Non-CTIMP Study Protocol

Coronagenes – An international population cohort to investigate genetic susceptibility to the novel coronavirus (COVID-19)

	The University of Edinburgh and Lothian Health Board ACCORD The Queen's Medical Research Institute 47 Little France Crescent Edinburgh EH16 4TJ
Protocol authors	Prof James F. Wilson, Prof Albert Tenesa, Dr Shona M. Kerr
Funder	Medical Research Council
Funding Reference Number	632QTL RH1016 (MC_UU_00007/10) 632ATE RA4626 (HDR-9004, HDR-9003) 338DNA RI1023 (BBS/E/D/30002276)
Chief Investigator	Prof Albert Tenesa
Principal Investigator	Prof James F Wilson
Sponsor number	AC20050
REC Number	TBC
Version Number and Date	1.0, 08/04/2020

CONTENTS

LIST OF ABBREVIATIONS

ACCORD	Academic and Clinical Central Office for Research & Development - Joint office for The University of Edinburgh and Lothian Health Board
COVID-19	Coronavirus Disease 19
COVID-19 hg	The COVID-19 host genetics initiative
CI	Chief Investigator
CRF	Edinburgh Clinical Research Facility
DPIA	Data Protection Impact Assessment
DTC	Direct to Consumer
EHR	Electronic Health Record
FAQ	Frequently Asked Question
GCP	Good Clinical Practice
GENOMICC	Genetics of Mortality in Critical Care study
HGU	Human Genetics Unit
ISARIC-4C	International Severe Acute Respiratory and Emerging Infection Consortium - Coronavirus Clinical Characterisation Consortium
MRC	Medical Research Council
PIS	Participant Information Sheet
PI	Principal Investigator
REC	Research Ethics Committee
SARS-Cov-2	Severe Acute Respiratory Syndrome Coronavirus 2
SHARE	Scottish Health and Research Register
SOP	Standard Operating Procedure
QTL	Quantitative Trait Locus/Loci
WGS	Whole Genome Sequence

1 INTRODUCTION

1.1 BACKGROUND

This “Coronagenes” study seeks to recruit new participants from the general population during the COVID-19 pandemic into a large genetic epidemiological cohort. Recruitment (and subsequent saliva sampling and antibody testing) will take place remotely. Coronagenes re-purposes much of the infrastructure and expertise from the VIKING II study. This launched in January 2020 and paused in March 2020 due to the pandemic, with over 4,000 participants registered. In Coronagenes, data will be collected at baseline through an online questionnaire and longitudinally for UK participants through linkage to routine NHS data (and where possible other healthcare system data) in electronic health records (EHR). Participants will have the option to upload genetic data they possess via direct-to-consumer testing. Once recruitment is complete, the research team will use state-of-the-art statistical tools to study the role of genetic variants in the susceptibility to SARS-CoV-2 infection, and the severity of symptoms and outcomes. In the longer term, EHR linkage will allow study of other diseases of public health importance, and co-morbidities. The research will also gather data on psychological influences of the pandemic. Other relevant factors such as the distribution of variants across the worldwide population and different ethnicities (population genetics) can be analysed. Collectively, the data and biobank of samples will form a strategic resource for health, disease and population studies.

The research aims to understand the genetic susceptibility to coronavirus infection, and the genetic influences on severity of symptoms and outcomes, particularly for less severe cases. These will be missed in hospital-based studies. Milder cases are important to understand more fully the genetic risk factors and, not least, as well-matched controls for severe cases. Finding the genes and variants which predispose or protect people from coronavirus and its severe and mild symptoms, and the biology thus revealed, is a first step on the road to new ways of preventing and treating the disease both now and in future coronavirus outbreaks. In recent years, genetic studies have had enormous success in identifying genes influencing the risk of many complex diseases including infectious diseases, such as HIV and influenza. The global COVID-19 pandemic has created an unprecedented need to better understand the biology of the host response to the virus, all the more so as it is likely there will be further pandemics in the future. The unprecedented global focus and media coverage, however, provides an opportunity to motivate large numbers of volunteers to participate in medical research. We propose to recruit tens to hundreds of thousands of volunteers, including both those who have experienced flu-like symptoms and those who have not.

A “baseline” health and symptoms survey will be carried out on all volunteers by online questionnaire. This will cover a wide range of questions including existing conditions, medication, family health, smoking, flu-like symptoms, behaviour and impact of the pandemic. Participants will then be asked to complete follow-up questionnaires, aimed at capturing their COVID-19 symptoms as they occur and the effects of the confinement policies on health.

When public health policies implemented to contain the spread of COVID-19 permit it, a subset of volunteers will receive saliva sampling kits for “spit and post” DNA extraction. Finger prick (capillary blood drop) antibody testing kits will also be sent, to determine if they were infected with SARS-CoV-2. DNA will be extracted from saliva samples and a genome-wide scan (using “gene chip” technology) and (in the longer term) DNA sequencing will be carried out on the chosen subset of participants.

To allow immediate genetic analysis of genetic susceptibility to coronavirus infection, participants can also upload existing genome-wide data, which they already have acquired from direct-to-consumer testing companies. It is estimated that over 30M people worldwide have one of these tests. These are often sold for ancestry inference, but are very similar to those used in complex trait genetics. A variety of genomics analyses, including genome-wide association studies, will be performed. The data will be analysed both on their own and together with other studies worldwide, for instance through the COVID-19 host genetics initiative and with the ISARIC-4C and GENOMICC collections of severe COVID-19 cases. Working with other cohorts provides increased statistical power to identify and replicate

genetic variants. Understanding the risk factors influencing complex and infectious disease is a major public health goal, so as to identify new pathogenic mechanisms.

Data from the Symptom Tracker (covid.joinzoe.com), with over 1.5 million registrations in a few days shows that there is an unprecedented public appetite for research in this area. It does not matter where in the world volunteers live. The risks of participation are low and relate to data privacy and security, managed as described below (Section 8). The benefits are largely the advancement of scientific knowledge.

