

# UK Biobank

## Accessing UK Biobank Data

---

Version 2.3

<http://www.ukbiobank.ac.uk/>

October 2020



This document details the means by which data supplied by UK Biobank can be obtained and manipulated once access has been approved.

## Contents

0 What's new in this version .....	4
1 Introduction .....	5
1.1 The notification email .....	5
1.2 The formats of data available .....	5
1.3 Multiple downloads and refreshes .....	6
1.4 Getting help .....	7
2 Downloading a main dataset.....	8
2.1 Helper programs & encoding file .....	8
2.2 Downloading a main dataset from Showcase .....	10
2.3 Open a command prompt / terminal .....	11
2.4 Validating the download (ukbmd5) .....	12
2.5 Decrypting the dataset (ukbunpack).....	13
2.6 Conversion of the dataset (ukbconv).....	14
2.6.1 Creating a data dictionary .....	15
2.6.2 Converting to a csv or statistics package format.....	16
2.6.3 Optional parameters for ukbconv.....	18
2.7 Decryption and conversion example .....	19
2.8 The structure of a main dataset.....	21
3 Bulk data.....	24
3.1 Connectivity and authentication.....	24
3.2 Using ukbfetch.....	25
3.2.1 Downloading a single bulk item .....	25
3.2.2 Creating and using a bulk file.....	26
4 Genetics data.....	28
4.1 Genotype fields .....	28
4.2 Using gfetch .....	29

4.2.1 Connectivity & authentication.....	29
4.2.2 A gfetch example .....	30
4.3 Exome sequences.....	31
5 Record-level data.....	32
5.1 Record-level data on the Data Portal .....	32
5.2 Gaining access to the Data Portal.....	32
5.3 Downloading tables from the Data Portal .....	33
5.4 Using SQL to query the tables.....	33
6 Returned datasets.....	35
6.1 Authentication.....	35
6.2 Using ukblink.....	35
7 Bridges.....	37
7.1 Linking to Genetic data.....	37
7.2 Bridge files for bulk fields .....	37
8 Appendix.....	38
8.1 Using a command prompt in Windows .....	38
8.2 Issues with helper files & utilities .....	40
8.2.1 General.....	40
8.2.2 ukbunpack .....	41
8.2.3 ukbconv .....	41
8.2.4 ukbfetch .....	41
8.2.5 gfetch.....	42
8.2.6 ukbfetch / ukblink / gfetch .....	42
8.3 Sizes of bulk fields.....	44
8.4 File types of returned datasets .....	45

## 0 What's new in this version

This section summarises the changes in this document between version 2.2 and this version 2.3.

- Section 4 on accessing genomics data has been revised to describe the new gfetch utility that replaces the old ukbgene utility for downloading certain genomics data.

# 1 Introduction

## 1.1 The notification email

This guide is intended for researchers who have had an application for access to UK Biobank data, or a request for additional data on an existing project, approved and have received a notification email that their data is now available for download.

Note that only the Principal Investigator (PI) of a project and those collaborators with “delegate” status on the Application Management System (AMS) will receive the notification email and be able to download the main dataset and access the Data Portal on Showcase. The project PI can assign delegate rights to other collaborators in the Collaborators tab on the AMS.

The notification email will contain a 32-character MD5 Checksum within the main body of the email. This is needed to download the main dataset.

It will also have an attachment, with a name of the form: k56789r23456.key where 56789 will be replaced by the relevant application id and 23456 the run id, called an (authentication) keyfile containing a 64-character password. This password is needed to decrypt the main dataset, and the keyfile itself is needed to use other utilities to download bulk and genetics data as well as returned datasets.

## 1.2 The formats of data available

The data available to download from UK Biobank comes in a variety of formats which need to be accessed in different ways:

- The main dataset – this is downloaded, decrypted and converted according to the instructions given in section 2.
- Bulk images/files (e.g. MRI Images, ECG data) – these are downloaded using the ukbfetch utility as explained in section 3.
- Genetics data – this is downloaded using either the gfetch utility or ukbfetch depending on the type of data. Also, some genetics data can be downloaded from the European Genome-Phenome Archive (EGA), and some genetics fields are included as part of the main dataset rather than needing a separate download. See section 4 for further details.

