UK10K Data Access Agreement

UK10K Project Data Access Agreement Version 9 (05/03/2015)

This agreement governs the terms on which access will be granted to the sequence and genotype data generated by the UK10K Consortium, together with accompanying phenotype data (the "Data" as defined below).

For the sake of clarity, the terms of access set out in this agreement apply to all of the User, Authorised Personnel within the User's research group, and the User Institution (as defined below). Within the agreement "You" and "Your" shall be construed to refer to all these. In signing this agreement, You are agreeing to be bound by the terms and conditions of access set out in this agreement.

Title of Research

Title of research (120 character maximum – this will be made public with your name and institution on access being granted)
Investigation of Stratified False Disovery Rate in Environments of Complex Correlation

Applicant

Name of applicant (User), including affiliation and contact details

Name with Title: Dr Joanne Knight

Position: Reader in Applied Data Science

Affiliation: Lancaster University

Institution's Legal Name (if differing from affiliation):

Institutional postal address: Lancaster University, South Dr, Lancaster University, Lancaster, LA1 4WA, United Kingdom

Institutional E-mail Address: ccole019@uottawa.ca

Date: 5/7/2016

Authorised Representative

Name of authorised representative of the User Institution, including affiliation and contact details:

Name with Title: Mrs Yvonne Fox

Position: Head of Research Services

Affiliation: Lancaster University

Institutional postal address: Lancaster University, B 58, Bowland Main, Lancaster, LA1 4YG, Lancashire, United Kingdom

Institutional E-mail Address: y.fox@lancaster.ac.uk

Data sets for which access is requested

I would like to apply for access to the following data-set(s)
UK10K_COHORT_ALSPAC REL-2012-06-02 (EGAD00001000740)
UK10K_COHORT_TWINSUK REL-2012-06-02 (EGAD00001000741)
UK10K_COHORT_TWINS REL-2011-12-01 (EGAD00001000194)
UK10K_COHORT_IMPUTATION REL-2012-06-02 (EGAD00001000776)

Description of proposed research

Please provide a clear description of the project and its specific aims in no more than 500 words. This should include specific details of what you plan to do with the data and include key references.

If applying to use datasets that have restrictions on the way that they may be used (e.g. must only be used to investigate a specific condition, or may not be used for control purposes), then please clearly state how you plan to use [named datasets] as controls, and that you will respect the [specified] constraints on the use of [named] non-control datasets.

The reproducibility and success of Genome Wide Association Studies (GWAS) is dependant upon the accurate and reliable correction for many millions of simultaneous tests; it has been theorized that stratifying tests based on prior information and applying False Discovery Rate Control can identify more real associations. This method, named Stratified False Discovery Rate (sFDR), has been gaining ground among genetics practitioners, though the conditions under which data may be optimally stratified to reduce the false discovery rate is not well understood. This investigation aims to use publicly available data and a novel R package (coR-ge) to improve our understanding of the mechanics behind stratified false discovery rate usage in GWAS studies, as well as the conditions under which

it may be optimally applied. We plan to use the control cohorts from the UK10K consortium in order to more realistically simulate human genomes which emulate true correlation patterns. We have broken this project into two specific aims.

Specific Aim 1: Determine whether stratification leads to a decreased false discovery rate under reasonable conditions.

Specific Aim 2: Estimate the effect of different genetic parameters on the efficacy of sFDR and on FDR in general.

The first of the specific aims will provide evidence for the use of sFDR in future genetic studies, while the second specific aim will allow resetters to better understand the mechanisms underpinning the use and misuse of this methodology.

User's publication record

Please list up to 5 relevant publications of which you were an author or co-author, demonstrating your experience and competence to analyse data sets of this type. If you do not have relevant publications please demonstrate your expertise and responsibility with respect to human subjects genetic data analysis.

Senior Author: Pouget JG, Gonçalves VF; Schizophrenia Working Group of the Psychiatric Genomics Consortium, Spain SL, Finucane HK, Raychaudhuri S, Kennedy JL, Knight J. "Genome-Wide Association Studies Suggest Limited Immune Gene Enrichment in Schizophrenia Compared to 5 Autoimmune Diseases." Schizophr Bull. 2016 May 30.

Senior Author: Masellis M et. al., "Dopamine D2 receptor gene variants and response to rasagiline in early Parkinson's disease: a pharmacogenetic study." Brain. 2016 May 13.

Senior Author: Gagliano SA, Ravji R, Barnes MR, Weale ME, Knight J. "Smoking Gun or Circumstantial Evidence? Comparison of Statistical Learning Methods using Functional Annotations for Prioritizing Risk Variants." Sci Rep. 2015 Aug 24;5:13373.

Knight J, et. al. "Conditional analysis identifies three novel major histocompatibility complex loci associated with psoriasis." Hum Mol Genet. 2012 Dec 1;21(23):5185-92.

(4th Placed author.) Genetic Analysis of Psoriasis Consortium & the Wellcome Trust Case Control Consortium 2 et. al. "A genome-wide association study

identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1." Nat Genet. 2010 Nov;42(11):985-90. PMCID: PMC3749730.

List of Authorised Personnel

List of Authorised Personnel (please refer to definitions section)

Registered User
Name: Christopher B. Cole
Title: Cooperative Education Student
Email: ccole019@uottawa.ca

Definitions

Consortium means the UK10K Consortium, a list of which can be found on the study website www.uk10k.org.