1.2 RATIONALE FOR STUDY

The rationale for this study was to make use of expertise, experience and infrastructure from the successful VIKING II project to create a new cohort for research on the coronavirus disease COVID-19. The scientific aims of this study are to understand the genetic susceptibility to coronavirus, particularly for less severe cases. A further aim is to assess the genetic contributions to the severity and duration of symptoms, and the rate of change of symptoms at the onset of symptoms. The mild cases will be an especially useful comparison set for severe cases coming from ISARIC-4C and GENOMICC. The detailed symptom information will also allow genetic analyses of phenotypes particular to COVID-19, such as anosmia (loss of sense of taste and smell). Identification of genetic risk factors is the first step to better understanding the host's underlying biology, revealing important pathways, some of which can potentially already be altered by approved medications for other conditions. Secondary aims are to use the detailed symptom tracking to test predictive models of outcomes in real time, and also to investigate the psychological effects of long term self-isolation during the pandemic.

In order to explore the genetic susceptibility across the broadest series of ancestral populations, we will recruit from around the world. It is not yet known whether certain populations are more susceptible or tend to experience a different course of the disease or if different genetic variants impart risk in different parts of the world. With sufficient samples we will be able to address all these questions.

This research will complement analysis in existing resources such as UK Biobank, by recruiting volunteers from across the globe, in many ethnic groups, experiencing different lockdown conditions at different times in different countries and cities. Moreover, the daily symptom questionnaire will provide detailed, quantitative phenotyping over and above the baseline questionnaire. Coronagenes will also complement other ongoing collections of patients in intensive care, such as GENOMICC, by covering the less severe end of the symptom spectrum, and with sufficient detail to investigate both the amplitude of symptoms, their duration and the rate of change in these symptoms prior to any hospitalisation. Analyses will also be conducted in collaboration with the COVID-19 Host Genetics Initiative to increase power (https://www.covid19hg.org/).

Participants can be located anywhere in the world. Two groups are of particular interest: (a) volunteers in the UK, where additional longitudinal information can be gathered through electronic health record (EHR) linkage; (b) volunteers with genotype data already available, through direct-to-consumer testing. These will allow genome-wide association analyses to be performed in real time, before lockdowns are relaxed, thereby allowing research that could otherwise not be performed. Volunteers who also belong to SHARE (Scottish Health Research Register (https://www.registerforshare.org/) are also of great interest as >80,000 have blood already stored, which can be accessed to extract DNA, obviating the need for saliva sampling. The Coronagenes volunteer cohort will be a useful resource for future studies into genetics and health. The outcomes expected from the research will be relevant to the overall strategy of both the MRC, HDR-UK and BBSRC (study funders) and to the NHS.

2 STUDY OBJECTIVES

2.1 OBJECTIVES

2.1.1 Primary Objective

The primary objective is to recruit tens to hundreds of thousands of volunteers, using online processes, who consent to be participants in epidemiological and genetic research. They will provide data and in some cases samples (for DNA and antibody testing for coronavirus) (Figure 1). Some participants will upload genetic data they already possess from direct-to-consumer testing companies. Their samples and data will be used to identify regions of the genome and variants in them that are associated with disease susceptibility, symptoms, and outcomes. Some analyses will be performed by meta-analysing data with other studies, such as cohort studies in the COVID-19 host genetics initiative (https://covid19hg.netlify.com/), ISARIC-C4 (https://isaric.tghn.org/) and GENOMICC (https://genomicc.org/).

2.1.2 Secondary Objectives

We shall also investigate the psychological effects of long term isolation. Coronagenes will create a resource for health and population genetic research with consent for recontact, ability to perform further phenotyping online, genome-wide genotypes, and a biobank of DNA samples. We will obtain consent for follow up and for linkage to routine electronic health records (UK only), thereby enriching the study phenotype and introducing a longitudinal element to the Study. Population-based cohorts such as Coronagenes, which combine genomic data with detailed phenotypes across many populations, provide a framework to study the population-specific landscape of coronavirus infection and the host's underlying genetic variation.

2.2 ENDPOINTS

2.2.1 Primary Endpoint

The primary end point of this study is to recruit to hundreds of thousands of people and use their samples and data to study correlations between genotypes and phenotypes, in particular related to COVID-19 (Figure 1). This endpoint will generate a platform resource for health and population genetics research with phenotyping, consent for recontact (for example through further online questionnaires), electronic health record linkage, genome-wide genotyping and a biobank of DNA. After the recruitment and questionnaire data collection phase of the study is complete (phenotyping), and genetic data have been gathered from direct-to-consumer participants or generated from the DNA samples (genotyping and eventually sequencing), the analysis phase will accelerate. The data will be analysed in a number of ways to reveal which genetic variants and hence which genes are influencing the risks health and disease, particularly COVID-19 infection, severity and outcomes. The analyses will be done both using Coronagenes data alone and also together with data from resources worldwide.

2.2.2 Secondary Endpoints

The secondary endpoint is to analyse collected data to better understand the psychological effects of long term isolation. Epidemiological models will be developed to predict outcomes (via electronic health records) from study phenotypes.

3 STUDY DESIGN

Coronagenes is primarily an observational study, with collection of saliva samples for DNA extraction and genetic analysis, and detailed phenotype data initially through questionnaires and longitudinally through NHS EHR data linkage (with consent). The phenotypes will be compared to the genetic data for each available individual using a variety of analyses. The Coronagenes study design is summarised in Figure 1. Green boxes are actions by participants, white by the Coronagenes management team, grey by the accredited laboratory and orange by the research group at the University of Edinburgh. Boxes with dashed outlines will occur in the future after lockdowns have been relaxed. The recruitment will take place online, is expected to last for up to three years and can continue until March 2023 with current funding. The minimum involvement by participants (following completion of the informed consent process) is filling in an online questionnaire. Broad consent is taken for use of data and DNA in future studies. Participants will also consent for contact, at a later date, by

members of the research team and linked to previous studies (e.g. UK Biobank). Separate applications for approval will be made for any future research project that requires re-contact of some or all Coronagenes participants as part of its protocol.