- Record-level hospital and primary care (GP) data – this is accessed via the Data Portal in the Downloads page of Showcase. See section 5 for details.
- Returned datasets – these are datasets returned from researchers who have used UK Biobank data in their research, but which have not been incorporated directly into the main resource. See section 6 for details.

### 1.3 Multiple downloads and refreshes

The main dataset can be downloaded multiple times without limit, but will become inaccessible after a year. This is in order to prevent the data of participants who have subsequently withdrawn from the study being released again.

Periodically, the UK Biobank Showcase resource will be updated with new data. Currently, this typically happens 2-3 times a year. Researchers will be notified by email when a Showcase update has been made.

In order to gain access to updated data for fields in a previous data basket a researcher can request a “refresh” of that basket through AMS. A refresh of a dataset is a new extraction of the fields in the basket, and will include any additional data added to Showcase when it was updated. It will also remove the data for participants who have withdrawn since the basket was last released.

In order to request a refresh of a basket, a researcher will need to login to the Access Management System (AMS), navigate to their project (click Projects then View/Update), then click on the Data tab, and then on the “Go to Showcase to refresh or download data” button which will lead to the Downloads page. Next click ‘Application’ (at the top of the page) and then select the basket to be refreshed.

It will only be possible to refresh a basket that contains new data subsequent to a Showcase Update. If the selected basket can be refreshed a button ‘Request Refresh’ will be visible. Clicking on this button will then show the refresh requested as ‘Queued’. A new notification email will be sent when the refreshed basket is available to download.

If a Showcase update includes new fields that were not previously included in a basket for the project, then a “Change Request” can be submitted for access to the new fields.

Periodically UK Biobank will send out an email to researchers containing a list of all participants who have withdrawn consent for their data to be used. These participants should be removed from any unpublished analyses.

## 1.4 Getting help

If you are having difficulties with any aspect of the data download process we have collected some previously encountered issues in the Appendix of this document.

If you are unable to find a solution then please contact the Access Management Team (AMT) at [access@ukbiobank.ac.uk](mailto:access@ukbiobank.ac.uk) quoting your Application ID and the Run ID to which the problem relates.

It will help us to solve your issue more quickly if you provide screenshots of your problem, the steps you have followed up until the point the issue occurred, including any error messages received, as well as (where appropriate) listings of the contents of the folder you are working from.

If you find any errors in this document, or any parts that are unclear or incomplete, we would be grateful if you would pass them on to the AMT at the address above.

## 2 Downloading a main dataset

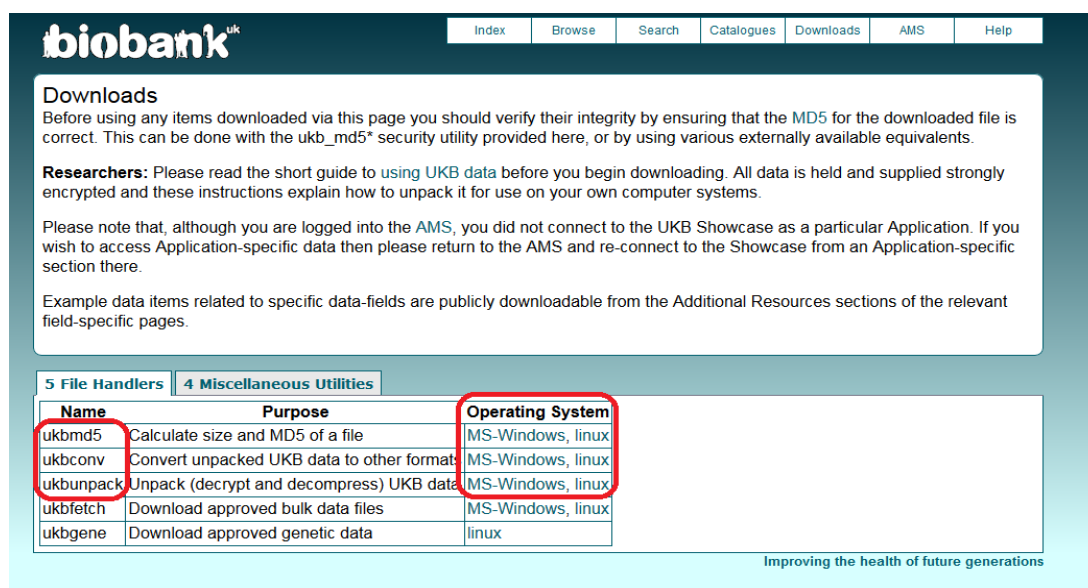
Downloading a main dataset requires several steps. The encrypted dataset must be downloaded through AMS. It must then be decrypted (“unpacked”), and then converted to a suitable format for use. A number of “helper programs” need to be downloaded to accomplish these steps.

