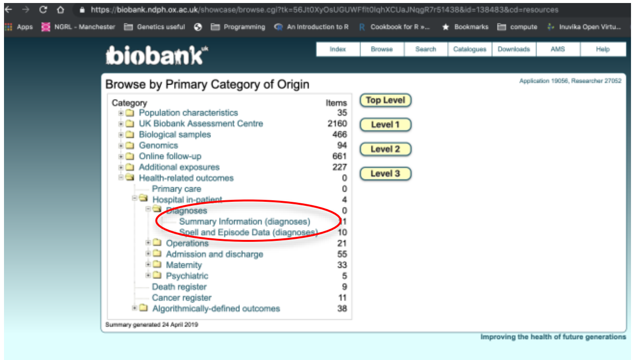
UK Biobank – UoM

Getting started with hospital episode statistics (HES) data

27/06/19

HES – what diagnoses are present?



https://biobank.ndph.ox.ac.uk/showcase/browse.cgi?tk=56Jt0XyOsUGUWFflt0lqhXCUaJNqgR7r51438&id=138483&cd=resources

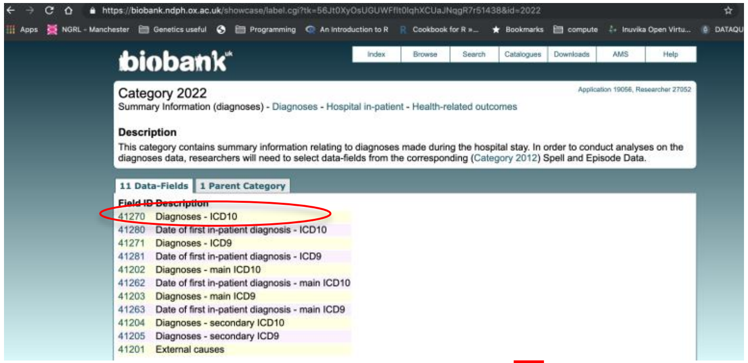
Browse by Primary Category of Origin

Application 19056, Researcher 27052

Category	Items
Population characteristics	35
UK Biobank Assessment Centre	2160
Biological samples	466
Genomics	94
Online follow-up	661
Additional exposures	227
Health-related outcomes	0
Primary care	0
Hospital in-patient	4
Diagnoses	0
Summary Information (diagnoses)	11
Spell and Episode Data (diagnoses)	10
Operations	21
Admission and discharge	55
Maternity	33
Psychiatric	5
Death register	9
Cancer register	11
Algorithmically-defined outcomes	38

Summary generated 24 April 2019

Improving the health of future generations



https://biobank.ndph.ox.ac.uk/showcase/label.cgi?tk=56Jt0XyOsUGUWFflt0lqhXCUaJNqgR7r51438&id=2022

Category 2022

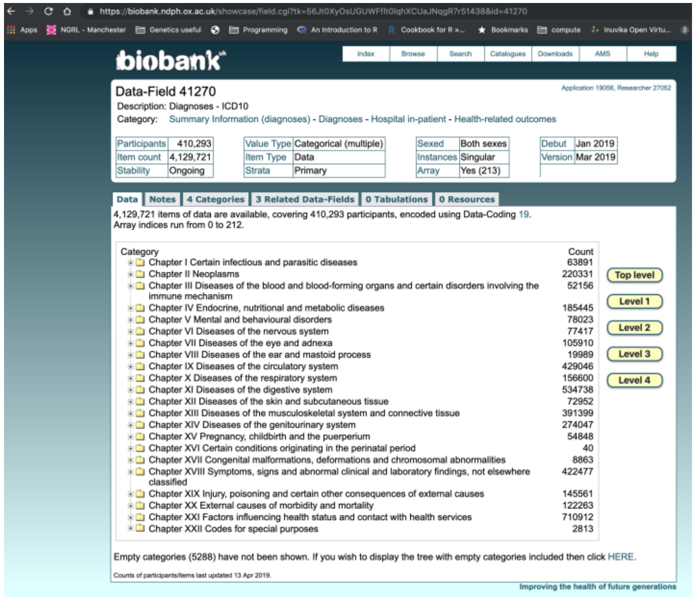
Application 19056, Researcher 27052

Summary Information (diagnoses) - Diagnoses - Hospital in-patient - Health-related outcomes

Description

This category contains summary information relating to diagnoses made during the hospital stay. In order to conduct analyses on the diagnoses data, researchers will need to select data-fields from the corresponding (Category 2012) Spell and Episode Data.

Field ID	Description
41270	Diagnoses - ICD10
41280	Date of first in-patient diagnosis - ICD10
41271	Diagnoses - ICD9
41281	Date of first in-patient diagnosis - ICD9
41202	Diagnoses - main ICD10
41262	Date of first in-patient diagnosis - main ICD10
41203	Diagnoses - main ICD9
41263	Date of first in-patient diagnosis - main ICD9
41204	Diagnoses - secondary ICD10
41205	Diagnoses - secondary ICD9
41201	External causes



https://biobank.ndph.ox.ac.uk/showcase/field.cgi?tk=56Jt0XyOsUGUWFflt0lqhXCUaJNqgR7r51438&id=41270

Data-Field 41270

Application 19056, Researcher 27052

Description: Diagnoses - ICD10

Category: Summary Information (diagnoses) - Diagnoses - Hospital in-patient - Health-related outcomes