The data will also be used for population genetic analyses, such as assessing the genetic history of various populations.

The saliva sample will be entirely used up in the process of extraction of DNA. The resulting DNA sample will be stored under the custodianship of the University of Edinburgh. Genomics data derived from the DNA will be stored in databases and data archives under the control of the University of Edinburgh. On completion of recruitment, CHI/NHS number assignment and linkage (with PBPP approval in Scotland) to the electronic health/medical record of the participants will add a longitudinal component to the study. The study questionnaire data will not be directly merged with the NHS medical record data. Instead, the two datasets will be held separately at the University of Edinburgh, with access permissions granted on an individual basis.

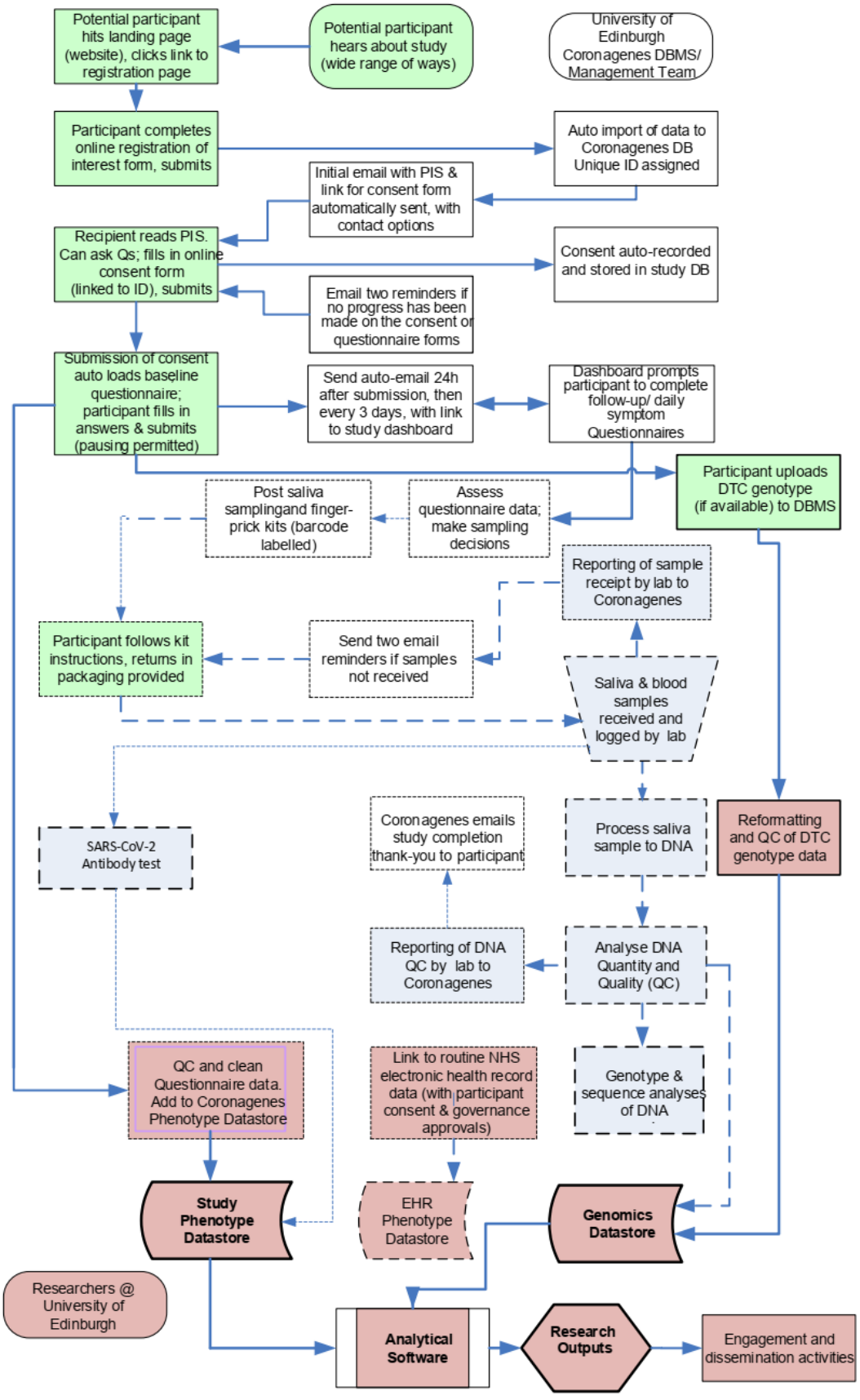

Figure 1 Schematic diagram of the Coronagenes study design

4 STUDY POPULATION

4.1 NUMBER OF PARTICIPANTS

The target is to recruit tens of thousands (up to hundreds of thousands) of volunteers from across the globe, who have access to the internet and are willing to report coronavirus symptoms. The recruitment period will begin immediately once all approvals and processes are in place, which is anticipated for April or early May 2020. The project is funded through a core award by the MRC to the QTL Programme at the MRC Human Genetics Unit which continues to 31st March 2023, and BBSRC and HDR-UK programmes that run up to early 2023. It is expected that the target number will be reached within 18 months of the start date.

All recruitment will take place online. We shall use all resources at our disposal to drive publicity about the study, for example via email lists at the University of Edinburgh and similar institutions, by an email to 280,000 participants in the Scottish Health Research Register (SHARE; https://www.registerforshare.org/; looked upon favourably by Prof Brian McKinstry, subject to formal application), Public Health Scotland, the European Respiratory Consortium (PI Chris Brightling), the Spirometa consortium (PI Martin Tobin), the ROHgen consortium (PI Jim Wilson), HDR-UK (PI Andrew Morris) and hopefully other public bodies. Through social (Facebook, Twitter and Instagram) and traditional media we hope to spread the word far and wide. Many may be willing to volunteer.