Data: means all and any human genetic data including phenotype data obtained from the managed access datasets of the UK10K Project.

Data Producer(s): The Wellcome Trust Sanger Institute (WTSI) and collaborators within the UK10K project, responsible for the development, organisation, and oversight of the Data.

Data Subject means a person, who has been informed of the purpose for which the Data is held and has given his/her informed consent thereto.

Collaborator: A collaborator of the User, including both someone working at a different institution from the User Institution and someone working in a separate research group in the same User Institution.

Publications: Includes, without limitation, articles published in print journals, electronic journals, reviews, books, posters and other written and verbal presentations of research.

Research Purposes: shall mean research that is seeking to advance the understanding of genetics and genomics, including the treatment of disorders, and work on statistical methods that may be applied to such research. Further specific conditions apply to particular data sets as listed in Appendix A.

User: An applicant having signed this Data Access Agreement, whose User Institution has co-signed this Data Access Agreement, both of them having received acknowledgement of its acceptance.

User Institution: Institution(s) at which the User is employed, affiliated or enrolled. A

representative of it has co-signed this Data Access Agreement with the User and received acknowledgement of its acceptance.

Authorised Personnel: Additional individuals who have an affiliation within the research group of the User at the User Institution including postdocs, students and any visitors. All Authorised Personnel must have an email address within the User Institution. If multiple research groups within the same institution require access they must each apply. Core IT and other administrative personnel at the User Institution who need to have access to the Data for data security and management purposes are automatically treated as Authorised Personnel and bound by the institutional acceptance of this agreement,

Terms and Conditions

In signing this agreement, the User and the User Institutions(s):

1. Agree to only use the Data for Research Purposes, subject to any data set specific conditions listed in Appendix A, according to the consent obtained from sample donors.

2. Agree to preserve, at all times, the confidentiality of the information and Data. In particular, you undertake not to use, or attempt to use the Data to compromise or otherwise infringe the confidentiality of information on Data Subjects.

3. Agree to protect the confidentiality of Data Subjects in any Publications that you prepare by taking all reasonable care to limit the possibility of identification.

4. Agree not to attempt to link or combine the data provided under this agreement to other information or archived data available for the data sets provided, even if access to that data has been formally granted to you, or it is freely available without restriction, unless specific permission to do so has been received from the relevant access committee(s) or sample custodians.

5. Agree not to transfer or disclose the Data, in whole or part, or any material derived from the Data beyond that in Publications, to any non-authorised personnel. Should the User or the User Institution(s) wish to share the Data with a Collaborator, the Collaborator must complete a separate *Application for Access to the Data*.

6. Agree to use the data for the approved research purpose and project described in your application; use of the data for a new purpose or project will require a new application and approval.

7. Accept that Data may be reissued from time to time, with suitable versioning. If the reissue is at the request of sample donors and/or other ethical scrutiny, You will

remove earlier versions of the Data from subsequent analysis and publication, and destroy/discard the earlier version unless obliged to retain data for archival purposes in conformity with Institutional policy.

8. Agree to abide by the terms outlined in the “UK10K Project Publications Policy” (Appendix B). This includes respecting the moratorium period for Data Producers to first Publication of report(s) describing and analyzing the Data.

9. You agree to acknowledge in any work based in whole or part on the Data, the published paper from which the Data derives, the version of the Data, and the role of the UK10K Consortium and the relevant primary collectors and their funders. Suitable wording for such acknowledgement is provided in the “UK10K Publications Policy”.

10. Agree that the UK10K Consortium, the original Data producers, Data depositors, copyright holders, and all other parties involved in the creation, funding or protection of any part of the Data supplied:

a) make no warranty or representation, express or implied as to the accuracy, quality or comprehensiveness of the Data;

b) exclude to the fullest extent permitted by law all liability for actions, claims, proceedings, demands, losses (including but not limited to loss of profit), costs, awards damages and payments made by the Recipient that may arise (whether directly or indirectly) in any way whatsoever from the Recipient’s use of the Data or from the unavailability of, or break in access to, the Data for whatever reason and;

c) bear no responsibility for the further analysis or interpretation of these Data.

11. Understand and acknowledge that the Data is protected by copyright and other intellectual property rights, and that duplication, except as reasonably required to carry out Your research with the Data, or sale of all or part of the Data on any media is not permitted.

12. Recognise that nothing in this agreement shall operate to transfer to the User Institution any intellectual property rights relating to the Data.

13. Accept that the User Institution has the right to develop intellectual property based on comparisons with their own data, but may not make intellectual property claims on the Data nor use intellectual property protection in ways that would prevent or block access to, or use of, any element of the Data, or conclusion drawn directly from the Data.

14. You agree that you will submit a report to the Data Access Committee, if requested, on completion of the agreed purpose. The Data Access Committee agrees to treat the report and all information, data, results, and conclusions

contained within such report as confidential information belonging to the User Institution.

15. If results arising from the User and the User Institution(s) use of the Data could provide health solutions for the benefit of people in the developing world, the User and the User Institution(s) agree to offer non-exclusive licenses to such results on a reasonable basis for use in low income and low-middle income countries (as defined by the World Bank) to any party that requests such a license solely for uses within these territories.

16. Agree to destroy/discard the Data held, once it is no longer used for the approved research, unless obliged to retain the data for archival purposes in conformity with Institutional policy.

17. Agree to update the list of Authorised Personnel to reflect any changes or departures in affiliated researchers and personnel within 30 days of the changes made. These changes can be made by emailing info@uk10k.org.