Participants	410,293	Value Type	Categorical (multiple)	Sexed	Both sexes	Debut	Jan 2019
Item count	4,129,721	Item Type	Data	Instances	Singular	Version	Mar 2019
Stability	Ongoing	Strata	Primary	Array	Yes (213)		

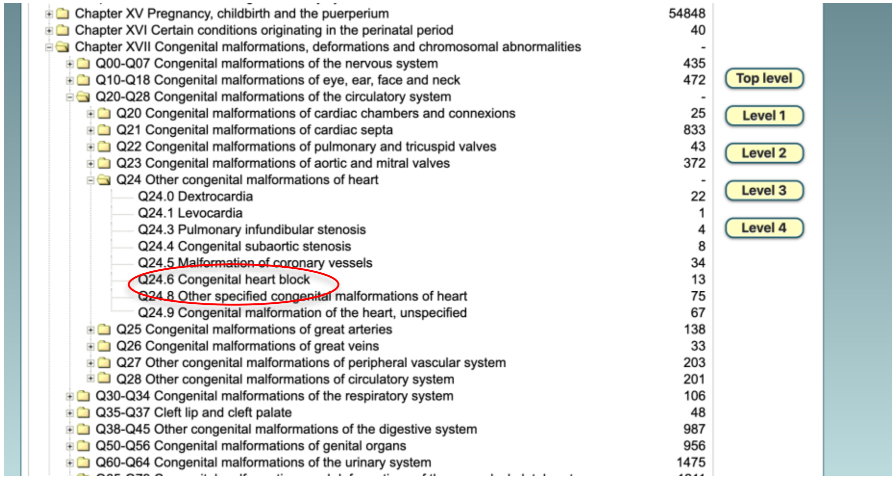
4,129,721 items of data are available, covering 410,293 participants, encoded using Data-Coding 19.
Array indices run from 0 to 212.

Category	Count
Chapter I Certain infectious and parasitic diseases	63891
Chapter II Neoplasms	220331
Chapter III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	52156
Chapter IV Endocrine, nutritional and metabolic diseases	185445
Chapter V Mental and behavioural disorders	78023
Chapter VI Diseases of the nervous system	77417
Chapter VII Diseases of the eye and adnexa	105910
Chapter VIII Diseases of the ear and mastoid process	19989
Chapter IX Diseases of the circulatory system	429046
Chapter X Diseases of the respiratory system	156600
Chapter XI Diseases of the digestive system	534738
Chapter XII Diseases of the skin and subcutaneous tissue	72952
Chapter XIII Diseases of the musculoskeletal system and connective tissue	391399
Chapter XIV Diseases of the genitourinary system	274047
Chapter XV Pregnancy, childbirth and the puerperium	54848
Chapter XVI Certain conditions originating in the perinatal period	40
Chapter XVII Congenital malformations, deformations and chromosomal abnormalities	8863
Chapter XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	422477
Chapter XIX Injury, poisoning and certain other consequences of external causes	145561
Chapter XX External causes of morbidity and mortality	122263
Chapter XXI Factors influencing health status and contact with health services	710912
Chapter XXII Codes for special purposes	2813

Empty categories (5288) have not been shown. If you wish to display the tree with empty categories included then click HERE.

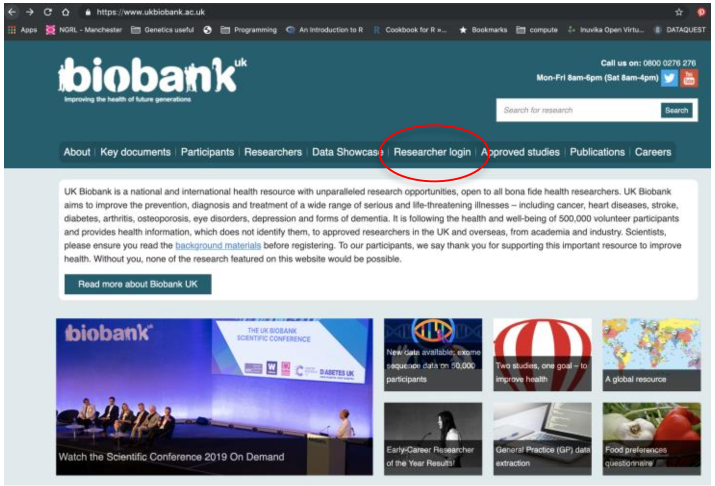
Counts of participants/items last updated 13 Apr 2019.