4.2 INCLUSION CRITERIA

Study participants will be volunteers either having shown flu-like symptoms or not. All participants will be aged 16 years or more. There is no upper age limit. Males and females are eligible, but there is no need for equal numbers of each sex. The inclusion criteria are made clear to participants in several places, including the Registration of Interest Form on the study website.

- Participant is willing and able to give informed consent for participation in the study
- Male or Female, aged 16 years or above
- Has access to the internet, to complete the questionnaires

4.3 EXCLUSION CRITERIA

Exclusion criteria are the lack of capacity to provide informed consent or being under 16 years of age. Those without access to the internet will also be excluded, as the consent and questionnaire will be completed online, so this is an unavoidable consequence of the recruitment method.
Potential volunteers who are already part of existing cohort studies or who have completed the COVID Symptom Tracker will still be eligible to join Coronagenes. Participants will be excluded if they do not meet all inclusion criteria.

4.4 CO-ENROLMENT

No co-enrolment is planned.

5 PARTICIPANT SELECTION AND ENROLMENT

5.1 IDENTIFYING PARTICIPANTS

The processes for identifying participants, responding to their interest, managing registration and consent for participation in the Coronagenes Study are illustrated in Figure 2.

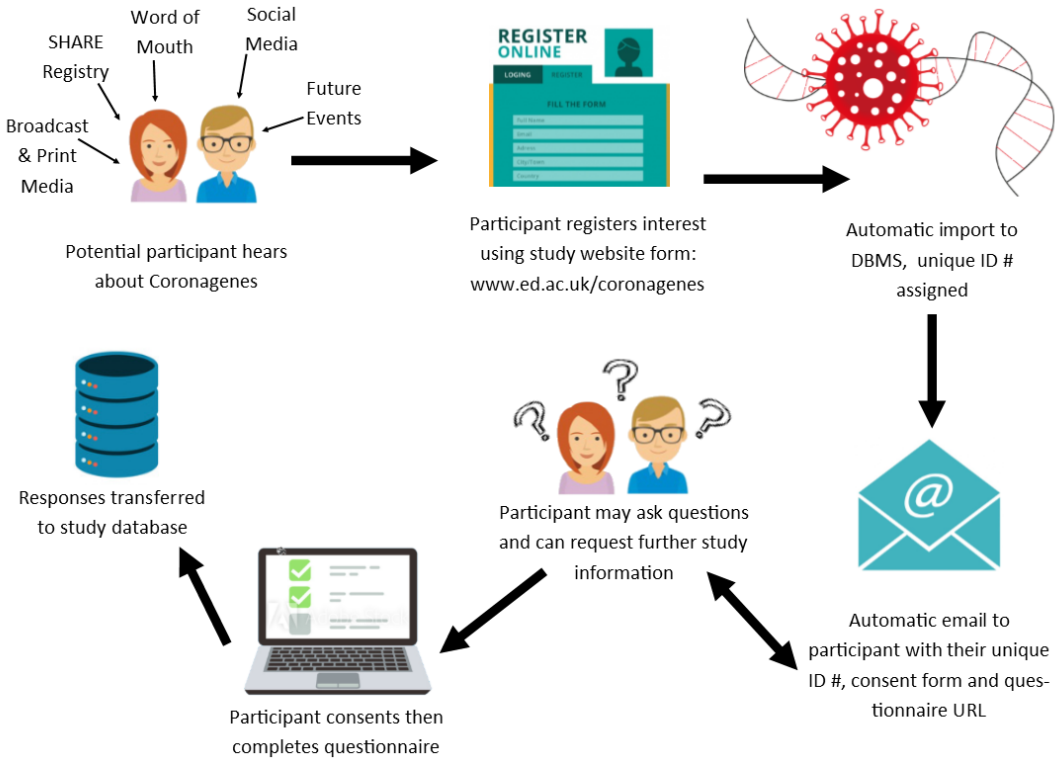

Figure 2. Schematic of Participant Identification and Enrolment in Coronagenes

The internet and social media increasingly offer ways to improve the reach, efficiency and effectiveness of recruitment efforts, at low cost. There are some privacy risks associated with online recruitment, such as inadvertent disclosure of information to companies that track online behaviour. However, these are mitigated by the secure and robust GDPR-compliant IT system for Coronagenes that has been established through the University of Edinburgh. Our system architecture is based on that of the REC-approved systems we have established in the VIKING II study (www.ed.ac.uk/viking/; CI Jim Wilson) and which have successfully been used to recruit thousands of individuals. This is also similar to the process successfully employed by other research projects using online recruitment in Scotland and throughout the UK, for example the GLAD Study of mental health (www.gladstudy.org.uk).

Information about the study, including the eligibility criteria, will be widely circulated in a variety of ways. A broad range of methods will be employed to explain the study and invite eligible people to engage and find out more. This will include print/online (e.g. national newspapers), broadcast media (helped by press release from the University of Edinburgh press office) and social media. A range of email lists will be used, with permission, to publicise the study. Twitter (@coronagenes), Facebook (@coronagenes) and Instagram (coronagenes) will be used by the Coronagenes management team to share information about the project (but not health or genetic information) with participants and potential participants.

We hope to recruit from SHARE, the Scottish Health Research Register, now with more than 280,000 people signed up. The direct volunteer contact allowed by SHARE registration means that the contact process is both rapid and efficient. We also plan to have Coronagenes featured in an email to all SHARE participants and propose to put an entry on the SHARE “Get Involved” tab (https://www.registerforshare.org/21_GetInvolved.html). Items publicising the study will also be posted on the Usher Institute, IGMM and MRC HGU websites (all University of Edinburgh), as well as the study website pages at www.ed.ac.uk/coronagenes/.