18. Agree to distribute a copy of this agreement and explain its content to any person mentioned in the list of Authorised Personnel, including any additions made according to paragraph 17.

19. You will notify the Data Access Committee as soon as You become aware of a breach of the terms or conditions of this agreement.

20. Accept that this agreement will terminate upon any breach of this agreement by the User, the User Institution(s) or any Authorised Personnel listed in this application document. In this case, You will be required to destroy/discard any Data held, including copies and backup copies.

21. Accept that it may be necessary for the UK10K Consortium or its appointed agent to alter the terms of this agreement from time to time. In this event, the UK10K Consortium or its appointed agent will contact You to inform You of any changes, and You may be required to enter into a new version of the Agreement.

22. If requested, You will allow data security and management documentation to be inspected to verify that they comply with the terms of this agreement.

23. Understand that this agreement (and any dispute, controversy, proceedings or claim of whatever nature arising out of this agreement or its formation) shall be construed, interpreted and governed by the laws of England and Wales and shall be subject to the exclusive jurisdiction of the English courts.

Agreement

I have read and agree to abide by the terms and conditions outlined in the Data Access Agreement.

Appendices

I have read and understood the Appendices listed at the end of this page.

Appendix A: Data Set Specific Conditions

Version 8

The sample sets for the UK10K Project were pre-existing prior to the start of the project. While they have all been approved for sequencing and use in the project, and for the resulting data to be used by others according to the UK Data Sharing Policy which this access agreement implements, in some cases the original consents restrict the uses to which the data can be put.
For the sake of clarity we list here all data sets, and any restrictions on use that apply. We also list the correct way to reference the origin of each sample set, as required in acknowledgements (see Appendix B).

Some of the UK10K disease (not cohort) studies have Research Ethical Committee approval to feedback to individual research participants genetic results that cause the clinical phenotype that is being studied. We encourage researchers who believe that they have identified a causal variant(s) for the disease under investigation by the UK10K project to contact the UK10K Project at info@uk10k.org who will ensure that the information is passed on to the relevant sample custodian, for their consideration.

Please note that NONE of the UK10K projects have Research Ethical Committee approval to feedback to individual research participants genetic results that do not pertain to the clinical phenotype under investigation (so-called 'Incidental Findings'), and so such results SHOULD NOT be returned to the Data Access Committee, or directly to members of the UK10K project or sample custodians.

If the data sets that you are requesting access for are not listed below, then you must obtain a more recent version of this appendix from http://www.uk10k.org/data_access.html and refer to that in your access application.

UK10K_COHORTS_TWINSUK
EGA Study ID: EGAS00001000108. Please refer to the EGA for this study's Dataset IDs.

Brief description:	The TwinsUK samples will be part of the cohort study and will undergo whole genome sequencing.
Conditions:	No additional constraints.
Data can be used as controls:	Yes.
Acknowledgement:	The TwinsUK Cohort

UK10K_COHORT_ALSPAC
EGA Study ID: EGAS00001000090. Please refer to the EGA for this study's Dataset IDs.

Brief description: The Avon Longitudinal Study of Parents and Children (ALSPAC) is a two-generation prospective study. Pregnant women living in one of three health districts in the former county of Avon with an expected delivery date between April 1991 and December 1992 were eligible to be enrolled in the study, and this formed the initial point of contact for the development of a large, family based resource. Information has been collected on the children and the mothers through retrieval of biological materials (e.g. antenatal blood samples, placentas), biological sampling (e.g. collection of cord blood, umbilical cord, milk teeth, hair, toenails, blood and urine), self-administered questionnaires, data extraction from medical notes, linkage to routine information systems and at repeat research clinics.

http://www.bristol.ac.uk/alspac/researchers/data-access/

Conditions: As above, users may not pass the data to any other party, attempt to identify any individual or merge the data with ALSPAC data from any other source.

Data can be used as controls: Yes.

Acknowledgement: As on publication check list (http://www.bristol.ac.uk/media-library/sites/alspac/documents/research/ALSPAC%20Publishing%20papers%20checklist_v21_190215.docx)

UK10K_NEURO_MUIR

EGA Study ID: EGAS00001000122. Please refer to the EGA for this study's Dataset IDs.

Brief description: The sample selection consists of subjects with schizophrenia (SZ), autism, or other psychoses all with mental retardation (learning disability). The samples were initially collected under the leadership of Walter J Muir (deceased), now with Prof. Blackwood, Dr McKechanie and Prof McIntosh as custodians. These subjects represent the intersection of severe forms of neurodevelopmental disorders, appear to have a higher rate of familiality of SZ than typical, and are likely to have more serious and penetrant forms of mutations.

Conditions: There are no additional constraints on the analyses that can be carried out.

Data can be used as controls: Yes.

Acknowledgement: Edinburgh MR-psychosis samples.

UK10K_NEURO_EDINBURGH

EGA Study ID: EGAS00001000117. Please refer to the EGA for this study's Dataset IDs.

Brief description: This sample set consists of subjects with schizophrenia recruited from psychiatric in-patient and out-patient facilities in Scotland. All diagnoses are based on standard research procedures and family histories are available. Patients have IQ>70 and the cohort includes the following groups: 100 cases with detailed clinical, cognitive and structural and functional neuroimaging phenotypes; 138 familial cases who are the probands of families where DNA has been collected from other affected members; 162 unrelated individuals. In most cases patients and their families may be re contacted to take part in further studies.