Improving the health of future generations



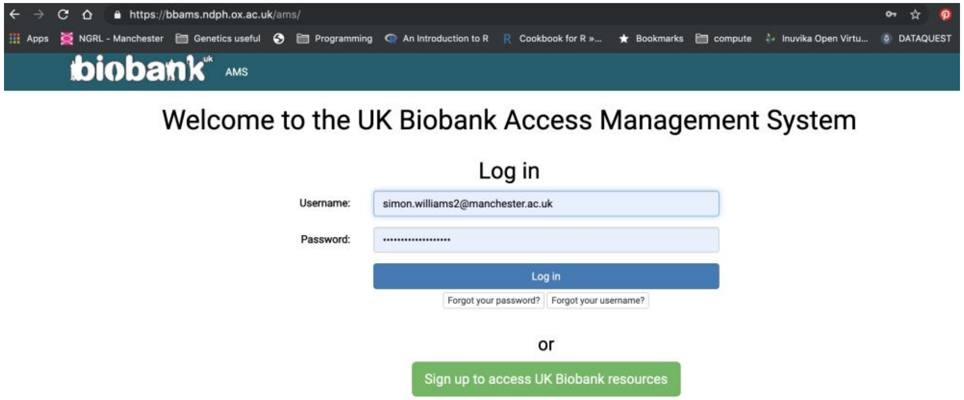
Category	Count
Chapter XV Pregnancy, childbirth and the puerperium	54848
Chapter XVI Certain conditions originating in the perinatal period	40
Chapter XVII Congenital malformations, deformations and chromosomal abnormalities	-
Q00-Q07 Congenital malformations of the nervous system	435
Q10-Q18 Congenital malformations of eye, ear, face and neck	472
Q20-Q28 Congenital malformations of the circulatory system	-
Q20 Congenital malformations of cardiac chambers and connexions	25
Q21 Congenital malformations of cardiac septa	833
Q22 Congenital malformations of pulmonary and tricuspid valves	43
Q23 Congenital malformations of aortic and mitral valves	372
Q24 Other congenital malformations of heart	-
Q24.0 Dextrocardia	22
Q24.1 Levocardia	1
Q24.3 Pulmonary infundibular stenosis	4
Q24.4 Congenital subaortic stenosis	8
Q24.5 Malformation of coronary vessels	34
Q24.6 Congenital heart block	13
Q24.8 Other specified congenital malformations of heart	75
Q24.9 Congenital malformation of the heart, unspecified	67
Q25 Congenital malformations of great arteries	138
Q26 Congenital malformations of great veins	33
Q27 Other congenital malformations of peripheral vascular system	203
Q28 Other congenital malformations of circulatory system	201
Q30-Q34 Congenital malformations of the respiratory system	106
Q35-Q37 Cleft lip and cleft palate	48
Q38-Q45 Other congenital malformations of the digestive system	987
Q50-Q56 Congenital malformations of genital organs	956
Q60-Q64 Congenital malformations of the urinary system	1475

HES – where to access the data once project has been approved



UK Biobank is a national and international health resource with unparalleled research opportunities, open to all bona fide health researchers. UK Biobank aims to improve the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses – including cancer, heart diseases, stroke, diabetes, arthritis, osteoporosis, eye disorders, depression and forms of dementia. It is following the health and well-being of 500,000 volunteer participants and provides health information, which does not identify them, to approved researchers in the UK and overseas, from academia and industry. Scientists, please ensure you read the background materials before registering. To our participants, we say thank you for supporting this important resource to improve health. Without you, none of the research featured on this website would be possible.

Read more about Biobank UK
THE UK BIOBANK SCIENTIFIC CONFERENCE
Watch the Scientific Conference 2019 On Demand
New data available: exome sequence data on 50,000 participants
Two studies, one goal – to improve health
A global resource
Early-Career Researcher of the Year Results!
General Practice (GP) data extraction
Food preferences questionnaire


https://bbams.ndph.ox.ac.uk/ams/
Apps
NGRL - Manchester
Genetics useful
Programming
An Introduction to R
Cookbook for R »...
Bookmarks
compute
Inuvika Open Virtu...
DATAQUEST
biobank uk AMS

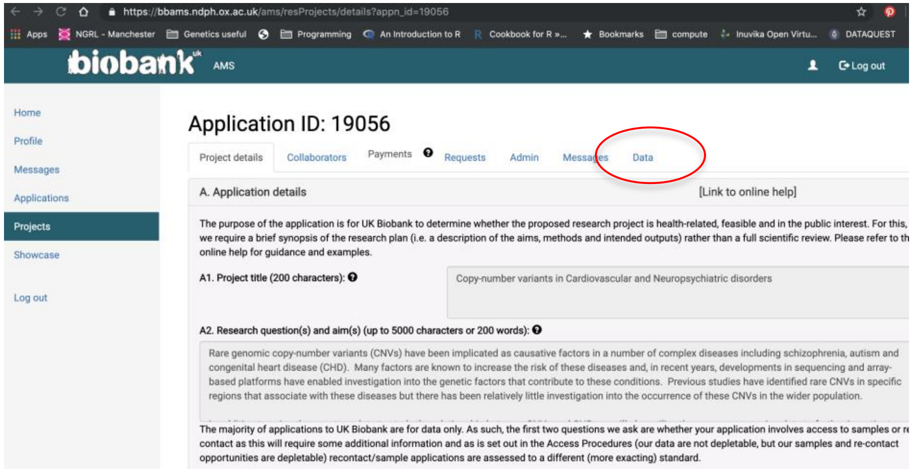
Welcome to the UK Biobank Access Management System

Log in

Username: simon.williams2@manchester.ac.uk

Password: ••••••••••••••••••••

Log in
Forgot your password?
Forgot your username?
or
Sign up to access UK Biobank resources