Further publicity will involve emails and social media posts through clubs, sports teams, university associations, international consortia, etc. and help to make potential participants aware of Coronagenes. Where possible and with permission, we shall seek to further increase awareness of our study using notifications from apps on mobile devices.

Potential participants will self-identify and as the first step will complete a short registration of interest in the study form on the Coronagenes website (screenshot enclosed). The Participant Information Sheet (enclosed) and a unique link for the Consent Form (screenshot enclosed) will then be emailed in response, automatically by the Coronagenes Database Management System (DBMS; Figures 1 and 2). If no action is made by the participant, up to two reminder emails will be sent by the Coronagenes team.

From previous experience in our recruitment studies, we anticipate that many people will come forward directly to us to enquire about participation after hearing about the study by any of the routes outlined above and word of mouth.

5.2 CONSENTING PARTICIPANTS

In September 2018, the Health Research Authority and the Medicines and Healthcare products Regulatory Agency published a joint statement on seeking consent by electronic methods (Ref 1). The basic principles outlined in this statement will be applied to the Coronagenes research study.

- Electronic methods for seeking, confirming and documenting informed consent are increasingly being adopted by sponsors and researchers. There is evidence that multimedia information is preferred by potential participants and can help to reinforce participant comprehension.
- Potential volunteers will be encouraged to consider the information sheets (and relevant supplementary information such as FAQs on the study website) for as long as they like before consenting, until the study ceases recruitment.
- 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 providing any additional information or answering queries about participation in the study will understand the ethical principles underpinning informed consent. The participant information sheet and supplementary information available online on the study website are designed to explain the study clearly and in lay terms.
- The consent mechanism online will take no more than ten minutes to read each item and tick a box, although it will take considerably longer than that to read and digest the study information provided and (if desired) ask questions or consider supplementary information.
- Participants will be consented after being provided the PIS through email and given the opportunity to ask questions (Figure 2). In the email (copy enclosed with the IRAS application) there will be a link, unique to the participant. This will take them to the entry/welcome page, then the consent form. Participants are asked to check their name, date of birth and sex on the welcome page, to confirm their identity is correct and matches the details they provided in the Registration of Interest form. They then provide their consent. Following consent, they move directly on to the questionnaire. In the first pages they provide some personal details, before answering a range of personal and family health and other phenotype questions including about flu-like symptoms (screenshots of the questionnaire enclosed).
- Any form of simple electronic signature is normally considered adequate to document consent in a non-CTIMP. In Coronagenes, we will use signatures that are tick box plus declarations, with the date recorded. The Consent Form is enclosed.
- Participants can click a link to view, save or print a copy of their consent form on the last page of the questionnaire. This link will also be available to each participant on the entry page of the questionnaire, if they return to it for any reason (for example pausing part way through completion) after they have completed the consent form. If the email has been mislaid by the participant, the study management team can send the participant a link to view the copy of their online consent on request.

5.2.1 Withdrawal of Study Participants

Participants are free to withdraw from the study at any point, by contacting the study team. Contact details are provided in the PIS, and online in the study website. If withdrawal occurs,

the primary reason for withdrawal (if provided) will be documented in the participant's entry in the participant management database. A reason does not have to be given.

There are two withdrawal options:

- 'No further contact'': This means the study team will no longer contact the participant with study updates or requests to join future studies. However, permission would still be in place to use the information and sample provided. The research team would also still be able to receive information from the health records of that participant.

- ''No further use'': In addition to ''no further contact,'' the study team would no longer be able to make data or samples from that participant available for future research. It will not be possible to remove results from research already performed or projects currently being performed.

6 STUDY ASSESSMENTS

6.1 STUDY ASSESSMENTS

The study procedures are illustrated schematically in Figure 1. There are no specific assessments; instead, participants complete questionnaires online. This is done at their pace, with an option to pause and resume at the same point subsequently. Volunteers are invited to complete short daily questionnaires on symptoms in addition to the baseline questionnaire. Volunteers with DTC genotypes will also be able to upload their data at the end of the questionnaire.

6.2 LONG TERM FOLLOW UP ASSESSMENTS

There is no active treatment phase in this study. The health of UK participants will be followed remotely through linkage to routine NHS data collected in the electronic health record.

6.3 STORAGE AND ANALYSIS OF SAMPLES

- For participants in SHARE, we shall access their stored blood sample to extract DNA and will not require a saliva sample.
- For a subset of volunteers without DTC genotyping available or a stored blood sample in SHARE, after the lockdowns have been relaxed, one sample of **saliva** will be collected, using a specialised container manufactured for this purpose and provided to the participant by post. This is a proven, all-in-one system for reliable, non-invasive collection, stabilisation and transportation of high quality DNA.
- Approximately 2 ml of saliva (roughly equivalent to a teaspoon) will be provided by participants, following the detailed instructions provided in the kit.
- Saliva samples in approved packaging will be sent, by participants, to an accredited laboratory, working to the principles of Good Clinical Practice for Laboratories.
- On receipt, the sample will be logged in the laboratory information management system (LIMS) and stored until extraction of DNA.
- The extracted DNA will be QC'd and concentration measured in the laboratory, then stored frozen in aliquots of master and working stocks. There should be enough material for a wide range of genetic analysis over many years, including whole genome sequencing.
- Following their established QC procedures, the lab will report any failures to extract sufficient quantity and quality of DNA for genetic analysis. These samples will not be used.
- Samples will be retained indefinitely (subject to renewed approval as and when required); consent is sought for storage of samples on this basis, under the control of The University of Edinburgh. Samples will be stored at the Clinical Research Facility at the University of Edinburgh,

- The research analysis of DNA samples will either take place in the Edinburgh Clinical Research facility lab or in collaborating research laboratories or technology service providers worldwide (with a Material Transfer Agreement in place where appropriate).
- For a subset of volunteers with no SARS-CoV-2 antibody test results, after the lockdowns have been relaxed, one capillary **blood drop** sampling kit will be collected using a specialised finger-prick system similar to those used by diabetics to monitor glucose levels. The blood testing kits shall be posted to the volunteers. We shall use a proven system for the collection, stabilisation and transportation of whole blood samples, once such a system has been validated, for instance for use in the NHS.
- About two drops of blood (about 100 microlitres) will be provided by participants, following the detailed instructions provided in the kit.
- Blood drop samples in approved packaging will be sent, by participants, to an accredited laboratory, working to the principles of Good Clinical Practice for Laboratories.
- On receipt, the sample will be logged in the laboratory information management system (LIMS) and stored until antibody testing.
- Following their established QC procedures, the lab will report any failures to measure the antibody titre accurately. These samples will not then be used (Figure 1).
- Blood drop samples will be entirely used up in the antibody testing.