Conditions: There are no additional constraints on the analyses that can be carried out

Data can be used as controls: Yes.

Acknowledgement: Edinburgh Schizophrenia Samples.

UK10K_NEURO_ASD_SKUSE

EGA Study ID: EGAS00001000114. Please refer to the EGA for this study's Dataset IDs.

Brief description: This sample set of UK origin consists of clinically identified subjects with Autism Spectrum Disorders, mostly without intellectual disability (ie. Verbal IQs >70). The subjects represent children and adults with Autism, Asperger syndrome or Atypical Autism, identified according to standardized research criteria (ADI-algorithm, ADOS). A minority has identified comorbid neurodevelopmental disorders (e.g. ADHD). Family histories are available, with measures of broader phenotype in first-degree relatives.

Conditions: The data can only be used for research of autism spectrum disorders.

Data can be used as controls: No.

Acknowledgement: Institute of Child Health & Great Ormond Street Hospital Autism Families Study.

UK10K_NEURO_ASD_TAMPERE

EGA Study ID: EGAS00001000115. Please refer to the EGA for this study's Dataset IDs.

Brief description: The Tampere Autism sample set consists of samples from Finnish subjects with ASD (autism spectrum disorders) with IQs over 70 recruited from a clinical centre for the diagnosis and treatment of children with ASD.

Conditions: These Finnish autism samples can only be used for the research of autism spectrum disorders and no other studies.

Data can be used as controls: No.

Acknowledgement: Tampere University Hospital Autism Study data set.

UK10K_NEURO_ASD_BIONED

EGA Study ID: EGAS00001000111. Please refer to the EGA for this study's Dataset IDs.

Brief description: The BioNED (Biomarkers for Childhood onset neuropsychiatric disorders) study has been carrying out detailed phenotypic assessments evaluating children with an autism spectrum disorder. These assessments included ADI-R, ADOS, neuropsychology, EEG etc. There are 56 DNA samples from this study (25 extracted from blood).

Conditions: The data can only be used for Autism spectrum disorder and ADHD research. The data must be kept confidentially and securely.

Data can be used as controls: No.

Acknowledgement: Biomarkers in Neurodevelopmental Disorders (BioNed) Study

UK10K_NEURO_ASD_MGAS

EGA Study ID: EGAS00001000113. Please refer to the EGA for this study's Dataset IDs.

Brief description: The MGAS (Molecular Genetics of Autism Study) samples are from a clinical sample seen by specialists at the Maudsley hospital and who have had detailed phenotypic assessments with ADI-R and ADOS.

Conditions: The data can only be used to identify genes for Autism Spectrum Disorders. The data must be kept confidentially and securely.

Data can be used as controls: No.

Acknowledgement: The Molecular Genetics of Autism Study.

UK10K_NEURO_FSZNK

EGA Study ID: EGAS00001000119. Please refer to the EGA for this study's Dataset IDs.

Brief description: *Non-Kuusamo samples*

This Finnish schizophrenia sample set has been collected from a population cohort using national registers. The entire sample collection consists of 2756 individuals from 458 families of whom 931 are diagnosed with schizophrenia spectrum disorder. Families outside Kuusamo (n=288) all had at least two affected siblings. All diagnoses are based on DSM-IV and for a large fraction of cases there is cognitive data.

Conditions: Only risk-increasing or protective factors that may be associated with severe mental disorders may be studied using this data set. "Severe" refers to functional limitations caused by the disorder, not to any specific diagnostic group within mental disorders.

Data can be used as controls: No.

Acknowledgement: National Institute for Health and Welfare (THL) Finnish Schizophrenia Families from the "*The genetic etiology of severe mental disorders in Finland*" study.

UK10K_NEURO_FSZ

EGA Study ID: EGAS00001000118. Please refer to the EGA for this study's Dataset IDs.

Brief description: *Kuusamo samples*

These Finnish schizophrenia samples have been collected from a population cohort using national registers. The entire sample collection consists of 2756 individuals from 458 families of whom 931 are diagnosed with schizophrenia spectrum disorder, each family having at least two affected siblings. 170 families originate from an internal isolate (Kuusamo) with a three-fold lifetime risk for the trait. The genealogy of the internal isolate is well documented and the individuals form a "megapedigree" reaching to the 17th Century. All diagnoses are based on DSM-IV and for a large fraction of cases there is cognitive data.

Conditions:	Only risk-increasing or protective factors that may be associated with severe mental disorders may be studied using this data set. “Severe” refers to functional limitations caused by the disorder, not to any specific diagnostic group within mental disorders.
Data can be used as controls:	No.
Acknowledgement:	National Institute for Health and Welfare (THL) Finnish Schizophrenia Families from the “*The genetic etiology of severe mental disorders in Finland*” study.

UK10K_NEURO_ASD_FI

EGA Study ID: EGAS00001000110. Please refer to the EGA for this study’s Dataset IDs.

Brief description: These samples are a subset of a nationwide collection of Finnish autism spectrum disorder (ASD) samples. The samples have been collected from Central Hospitals across Finland in collaboration with the University of Helsinki. The samples consist of 93 individuals with a diagnosis of autistic disorder or Asperger syndrome from 36 families with at least two affected individuals. Of these individuals, 16 can be genealogically connected to form two large pedigrees originating from Central Finland, suggesting possible genetic risk factors shared identical by descent within the pedigrees. All diagnoses are based on ICD-10 and DSM-IV diagnostic criteria for ASDs. Additional phenotypic data is available for a subset of the individuals.