Application ID: 19056

A. Application details [Link to online help]

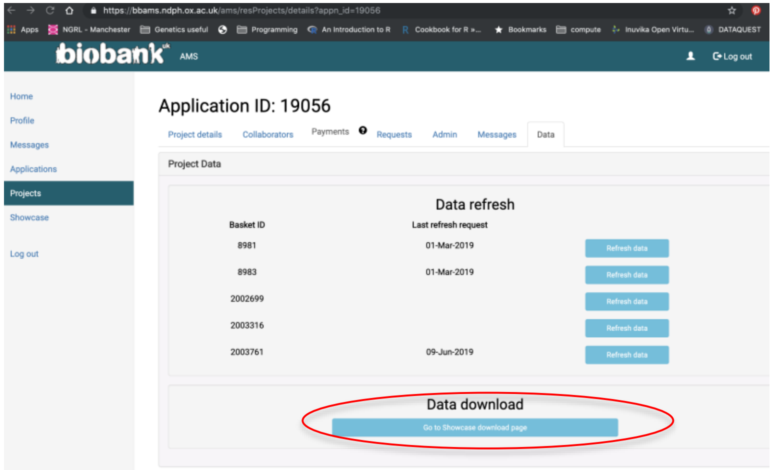
The purpose of the application is for UK Biobank to determine whether the proposed research project is health-related, feasible and in the public interest. For this, we require a brief synopsis of the research plan (i.e. a description of the aims, methods and intended outputs) rather than a full scientific review. Please refer to th online help for guidance and examples.

A1. Project title (200 characters): Copy-number variants in Cardiovascular and Neuropsychiatric disorders

A2. Research question(s) and aim(s) (up to 5000 characters or 200 words):

Rare genomic copy-number variants (CNVs) have been implicated as causative factors in a number of complex diseases including schizophrenia, autism and congenital heart disease (CHD). Many factors are known to increase the risk of these diseases and, in recent years, developments in sequencing and array-based platforms have enabled investigation into the genetic factors that contribute to these conditions. Previous studies have identified rare CNVs in specific regions that associate with these diseases but there has been relatively little investigation into the occurrence of these CNVs in the wider population.

The majority of applications to UK Biobank are for data only. As such, the first two questions we ask are whether your application involves access to samples or re contact as this will require some additional information and as is set out in the Access Procedures (our data are not depletable, but our samples and re-contact opportunities are depletable) recontact/sample applications are assessed to a different (more exacting) standard.



Application ID: 19056

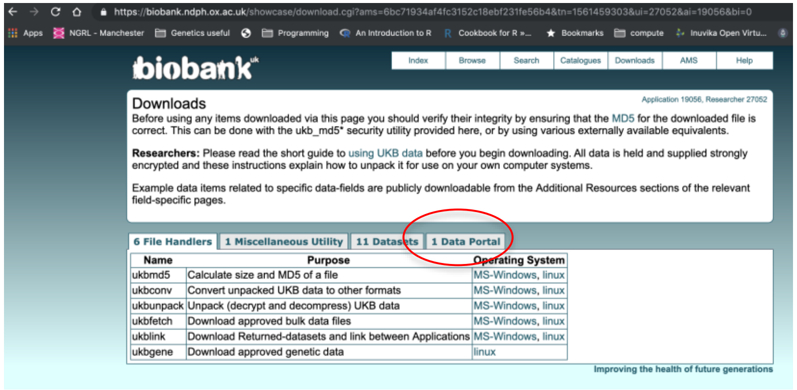
Project Data

Data refresh

Basket ID	Last refresh request	
8981	01-Mar-2019	Refresh data
8983	01-Mar-2019	Refresh data
2002699		Refresh data
2003316		Refresh data
2003761	09-Jun-2019	Refresh data

Data download

HES tables



Downloads

Application 19056, Researcher 27052

Before using any items downloaded via this page you should verify their integrity by ensuring that the MD5 for the downloaded file is correct. This can be done with the ukb_md5* security utility provided here, or by using various externally available equivalents.

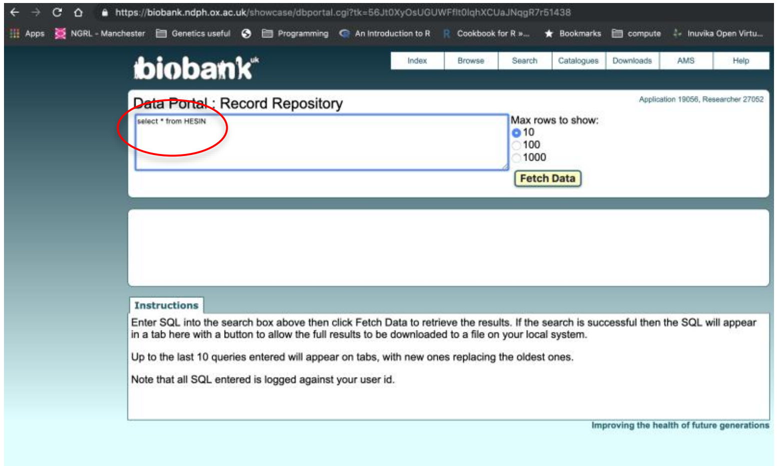
Researchers: Please read the short guide to using UKB data before you begin downloading. All data is held and supplied strongly encrypted and these instructions explain how to unpack it for use on your own computer systems.

Example data items related to specific data-fields are publicly downloadable from the Additional Resources sections of the relevant field-specific pages.

