP & Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies, Genet Epidemiol 34: 188-93.

Neale BM et al (2011) Testing for an unusual distribution of rare variants, PLoS Genet 7: e1001322.

Palin K et al (2011) Identity-by-descent-based phasing and imputation in founder populations using graphical models, *Genet Epidemiol* 35: 853-60.

Vitart V et al (2005) Increased level of linkage disequilibrium in rural compared with urban communities: a factor to consider in association-study design, *Am J Hum Genet* 76: 763-72.