Conditions:	These Finnish autism samples can only be used for research of autism spectrum disorders and no other studies.
Data can be used as controls:	No.
Acknowledgement:	Finnish Autism Families

UK10K_NEURO_IOP_COLLIER

EGA Study ID: EGAS00001000121. Please refer to the EGA for this study’s Dataset IDs.

Brief description:

The Genetics and Psychosis (GAP) set consists of samples from subjects with schizophrenia, ascertained as a new-onset sample. This set is of UK origin, with data on

cognition, brain imaging and other endophenotypes.

The Maudsley twin series consists of probands ascertained from the Maudsley Twin Register, defined as patients of multiple birth who had suffered psychotic symptoms. This set is of UK origin, with data on cognition, brain imaging and other endophenotypes, with DNA available from an MZ or DZ affected or unaffected co-twin.

The Maudsley family study (MFS) consists of over 250 families who have a history of schizophrenia or bipolar disorder Within the Maudsley Family Study, biological markers of psychosis include neuropsychological tests, Evoked Response Potentials Tests (ERPs), MRI scans, dermatoglyphics and eye tracking. Early risk factors for psychosis and clinical symptoms are also investigated. This set is of UK origin, with DNA available from both affected and unaffected relatives in many of the probands.

Conditions: No additional constraints. “Users" will be able to access unlinked, anonymised data with basic phenotypes (primary diagnostic information) and genome sequence data, in accordance with the UK10K data access agreement.

Data can be used as controls: Yes.

Acknowledgement:

Maudsley family study

Distribution of symptom dimensions across Kraepelinian divisions. Dikeos DG, Wickham H, McDonald C, Walshe M, Sigmundsson T, Bramon E, Grech A, Toulopoulou T, Murray R, Sham PC. Br J Psychiatry. 2006 Oct;189:346-53.

GAP study

High-potency cannabis and the risk of psychosis. Di Forti M, Morgan C, Dazzan P, Pariante C, Mondelli V, Marques TR, Handley R, Luzi S, Russo M, Paparelli A, Butt A, Stilo SA, Wiffen B, Powell J, Murray RM. Br J Psychiatry. 2009 Dec;195(6):488-91.

The Maudsley Twin Study.

Genetic overlap between episodic memory deficits and schizophrenia: results from The Maudsley Twin Study. Owens SF, Picchioni MM, Rijsdijk FV, Stahl D, Vassos E, Rodger AK, Collier DA, Murray RM, Toulopoulou T. Psychol Med. 2011 Mar;41(3):521-32

UK10K_NEURO_UKSCZ

EGA Study ID: EGAS00001000123. Please refer to the EGA for this study’s Dataset IDs.

Brief description: These samples have been collected from throughout

the UK and Ireland. The samples fall into two main categories, approximately 500 have a full diagnostic work up. A proportion of these are cases with a positive family history of schizophrenia, either collected as sib-pairs or from multiplex kindred's. The second group consist mainly of >300 samples that have been systematically collected within South Wales and in addition to full diagnostic work up have undergone detailed cognitive testing. All samples have obtained a DSM IV diagnosis of schizophrenia or schizoaffective disorder.

Conditions: No additional constraints, provided that the only data accessible will be sequence data, month/year of birth, diagnosis and family history.

Data can be used as controls: No.

Acknowledgement: CardiffScz

UK10K_NEURO_IMGSAC

EGA Study ID: EGAS00001000120. Please refer to the EGA for this study's Dataset IDs.

Brief description: The IMGSAC cohort is an international collection of families containing children ascertained for ASDs (autism spectrum disorders). The affected individuals are have been phenotyped, including using the ADI-R and ADOS instruments. Individuals with a past or current medical disorder of probable etiological significance or TSC have been excluded. Where possible, karyotyping has been performed on one affected individual per family to exclude Fragile X syndrome. Many of the samples have been genotyped, using the Affymetrix 10k and Illumina 1M platforms. All samples to be included in the current study are of UK origin.

Conditions: Yes – the use of the data must be restricted to the identification of susceptibility alleles for autism and related disorders (i.e. the broader autism phenotype).

Data can be used as controls: No.

Acknowledgement: "The International Molecular Genetic Study of Autism Consortium (IMGSAC)". A full list of consortium members can be provided as required.

UK10K_NEURO_ABERDEEN

EGA Study ID: EGAS00001000109. Please refer to the EGA for this study's Dataset IDs.

Brief description: This sample set comprises cases of schizophrenia with additional cognitive measurements, collected in Aberdeen, Scotland.

Conditions: No additional constraints, if anonymised.

Data can be used as controls: Yes, if anonymised.

Acknowledgement: Scottish schizophrenia cases.

UK10K_NEURO_ASD_GALLAGHER

EGA Study ID: EGAS00001000112. Please refer to the EGA for this study's Dataset IDs.

Brief description: This is an Irish sample set of individuals with ASD (approximately 50% with comorbid intellectual disability). Individuals have been diagnosed with ADI/ ADOS, measures of cognition/ adaptive function. They represent a more severe, narrowly defined cohort of ASD subjects. Family histories are available for some with measures of broader phenotype.

Conditions: No additional constraints.

Data can be used as controls: Yes.

Acknowledgement: Trinity College Dublin Autism Genetics Collection.