6 File Handlers | 1 Miscellaneous Utility | 11 Datasets | 1 Data Portal

Name	Purpose	Operating System
ukbmd5	Calculate size and MD5 of a file	MS-Windows, linux
ukbconv	Convert unpacked UKB data to other formats	MS-Windows, linux
ukbunpack	Unpack (decrypt and decompress) UKB data	MS-Windows, linux
ukbfetch	Download approved bulk data files	MS-Windows, linux
ukblink	Download Returned-datasets and link between Applications	MS-Windows, linux
ukbgene	Download approved genetic data	linux

Improving the health of future generations



Data Portal : Record Repository

Application 19056, Researcher 27052

```
select * from HESIN
```

Max rows to show:
● 10
○ 100
○ 1000

Instructions

Enter SQL into the search box above then click Fetch Data to retrieve the results. If the search is successful then the SQL will appear in a tab here with a button to allow the full results to be downloaded to a file on your local system.

Up to the last 10 queries entered will appear on tabs, with new ones replacing the oldest ones.

Note that all SQL entered is logged against your user id.

Improving the health of future generations

HESIN

- main primary hospital diagnoses
- episode/admission dates
- primary ICD9/ICD10/OPCS4 codes

HESIN_DIAG10

- secondary ICD10 diagnoses

HESIN_OPER

- secondary OPCS4 codes

HESIN_DIAG9

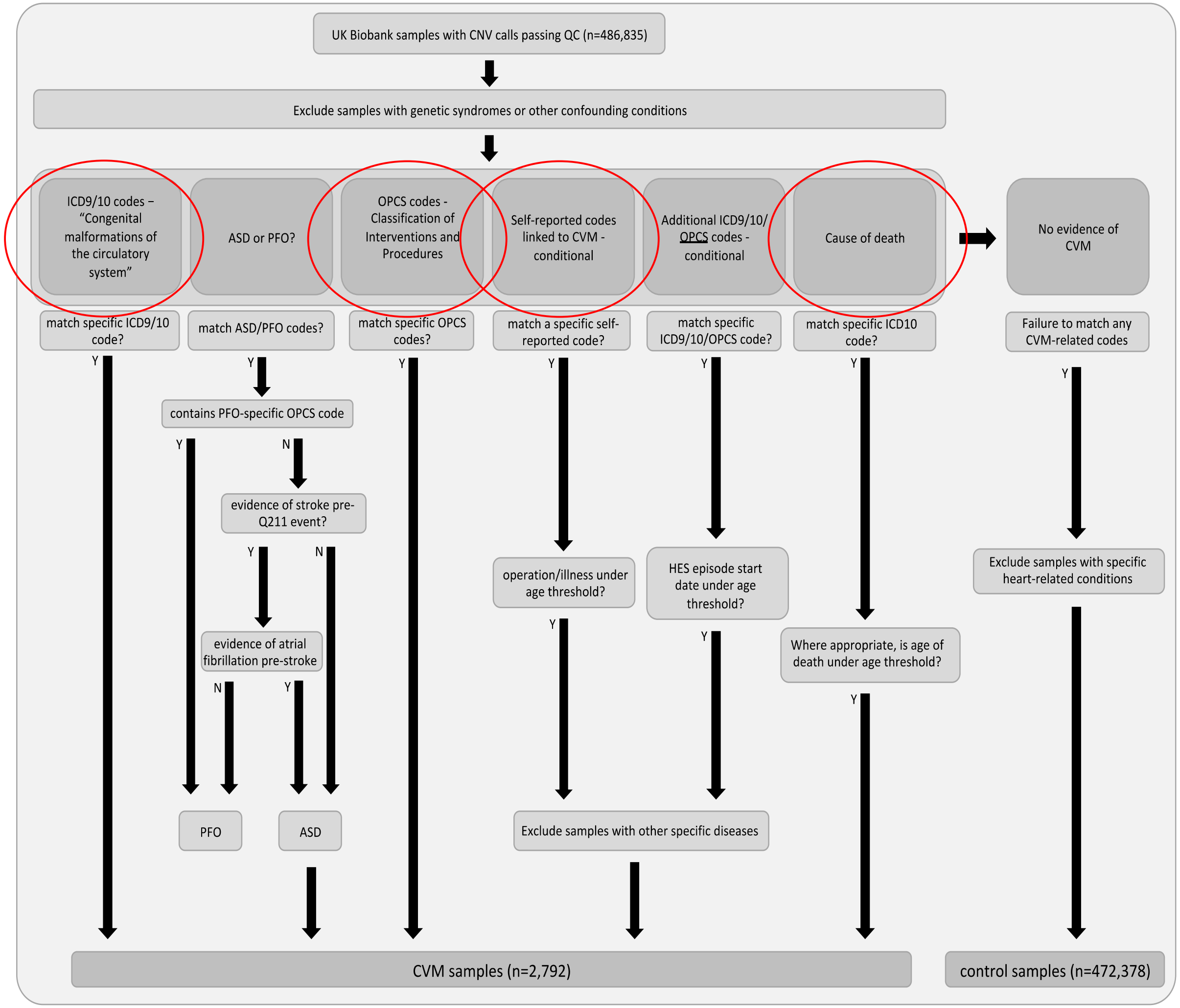
- secondary ICD9 codes

Data extracted by SQL queries:

- Query specific code e.g. SELECT eid FROM hesin WHERE diag_icd10 = 'Q256’
- Get everything and export e.g. SELECT * FROM hesin