7 DATA COLLECTION

- There is a single time point for direct initial data collection from each participant, at baseline when the volunteer completes the study questionnaire online.
- Subsequent questionnaires are offered to participants through the study dashboard (Figure 1)
- In the Coronagenes baseline and subsequent daily questionnaires (enclosed), standardised and validated tools and wording harmonised with other genetics research cohorts (e.g. ALSPAC, UK Biobank, Generation Scotland) are used wherever possible.
- The questions are clearly worded, with appropriate scales of measurement. Question ordering and style are carefully considered.
- The questionnaire is designed to work well on a range of devices (PC, tablet, smartphone).
- Completeness of data collection is ensured by automatic prompts for missing fields.
- Indirect data collection will take place longitudinally for the duration of the research, following linkage to the electronic health record of the participant

- Upload of direct-to-consumer genetic data

  Volunteers who have direct-to-consumer genetic data from gene chip arrays including hundreds of thousands of autosomal markers will be asked to upload the raw data files to the Coronagenes servers in order that they can be used for GWAS and other analyses.

  Over 30 million people worldwide have such genetic testing from companies including Ancestry, 23andMe or FTDNA. These were mostly for uses in genetic genealogy. A considerable ecosystem has developed in third party applications using these files, for example the now closed GEDmatch (gedmatch.com/), with over 1 million uploads, and DNA.land (https://dna.land/; Yuan et al. 2018 Nat Genet 50: 160-5). Active services include impute.me (https://www.biorxiv.org/content/10.1101/861831v1) and Sano Genetics (sanogenetics.com/). The International Society for Genetic Genealogy maintains a summary page of these services (https://isogg.org/wiki/Raw_DNA_data_tools).

  Raw data files (typically of 5-10 Mb in size after compression) will be uploaded to the Coronagenes servers, checked for computer viruses and then read into our quality control system prior to genetic analyses.

- Antibody test data will be returned from the testing lab directly to the Coronagenes team by a secure method using a unique barcode ID, rather than personal identifiers.

7.1 Source Data Documentation

For Coronagenes, source documents are the online questionnaire and the EHR.

Data (up to and including whole genome sequence) will also be derived from laboratory analysis of DNA samples isolated from the saliva samples provided by the participants.

Some participants will already have genome-wide genotype data and will upload this to our servers.

SARS-CoV-2 antibody test results will be generated in an accredited laboratory.

7.2 Case Report Forms

There is no case report form in Coronagenes. Data will be collected electronically and held in a secure study database.

8 DATA MANAGEMENT

8.1.1 Personal Data

The following personal data will be collected as part of the research:

- Participant name
- Participant postal address
- Participant work address
- Date of birth
- Email address
- Genetic data (genome-wide)
- Detailed health information including existing conditions, flu-like symptoms and psychological phenotypes

Personal data will be managed by the Coronagenes team at the MRC HGU. The data will be stored on a secure database within the University of Edinburgh. Access to the database is controlled through a secure application. Only authorised staff have access to the application. Access to the data is managed by user accounts and user level permissions within the application.

All volunteers will connect to the web data portal over the public internet and are forced to use a secure https connection to ensure that all traffic is encrypted. The only server visible to the internet and thus volunteers is the head node of the application server farm. We have achieved an A grade from the SSLLABS security report for the webserver. There are three firewalls involved: The University of Edinburgh's boundary firewall, the VMWare 'virtual server' firewall, and the local server firewall.
The web application server farm is comprised of five identical servers. The all run Windows Server 2019 and are hosted on the University of Edinburgh 'virtual server' platform. Using the Internet Information Server (IIS) Web Server Software, alongside additional features such as Application Request Routing; URL rewriting; and Web Proxying, we have built a load balancing farm of one head node and four application processing nodes. We believe that this will help us manage and meet the potential load the system will be under. We can easily add processing nodes if required. All servers have data snapshots and daily data backups taken.
The registration and participation data collection, from our volunteers, will be stored within an MS SQL database server. Each participant will have a unique ID assigned to them which will be used to identify their records within the database. Participants can only submit data to the system; they do not have any permissions to fulfil any other task. The application does not contain any administration or management functionality; this functionality has been hosted within a separate web application that has stricter security settings. The database server

maintains transaction logs and is backed up nightly. It is also hosted within the same 'virtual server' platform as the web application server farm.
During the project, Participants are asked to upload any relevant genetic data files that they own. For example, this can include their genomic data from public sequencing services such as "23andme.com". The web application supports file uploads. It restricts risky file types such as executable and scripts files. The application uses a functional account to connect to DataStore and receive, register, and store the uploads for each Participant. The application uses a strict file and folder structure to segment data from each participant.

Only approved and certified named staff have access to the data. All staff with access to personal data have completed GCP, MRC Research, GDPR and Confidentiality training.

Personal data will be stored for at least 20 years (as mandated by the funder, the MRC).

Once recruitment is complete, a subset of personal data for all participants will be securely transferred from the University of Edinburgh to NHS ISD through the eDRIS service, for matching to CHI numbers and subsequent EHR data linkage for participants in Scotland. The process for this will have PBPP approval. A similar linkage process will be adopted for participants in the rest of the UK, following guidelines from HDR-UK and other relevant bodies.