UK10K_NEURO_GURLING

EGA Study ID: EGAS00001000225. Please refer to the EGA for this study's Dataset IDs.

Brief description: This sample set consists of DNA from multiply affected schizophrenia families. The families have been diagnosed using the SADS-L clinical instrument which gives diagnoses at the probable level of the research diagnostic criteria (RDC). In addition all diagnoses are available using DSMIIIR criteria. These criteria are widely accepted as being valid and reliable for the diagnosis of schizophrenia. All families have been collected to ensure that they are uni-lineal for transmission of schizophrenia, i.e. they have only one affected parent with schizophrenia, or a relative of only one transmitting or obligate carrier parent with schizophrenia. Families with bi-lineal transmission of schizophrenia (i.e. with both parents being affected) were not sampled for this study. All families have multiple cases of schizophrenia and related disorders. The families have been selected to ensure there are no cases of bipolar disorder within them and that they do not contain bipolar disorder in any relatives on either side of the family.

Conditions: No additional constraints.

Data can be used as controls: Yes.

Acknowledgement: University College London Schizophrenia Family Samples.

UK10K_OBESITY_SCOOP

EGA Study ID: EGAS00001000124. Please refer to the EGA for this study's Dataset IDs.

Brief description: The Severe Childhood Onset Obesity Project (SCOOP) is a sub-cohort of the Genetics Of Obesity Study (GOOS) cohort, established by Sadaf Farooqi and Steve O'Rahilly at the University of Cambridge. The GOOS cohort contains >4000 patients of diverse geographic origin, many of whom have monogenic and syndromic forms of obesity, and includes patients that are offspring of consanguineous union. SCOOP is a subset of >1500 UK Caucasian patients with severe, early onset obesity (all patients have a BMI Standard Deviation Score (SDS) > 3 and obesity onset before the age of 10 years), in whom known monogenic causes of obesity have been excluded.

Conditions: "Users" can only use this data to look at genetic variation associated with obesity. In accordance with ethical committee approval, data cannot be analysed in relation to ancestry or other genetic diseases.

Data can be used as controls: No.

Acknowledgement: The Severe Childhood Onset Obesity Project (SCOOP) UK includes patients with severe early onset obesity that were originally recruited to the Genetics of Obesity Study (GOOS).

UK10K _OBESITY_GS

Please refer to the EGA for Study and Dataset IDs.

Brief Description: The *Generation Scotland: Scottish Family Health Study* (GS:SFHS) is a family-based genetic study with more than 24,000 volunteers across Scotland, consisting of DNA, clinical and socio-demographic data. This sample set consists of individuals from informative families with extreme obese subjects, including trios of extreme obese subjects with non-obese patents and multiple obese subjects within the same family.

Conditions: Users can only use GS data for biomedical research.

Data can be used as controls: Yes.

Acknowledgement:	Generation Scotland:Scottish Family Health Study (GS:SFHS).

UK10K_OBESITY_TWINSUK

Please refer to the EGA for Study and Dataset IDs.

Brief description:	This sample set consists of individuals from the TwinsUK study with a BMI >40.
Conditions:	No additional constraints.
Data can be used as controls:	Yes.
Acknowledgement:	The TwinsUK Cohort

UK10K_RARE_SIR

EGA Study ID: EGAS00001000130. Please refer to the EGA for this study's Dataset IDs.

Brief description:	The Severe Insulin Resistance (SIR) sample set will form part of the "rare disease" group, and will undergo exome sequencing.
Conditions:	No constraints.
Data can be used as controls:	Yes.
Acknowledgement: Cohort.	The Cambridge Severe Insulin Resistance Study

UK10K_RARE_NEUROMUSCULAR

EGA Study ID: EGAS00001000101. Please refer to the EGA for this study's Dataset IDs.

Brief description:	The samples for genetic neuromuscular diseases will be part of the "rare disease" group, and will undergo exome sequencing. They fall into the following groups: 1. Congenital muscular dystrophies and congenital myopathies.

2. Neurogenic conditions.
3. Mitochondrial disorders.
4. Periodic paralysis.

Conditions: No additional constraints.

Data can be used as controls: Yes.

Acknowledgement: The Molecular Genetics of Neuromuscular Disorders Study

UK10K_RARE_COLOBOMA

EGA Study ID: EGAS00001000127. Please refer to the EGA for this study's Dataset IDs.

Brief description: Ocular coloboma is the most common significant developmental eye defect with an incidence of ~1 in 5,000 live births. It results from failure of optic fissure closure during embryogenesis. The position and extent of the fusion failure dictates the clinical appearance and functional effect. ~30% of coloboma cases are associated with other systemic malformations. These UK10K samples will mostly comprise isolated coloboma cases without systemic involvement (aka non-syndromal coloboma). There is strong evidence from family studies that coloboma has a major genetic component with autosomal dominance being the most common pattern of inheritance. However, many cases are isolated or show complex patterns of familial clustering. The genes responsible for isolated coloboma are largely unknown, but in a small number of families mutations in SHH, CHX10, and PAX6 have been identified indicating marked genetic heterogeneity. Thus in addition to the clinical benefits of achieving a molecular diagnosis there are also major scientific advantages to identifying coloboma genes, as these are likely to provide insights into the complex process of optic fissure closure, that is critical to normal eye development. In the longer term, understanding the molecular basis of the disease may provide clues to therapeutic strategies.