Rscript to open HES tables and find samples with matching diagnosis – ‘Heart block (Q246)’

```
library(data.table)

###############################################################################
# load the HES data tables you’ve exported from UKB SQL page
###############################################################################

#read hesin table
hesin=fread("HESIN.tsv")

#read hesin_diag10 table (secondary diagnosis table)
hesin_diag10=fread("HESIN_DIAG10.tsv")

#read hesin_oper table (secondary operations table)
hesin_oper=fread("HESIN_OPER.tsv")


###############################################################################
# extract samples that might be of interest based on ICD10 codes
###############################################################################

#read list of ICD codes we're interested in – here jut a file with ‘Q246’ under the ‘ICD10’ header.
codes<-read.csv(”heart_block_phenotype_codes", header=TRUE)

#subset these samples from HESIN table
hit1<-subset(hesin, hesin$diag_icd10 %in% codes$ICD10)

#then HESIN_DIAG10 table - this table contains the secondary diagnosis (only ICD10)
hit2<-subset(hesin_diag10, hesin_diag10$diag_icd10 %in% codes$ICD10)

#you might also search for related ICD9/operation codes in the same way

#extract the eids
eid<-c(hit1$eid,hit2$eid)

#get unique eids
heart_block_samples<-unique(eid)
```


UK Biobank samples with CNV calls passing QC (n=486,835)
Exclude samples with genetic syndromes or other confounding conditions
ICD9/10 codes – "Congenital malformations of the circulatory system"
ASD or PFO?
OPCS codes - Classification of Interventions and Procedures
Self-reported codes linked to CVM - conditional
Additional ICD9/10/ OPCS codes - conditional
Cause of death
No evidence of CVM
match specific ICD9/10 code?
match ASD/PFO codes?
match specific OPCS codes?
match a specific self-reported code?
match specific ICD9/10/OPCS code?
match specific ICD10 code?
Failure to match any CVM-related codes
Y
Y
Y
Y
Y
Y
Y
contains PFO-specific OPCS code
Y
N
evidence of stroke pre-Q211 event?
Y
N
evidence of atrial fibrillation pre-stroke
N
Y
PFO
ASD
operation/illness under age threshold?
Y
HES episode start date under age threshold?
Y
Where appropriate, is age of death under age threshold?
Y
Exclude samples with specific heart-related conditions
Exclude samples with other specific diseases
CVM samples (n=2,792)
control samples (n=472,378)

Getting started genetic data

What is available?

Arrays

- 488,766 individuals
- 820,967 SNP and indel markers included
- 2 arrays
  - Affymetrix Axiom UK BiLEVE array (~50,000)
  - Affymetrix Axiom UK Biobank array (~450,000)

- Description of the files available:

http://www.ukbiobank.ac.uk/wp-content/uploads/2017/07/ukb_genetic_file_description.txt

Versions

- V1 – 1st release ~150,000 samples
- V2 – full release ~500,000 samples
- V3 – for imputation files only – due to an error in the initial V2 imputation release these files were re-processed and re-released

CSF3 central dataset

To access you must be added to a group with correct permissions (dataset-ukbiobank-full group)

Email the research IT team with confirmation of approved UKB data access

```
module load tools/env/ukbiobank-full-release-2018
```

Sets a number of environment variables:
UKBB_FILELIST
/mnt/data-sets/ukbiobank/full-release/filelist.2018.txt – list paths to files

More info:
http://ri.itservices.manchester.ac.uk/csf-apps/software/applications/ukbiobank/

Calls

- The genotype calls are in binary PLINK format (.bed, .bim, .fam) - see https://www.cog-genomics.org/plink/1.9/formats for details of the formats.

- The **BIM** file determines the order of markers in the calls and all of the other genotype data sets. The SNP_id is the rsid where it is available or the Affymetrix_SNP_id otherwise.

- The positions are **GRCh37** coordinates.

- The **FAM** file determines the order of samples in the calls and all of the other genotype data sets. The FAM file includes 'Batch' in the Phenotype field (6th column). - this file is project specific – the eids are different between projects

Imputed data

- The imputed genotype calls are in BGEN v1.2 format (.bgen, .sample, .bgi)