8.1.2 Transfer of Data

Identifying personal data collected or generated by the study will not be transferred to any external individuals or organisations outside of the Sponsoring organisation, apart from the transfer to NHS for record linkage described above (8.1.1).

De-identified data or samples are likely to be transferred outside University of Edinburgh/NHS Lothian, as part of research collaborations across the globe. Such transfers will comply with all legal requirements including GDPR. A data/material transfer agreement will be implemented through the University of Edinburgh Research Support Office contracts team when appropriate.

8.1.3 Data Controller

The University of Edinburgh is data controller.

8.1.4 Data Breaches

Any data breaches will be reported to the University of Edinburgh Data Protection Officer who will onward report to the relevant authority according to the appropriate timelines if required.

9 STATISTICS AND DATA ANALYSIS

9.1 SAMPLE SIZE CALCULATION

The investigators have many years' experience in the field of complex trait genetics and are thus well aware of the necessity for large sample sizes to discover the small genetic effects typically seen. Coronagenes will thus sample tens of thousands up to hundreds of thousands of individuals, many with DTC genotypes, to provide the necessary statistical power for association and other approaches. After the initial set-up costs, the collection of phenotypes is without significant cost, so the limiting factor is the number of DNA sampling kits, DNA extractions and genome scans which can be afforded. It is not known what proportion of participants will have DTC genotypes available: there is no per individual cost to genetic analyses for this subset. Symptoms may only be collected during the outbreak so the practical limit to study collection will be the length of the pandemic. We shall thus aim to recruit as many volunteers as reasonable within about 18 months or until the pandemic subsides. It is as yet unclear what proportion of the public will become infected and how many will be symptomatic and so very large numbers of recruits are required to provide sufficient cases of varying levels of severity, and examples of specific phenotypes such as anosmia.

The study has been peer-reviewed by the Director of the MRC Human Genetics Unit in April 2020.

Participants will be able to withdraw from the study at any time (see 5.2.1), but from previous experience we anticipate that the number of withdrawals will be small.

9.2 PROPOSED ANALYSES

A wide range of variables (phenotypes) collected from the Questionnaire and the EHR will be used for research. These will be analysed and reported in a variety of ways alongside the derived genetic data (genotype and DNA sequence), depending on the statistical design.

Reports of rates of recruitment, numbers of registrations of interest and numbers of participants completing each stage of recruitment will be compiled by the Coronagenes Management Team on a regular basis. In due course, the lab will report on receipt of saliva samples and the concentration and quality of the DNA prepared (see Figure 1).

- Missing data will be minimised through use of prompts on the online questionnaire. The Coronagenes Database Manager will flag any potentially spurious data to the CI and PI. If deemed spurious, it will be retained for the record, but removed from the main study dataset to be used in research.
- Data and samples from participants who withdraw will be removed from the Study Database, but it will not be possible to remove results from research already performed or projects in which analyses are already underway.
- There are no plans for pre-defined subgroup analyses, use of intention to treat analysis or any interim analysis other than the study monitoring described above.

Genome-wide association analyses will focus on case status (versus controls), and on mild case status (versus severe cases), as well as quantitative GWAS of symptom severity, symptom duration and rate of change of symptoms. Specific sub-analyses will focus on traits such as anosmia as well as psychological phenotypes.

10 ADVERSE EVENTS

This is a non-CTIMP online study, therefore physical adverse events are unlikely due to participation. Possible issues include discomfort during finger prick capillary blood sampling. Such blood drop sampling is performed millions of times daily by diabetics to assess glycaemic control.

Other risks of data-related adverse events are described in the Privacy Notice (https://www.ed.ac.uk/coronagenes/privacy-notice) and identified in the Research Data Protection Impact Assessment (DPIA, enclosed), alongside options for avoiding or mitigating each risk.

11 OVERSIGHT ARRANGEMENTS

11.1 INSPECTION OF RECORDS

Investigators and institutions involved in the study will permit study-related monitoring and audits on behalf of the sponsor, REC review, and regulatory inspection(s). In the event of audit or monitoring, the CI (Albert Tenesa) and PI (Jim Wilson) agree to allow the representatives of the sponsor direct access to all study records and source documentation. In the event of regulatory inspection, the CI and PI agree to allow inspectors direct access to all study records and source documentation.

11.2 STUDY MONITORING AND AUDIT

The ACCORD Sponsor Representative will assess the study to determine if an independent risk assessment is required. If required, the independent risk assessment will be carried out by the ACCORD Quality Assurance Group to determine if an audit should be performed before/during/after the study and, if so, at what frequency.

Risk assessment, if required, will determine if audit by the ACCORD QA group is required. Should audit be required, details will be captured in an audit plan.

Audit of the Investigator site, study management activities and study collaborative facilities may be performed.

12 GOOD CLINICAL PRACTICE

12.1 ETHICAL CONDUCT

The study will be conducted in accordance with the principles of the International Conference on Harmonisation Tripartite Guideline for Good Clinical Practice (ICH GCP).

Before the study can commence, all required approvals will be obtained and any conditions of approvals will be met.

12.2 INVESTIGATOR RESPONSIBILITIES

The Chief Investigator has overall responsibility for the conduct of the study and compliance with the protocol and any protocol amendments. The Principal Investigator is responsible for the conduct of the study at site. In accordance with the principles of ICH GCP, the following areas listed in this section are the responsibility of the CI or PI, as stated below. Responsibilities may be delegated to an appropriate member of study staff.

12.2.1 Informed Consent

The PI is responsible for ensuring informed consent is obtained before any protocol specific procedures are carried out. The decision of a participant to participate in clinical research is voluntary and should be based on a clear understanding of what is involved.