Conditions: "Users" are constrained to use the data to understand the genetic basis of eye malformations.

Data can be used as controls: No.

Acknowledgement: MRC HGU Coloboma Study Cohort.

UK10K_RARE_CHD

EGA Study ID: EGAS00001000125. Please refer to the EGA for this study's Dataset IDs.

Brief description:	The Congenital Heart Disease data set will be part of the "rare disease" group, and undergo exome sequencing.
Conditions:	The data must be used for CHD related research only.
Data can be used as controls:	No.
Acknowledgement:	Genetic Origins of Congenital Heart Disease Study (GOCHD Study).

UK10K_RARE_CILIOPATHIES

EGA Study ID: EGAS00001000126. Please refer to the EGA for this study's Dataset IDs.

Brief description: The ciliopathies are an emerging group of disorders that arise from some dysfunction of cilia both motile or immotile forms. It is predicted that over 100 known conditions are likely to fall under this category, but only a handful have thus far been studied in any depth. Most individual ciliopathies are rare with just a small number of cases having been reported, thereby presenting researchers with often insurmountable difficulties for causative gene identification.

Conditions:	"Users" are constrained to use the data for research on Ciliopathies.
Data can be used as controls:	No.
Acknowledgement:	Cilia in Disease and Development study (CINDAD)

UK10K_RARE_FIND

EGA Study ID: EGAS00001000128. Please refer to the EGA for this study's Dataset IDs.

Brief description: Familial INtellectual Disability (FIND) is a cohort of

families with intellectual impairment. Affected members in families are at the extreme end of the spectrum with the majority having moderate to severe mental retardation where the recurrence risks suggests most are likely to have monogenic causes. A subset of the cohort have undergone detailed analysis of the X chromosome by Sanger sequence analysis of exomes and more recently by detailed high resolution aCGH of the X c'some. Samples from the first study where no causal variant could be identified have been selected for this whole exome study. The cohort comprises of largely non-syndromic and a few syndromic cases and the samples have been selected to be biased towards families with male sib-pairs to enrich for non-X linked disease genes.

Conditions:	Consent is given to identify the cause of intellectual disability and does not state what else it can be used for. It is clear the data will be stored anonymously in a central database.
Data can be used as controls:	No for general use. Yes if the focus of the research is the genetic cause of intellectual disability.
Acknowledgement:	The Familial Intellectual Disability study or FIND study.

UK10K_RARE_THYROID

EGA Study ID: EGAS00001000131. Please refer to the EGA for this study's Dataset IDs.

Brief description: Two cohorts of subjects are being analysed: Individuals with Congenital Hypothyroidism (CH) due either to dysgenesis or dyshormonogenesis; Patients with Resistance to Thyroid hormone (RTH), a disorder characterized by elevated thyroid hormones and variable tissue refractoriness to hormone action. The CH cohort has been enriched for genetic aetiologies by recruiting cases that are familial, on a consanguineous background or syndromic. The RTH cohort consists of cases in which candidate gene analyses have been negative.

Conditions:	No additional constraints. Participants have consented to the anonymised data being accessible to other researchers subject to the constraints specified by the Data Access Committee (e.g. that researchers will not try to identify them).
Data can be used as controls:	Yes.
Acknowledgement:	Disorders of thyroid hormone synthesis and action cohort from Krishna Chatterjee (Wellcome Trust and NIHR supported).

UK10K_RARE_HYPERCHOL

EGA Study ID: EGAS00001000129. Please refer to the EGA for this study's Dataset IDs.

Brief description: Familial Hypercholesterolemia is a condition where the affected person has a consistently high level of LDL, which can lead to early clogging of the coronary arteries. All patients selected for this study will have been found not to carry the common APOB and PCSK9 mutations, and to have no detectable LDLR mutations by testing for 18 common mutations and by molecular screening methods including SSCP and HRM, and by MLPA for gross deletions/insertions.

Conditions: No additional constraints.

Data can be used as controls: Yes.

Acknowledgement: Simon Broome Register Familial Hypercholesterolaemia samples

Appendix B: UK10K Project Publications Policy

The primary purpose of the UK10K Project is to investigate in a systematic genome-wide fashion the contribution of low frequency and rare genetic variants to medical traits, based on genome-scale sequencing of phenotyped samples. We will study multiple groups of related phenotypes so as to explore rare variants in different types of disease process, and add exomes from extreme sample sets to increase power and compare clinical extremes to population cohort designs. The sequence data set we generate will provide a genotype/phenotype resource an order of magnitude deeper than the genetic-only 1000 Genomes Project data set for Europe, that will empower future human genetic research in the UK and beyond.

A publication moratorium will protect the first publication rights of the sample custodians and data generators. The protected publication rights prohibit submission of papers by groups of authors not including the UK10K consortium that describe genetic variants or their use in association tests for the named phenotypes for which the samples were selected into the project, until one of the following criteria is met.

All data will no longer be subject to the publication moratorium once the data has been published, or until one year has passed since the full data set required for analysis was released. This one-year moratorium period will expire on:

02 July 2013 for the Cohorts datasets:
EGAD00001000194 (1713 UK10K_COHORT_TWINSUK samples)
EGAD00001000195 (740 UK10K_COHORT_ALSPAC samples)

and 02 January 2014 for all other exome datasets.

In all cases data will no longer be subject to a publication moratorium two years after its initial release.