Files

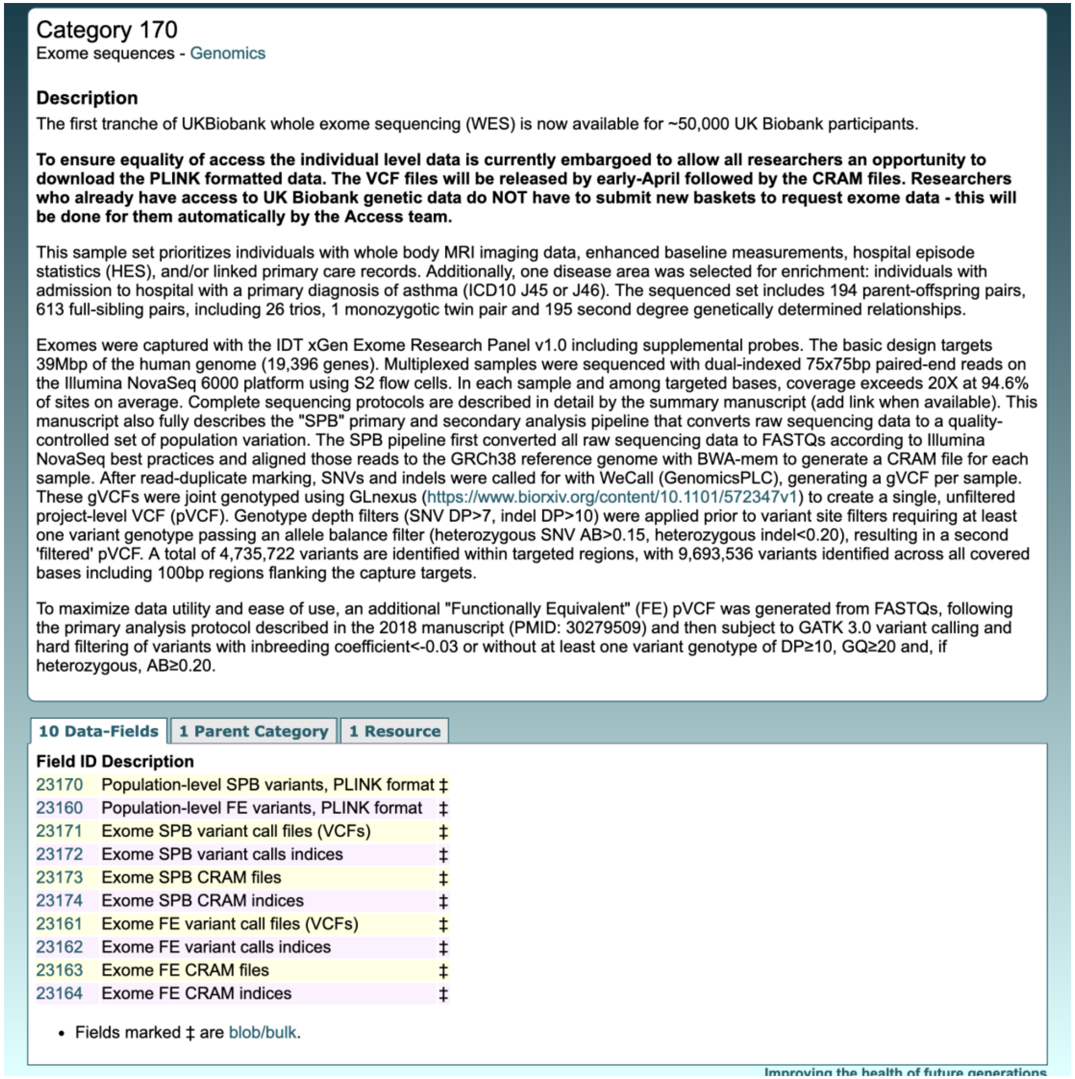
Calls BED	ukb_cal_chrN_v2.bed	
Calls BIM	ukb_snp_chrN_v2.bim	
Calls FAM	ukbA_cal_v2_sP.fam	→ Project-specific – download this yourself “ukbgene evc -c1 –m” from linux command line
Marker-QC	ukb_snp_qc.txt	
Sample-QC	ukb_sqc_v2.txt	
Relatedness	ukbA_rel_sP.txt	
Imputation BGEN	ukb_imp_chrN_v3.bgen	
Imputation BGI	ukb_bgi_chrN_v3.bgi	
Imputation MAF+info	ukb_mfi_chrN_v3.txt	
Imputation sample	ukbA_imp_autosome_v3_sP.sample	
Haplotypes BGEN	ukb_hap_chrN_v3.bgen	
Haplotypes BGI	ukb_hbg_chrN_v3.bgi	
HLA Imputation	ukb_hla_v2.txt	
Intensity	ukb_int_chrN_v2.bin	
Confidences	ukb_con_chrN_v2.txt	
CNV log2r	ukb_l2r_chrN_v2.txt	
CNV baf	ukb_baf_chrN_v2.txt	
SNP-posterior	ukb_snp_posterior_chrN.bin	
Batch	ukb_snp_posterior.batch	

Whole Exome Sequencing

- 1^{st} 50,000 released
- 39Mbp exome
- 75bp paired end reads
- Illumina NovaSeq 6000
- Mapped to GRCh38 reference

- Variant called through two pipelines:
  - 'FE' – 'Functional Equivalent' pipeline (GATK)
  - 'SPB' - Regeneron's Seal Point Balinese pipeline

- PLINK format release of all samples together
- Individual gVCFs can also be downloaded
- Download using 'ukbgene' utility



Category 170

Exome sequences - Genomics

Description

The first tranche of UKBiobank whole exome sequencing (WES) is now available for ~50,000 UK Biobank participants.

To ensure equality of access the individual level data is currently embargoed to allow all researchers an opportunity to download the PLINK formatted data. The VCF files will be released by early-April followed by the CRAM files. Researchers who already have access to UK Biobank genetic data do NOT have to submit new baskets to request exome data - this will be done for them automatically by the Access team.

This sample set prioritizes individuals with whole body MRI imaging data, enhanced baseline measurements, hospital episode statistics (HES), and/or linked primary care records. Additionally, one disease area was selected for enrichment: individuals with admission to hospital with a primary diagnosis of asthma (ICD10 J45 or J46). The sequenced set includes 194 parent-offspring pairs, 613 full-sibling pairs, including 26 trios, 1 monozygotic twin pair and 195 second degree genetically determined relationships.