Participants must receive adequate information – an appropriate Participant Information Sheet and an Informed Consent Form will be provided. An explanation by the PI or qualified delegated person will be given to any participant on request. This would cover any or all of the elements specified in the Participant Information Sheet and Consent Form.

Participants will be given the opportunity to clarify any points they do not understand and, if necessary, ask for more information. The participant will be given sufficient time to consider the information provided. It will be emphasized that the participant may withdraw their consent to participate at any time.

The electronic informed consent form signed and dated by the participant will be stored in the study database, to confirm that consent has been obtained. Participants can click a link to view, save or print a copy of their consent form on the last page of the questionnaire and the entry page of the questionnaire, if they return to it for any reason after they have completed the consent form.

12.2.2 Study Site Staff

All study site staff will be familiar with the protocol and the study requirements. It is the PI's responsibility to ensure that all staff assisting with the study are adequately informed about the protocol and their duties.

12.2.3 Data Recording

The PI is responsible for the quality of the data recorded in the study database.

12.2.4 Investigator Documentation

There is no need for multiple investigator site files (ISFs) as this is not a multi-site study.

12.2.5 GCP Training

The majority of members of the Coronagenes management team understand the principles of GCP and are safe researcher trained, with current certificates for Good Clinical Practice in Clinical Research (non-drug trials). Members who are not yet trained will undertake training as soon as possible. GCP training status for the Principal Investigator and Project Manager are indicated in their CVs.

12.2.6 Confidentiality

All laboratory samples and other records will be given identifiers (Unique IDs, Figure 1) in a manner designed to maintain participant confidentiality. All records will be kept in a secure storage area of the University of Edinburgh Datastore with limited access. All data sharing and access will be for research purposes, with a range of safeguards in place (see Section 8 Data Management) to maintain participant confidentiality.

The CI and PI and study site staff will not disclose or use for any purpose other than performance of the study, any data, record, or other unpublished information, which is confidential or identifiable, and has been disclosed to those individuals for the purpose of the study. Prior written agreement from the sponsor or its designee must be obtained for the disclosure of any said confidential information to other parties.

12.2.7 Data Protection

All Investigators and study site staff involved with this study must comply with the requirements of the appropriate data protection legislation (including the General Data Protection Regulation and Data Protection Act) with regard to the collection, storage, processing and disclosure of personal information.

Computers used to collate the data will have limited access measures via user names and passwords.

Published results will not contain any personal data and will be of a form where individuals are not identified and re-identification is not likely to take place. All manuscripts using Coronagenes data are assessed prior to being submitted for publication. These processes are described in the Privacy Notice and the DPIA.

STUDY CONDUCT RESPONSIBILITIES

12.3 PROTOCOL AMENDMENTS

Any changes in research activity, except those necessary to remove an apparent, immediate hazard to the participant in the case of an urgent safety measure, must be reviewed and approved by the CI.

Amendments will be submitted to a sponsor representative for review and authorisation before being submitted in writing to the appropriate REC, and local R&D, for approval prior to participants being enrolled into an amended protocol.

12.4 MANAGEMENT OF PROTOCOL NON COMPLIANCE

Deviations and violations are non-compliance events discovered after the event has occurred. Any protocol deviation will be recorded in a protocol deviation log. Any protocol violation will be reported to the sponsor within 3 days of becoming aware of the violation. All protocol deviation logs and violation forms will be emailed to QA@accord.scot

12.5 SERIOUS BREACH REQUIREMENTS

If a potential serious breach is identified by the CI or delegates, the co-sponsors (seriousbreach@accord.scot) will be notified within 24 hours. It is the responsibility of the co-sponsors to assess the impact of the breach on the scientific value of the trial (although Coronagenes is not a trial), to determine whether the incident constitutes a serious breach and report to research ethics committees as necessary.

12.6 STUDY RECORD RETENTION

All study documentation will be kept for a minimum of 10 years from the protocol defined end of study point. When the minimum retention period has elapsed, study documentation will not be destroyed without permission from the sponsor.

12.7 END OF STUDY

The end of study is defined as the last participant's completion of the recruitment process (Figure 1), and is anticipated to be in 2021. However, the samples and data are expected to be used in research for many years into the future (in effect indefinitely) and it is anticipated that participants will complete further questionnaires in the future after appropriate REC and governance approvals.

The Investigators or the sponsor have the right at any time to terminate the study for clinical or administrative reasons.

The end of the study will be reported to the REC, and R+D Office(s) and co-sponsors within 90 days, or 15 days if the study is terminated prematurely. End of study notification will be reported to the co-sponsors via email to resgov@accord.scot

A summary report of the study will be provided to the REC within 1 year of the end of the study.

12.8 CONTINUATION OF TREATMENT FOLLOWING THE END OF STUDY

N/A (there is no treatment).

12.9 INSURANCE AND INDEMNITY

The co-sponsors are responsible for ensuring proper provision has been made for insurance or indemnity to cover their liability and the liability of the Chief Investigator, Principal Investigator and staff.

The following arrangements are in place to fulfil the co-sponsors' responsibilities:

- The Protocol has been designed by the Chief Investigator, Principal Investigator and researchers employed by the University and collaborators. The University has insurance in place (which includes no-fault compensation) for negligent harm caused by poor protocol design by the Chief Investigator, Principal Investigator and researchers employed by the University.

13 REPORTING, PUBLICATIONS AND NOTIFICATION OF RESULTS

13.1 AUTHORSHIP POLICY

Ownership of the data arising from this study resides with the study team. For collaborative research projects, the Principal Applicant will acknowledge the cohort and include as authors members of the University of Edinburgh research team identified by the Coronagenes Data Access Committee to have played a key scientific role in the generation of the Data and/or the

Materials in all publications relating to the Research. All research papers reporting results from this study will normally include the CI and PI as authors.

14 REFERENCES

1. https://www.hra.nhs.uk/about-us/news-updates/hra-and-mhra-publish-joint-statement-seeking-and-documenting-consent-using-electronic-methods-econsent/ accessed July 2019