For the sake of clarity, for the whole genome cohorts (TwinsUK and ALSPAC) the moratorium protects the first association publication using UK10K sequence data for any phenotype, and association tests for the phenotypes made available with the genetic data through the EGA (European Genome-Phenome Archive, http://www.ebi.ac.uk/ega).

Details of the publication moratorium will be made available along with the data sets deposited in the EGA. Potential data users are encouraged to contact the UK10K Project if there is any doubt about how the data may be used. Email templates will be provided to encourage this dialogue between potential data users and the UK10K project. This dialogue will also demonstrate to journal editors whether data has been used appropriately. Where a breach of the moratorium takes place or is suspected, in addition to leading to potential termination of current and future data access, the Data Access Committee and/or the UK10K Management Committee may contact the appropriate journal editor with evidence that data use conditions have been breached and request that any manuscripts be withdrawn.

UK 10K

UK10K Project Data Access Agreement Version 8 (22/05/13)

Principles for publications and authorships

The UK10K Consortium plans to publish a number of manuscripts which vary widely in scope both of content and contributing authors. All of these publications aim to conform to a core set of consortium principles:

- Recognition of the contribution of the many individuals and groups who will have enabled the Consortium's projects. These contributors include clinicians, academic disease genetics groups, sample processing and data production groups, analysis groups and cohort groups. Due recognition should also be given to students, postdoctoral and junior academics to assist their career progression.
- Safeguarding the scientific credibility of the consortium by exercising caution in claims based on consortium data and analysis.
- Avoiding undue delays in publication and dissemination of prominent findings, in the interests of maximizing benefits to human health, maintaining priority of publication (and protection of intellectual property), and in accordance with Wellcome Trust policy.

Categories of publications arising from the UK10K Project

The UK10K Consortium envisages three broad groups of publications to arise from Consortium data, each with distinct guidelines for recognizing the Consortium's contribution:

I. At least one flagship 'UK10K' paper will be written describing data and analyses across all consortium samples under banner authorship: 'UK10K' with all Consortium participants listed in the end matter by working group (see e.g. 1000 Genomes Project paper, and additional details below).
II. Additional papers coming from members of the Consortium using Consortium data on specific phenotypes, analytical methods, etc. will follow a set of guidelines on authorship subject to oversight of the publications committee (details below).
III. Papers based on external users downloading UK10K data, which will not be subject to this publication policy. Data users will be given information about pre-publication embargo times and instructions for acknowledging use of UK10K data. No UK10K authorships will be requested on these papers, though a publications "checklist" will be provided on receipt of data.

Detailed Authorship guidelines

1. The following groups will be represented by appropriate individuals in all papers with consortium authorship:
    a. Management Committee
    b. Sample logistics & production teams
2. The following groups will be represented in flagship papers and other Consortium papers in which they directly participated:

   a. Clinicians and researchers who have made an important contribution to sample and/or phenotype collection for samples used in the paper
   b. Individuals involved directly in the analysis presented in the paper, both within WTSI and elsewhere
   c. Individuals involved in generating replication data
3. Individuals working in one phenotype or aspect of the project will generally not be authors on papers focusing on other phenotypes or aspects.
4. Authorship arrangements will vary on non-flagship papers, subject to consultation between authors and the Publications Committee. These might range from “a handful of individuals (primary and senior authors) on behalf of the UK10K Consortium” to “a relatively lengthy list of authors with the UK10K Consortium included somewhere”. As with flagship papers, if there is a sufficiently long list of representatives of the UK10K Consortium not directly involved in the analysis presented in the paper they will be listed by group in the end matter.
5. Number of authors, ordering and shared positions should be aimed at recognizing both disease focused researchers and Consortium-wide researchers, as well as a balance between Sanger and other collaborators.
6. The list of named authors from the UK10K Consortium should be in the order in which they will appear in the PubMed index. They should be identified specifically as authors in the text of the manuscript to ensure that they are indexed as such in PubMed.

Operational procedures

1. Primary responsibility for writing additional Consortium papers will lie with the appropriate sub-group or working group in close collaboration with the central UK10K team. Analyses in these papers should either be in sub-group analysis plans or else recorded via the ‘Secondary Analysis’ procedure.
2. Papers should be emailed to the Publications Committee at least 1 week prior to submission for comments (largely on compliance with this policy). At the same time the manuscript should be emailed to all authors on the understanding that there will be a short turnaround time for comments as many of the papers will be under competitive pressure. A recommendation will be made by the Publications Committee to the Management Committee for ultimate sign-off on the manuscript and authorship list.
3. Abstracts for scientific meetings should be emailed to the Publications Committee at least 1 week before submission for comment. The Publications Committee will approve or deny abstract submission on the Management Committee’s behalf.
4. The Management Committee will disband this committee when they deem the main work of the consortium to be complete, and any future publications from consortium members will not be governed by this policy.

Authors who use data from the project must acknowledge the UK10K Consortium using the following wording, or close equivalent containing the same elements, and cite the relevant primary UK10K publication (details of which can be found on the UK10K website):

UK10K Project Data Access Agreement Version 8 (22/05/13)

"*This study makes use of data generated by the UK10K Consortium, derived from samples from <list separately for each sample set>. A full list of the investigators who contributed to the generation of the data is available from www.UK10K.org. Funding for UK10K was provided by the Wellcome Trust under award* ***WT091310*** "

Users should note that the UK10K Consortium bears no responsibility for the further analysis or interpretation of these data, over and above that published by the Consortium.