Exomes were captured with the IDT xGen Exome Research Panel v1.0 including supplemental probes. The basic design targets 39Mbp of the human genome (19,396 genes). Multiplexed samples were sequenced with dual-indexed 75x75bp paired-end reads on the Illumina NovaSeq 6000 platform using S2 flow cells. In each sample and among targeted bases, coverage exceeds 20X at 94.6% of sites on average. Complete sequencing protocols are described in detail by the summary manuscript (add link when available). This manuscript also fully describes the "SPB" primary and secondary analysis pipeline that converts raw sequencing data to a quality-controlled set of population variation. The SPB pipeline first converted all raw sequencing data to FASTQs according to Illumina NovaSeq best practices and aligned those reads to the GRCh38 reference genome with BWA-mem to generate a CRAM file for each sample. After read-duplicate marking, SNVs and indels were called for with WeCall (GenomicsPLC), generating a gVCF per sample. These gVCFs were joint genotyped using GLnexus (https://www.biorxiv.org/content/10.1101/572347v1) to create a single, unfiltered project-level VCF (pVCF). Genotype depth filters (SNV DP>7, indel DP>10) were applied prior to variant site filters requiring at least one variant genotype passing an allele balance filter (heterozygous SNV AB>0.15, heterozygous indel<0.20), resulting in a second 'filtered' pVCF. A total of 4,735,722 variants are identified within targeted regions, with 9,693,536 variants identified across all covered bases including 100bp regions flanking the capture targets.

To maximize data utility and ease of use, an additional "Functionally Equivalent" (FE) pVCF was generated from FASTQs, following the primary analysis protocol described in the 2018 manuscript (PMID: 30279509) and then subject to GATK 3.0 variant calling and hard filtering of variants with inbreeding coefficient<-0.03 or without at least one variant genotype of DP≥10, GQ≥20 and, if heterozygous, AB≥0.20.

Field ID	Description	
23170	Population-level SPB variants, PLINK format	‡
23160	Population-level FE variants, PLINK format	‡
23171	Exome SPB variant call files (VCFs)	‡
23172	Exome SPB variant calls indices	‡
23173	Exome SPB CRAM files	‡
23174	Exome SPB CRAM indices	‡
23161	Exome FE variant call files (VCFs)	‡
23162	Exome FE variant calls indices	‡
23163	Exome FE CRAM files	‡
23164	Exome FE CRAM indices	‡

- Fields marked ‡ are blob/bulk.

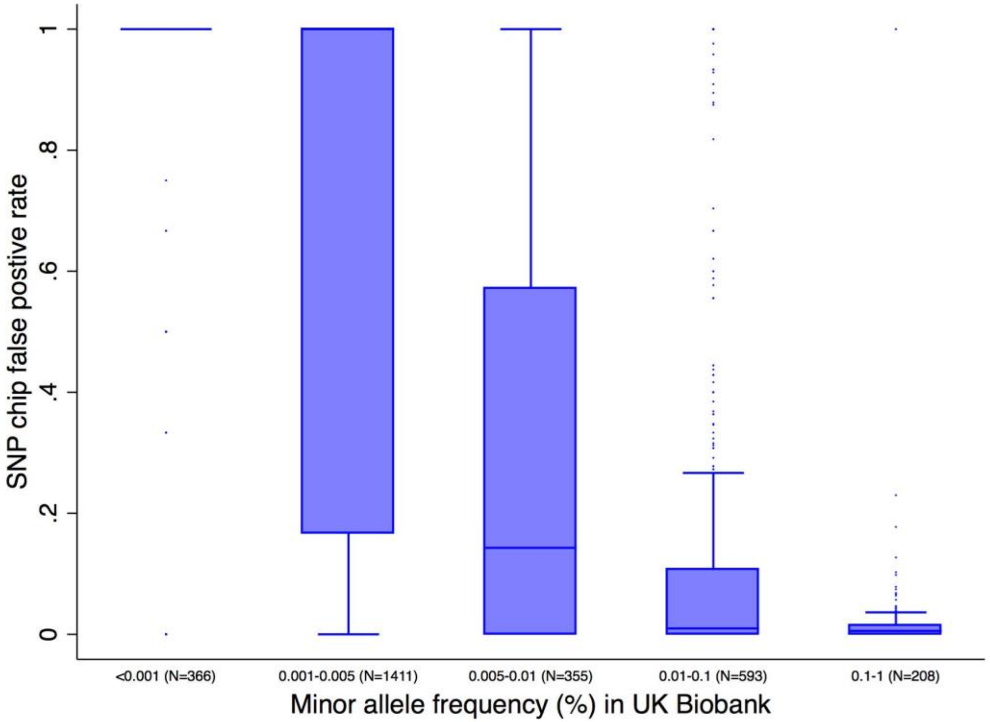
Variant call data for 50,000 exomes (sample level **VCF files**) ~10TB

CRAM files also available ~150TB

Capacity for downloading centrally…..

Initial comparison of arrays and exomes…

Mike Weedon (twitter) - analysis of 3000 QC'd exome SNPS reveals most SNPs on array with MAF <0.005% are FPs


SNP chip false postive rate
<0.001 (N=366)
0.001-0.005 (N=1411)
0.005-0.01 (N=355)
0.01-0.1 (N=593)
0.1-1 (N=208)
Minor allele frequency (%) in UK Biobank