### 2.1 Helper programs & encoding file

There are three helper programs required for decrypting and converting the main dataset:

- **ukbmd5** – for ensuring the encrypted main dataset has downloaded correctly;
- **ukbunpack** – for decrypting the downloaded main dataset;
- **ukbconv** – for converting the decrypted dataset into a suitable format.

These are provided in the File Handlers tab in the Download section of the Showcase website, as shown in figure 2.1.1.



The screenshot shows the UK Biobank website's 'Downloads' section. It includes a navigation bar with links like 'Index', 'Browse', 'Search', 'Catalogues', 'Downloads', 'AMS', and 'Help'. The main content area has a 'Downloads' heading and instructions for users. Below this, there are two tabs: '5 File Handlers' and '4 Miscellaneous Utilities'. The 'File Handlers' tab is active, displaying a table of helper programs. The table has three columns: 'Name', 'Purpose', and 'Operating System'. The programs listed are ukbmd5, ukbconv, ukbunpack, ukbfetch, and ukbgene. The 'Operating System' column lists 'MS-Windows, linux' for the first four and 'linux' for the last one. Red boxes highlight the 'Name' and 'Operating System' columns in the table.

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

**Figure 2.1.1: Helper programs**

The helper programs are supplied in two separate formats for compatibility with Windows or Linux operating systems. The Windows format is distinguished by the suffix ".exe".

The helper programs can be downloaded one at a time by selecting the required operating system version. This will open a new page, where the download can be found (figure 2.1.2).



Each program can be downloaded by clicking on it. We recommend that the helper programs are saved in a single file folder. A Linux command is also provided to perform the download.

**biobank<sup>uk</sup>** Index Browse Search Catalogues Downloads AMS Help

**Download 2**

Name ukbmd5

Description Calculate size and MD5 of a file

Platform/OS MS-Windows

Size 181248 bytes

MD5 6d7cf83d7bb8565f4a3ff6819d71e1b1

Instructions Run program from the command line with the name of the file as the only parameter. Running without a parameter will perform a self-test examination, the results of which should be compared against the values listed here.

Right-click or option-click filename and choose "Save As..." to download: ukbmd5.exe

If you have wget available (typically on linux systems), then you can also obtain the file using the command

```
wget -nd biobank.ndph.ox.ac.uk/showcase/util/ukbmd5.exe
```

Improving the health of future generations

Figure 2.1.2: Download page

As part of the conversion process into certain formats (section 2.6), the converter program “ukbconv” will look for a file called “encoding.ukb”, which is used to assign coded definitions to variables in the dataset, and is compatible for use with both Windows and Linux systems.

The file encoding.ukb is provided in the Miscellaneous Utility tab in the Download section of the Showcase website, as shown in figure 2.1.3. We recommend that you download “encoding.ukb” and save it along with the helper programs in the same folder.

**biobank<sup>uk</sup>** Index Browse Search Catalogues Downloads AMS Help

**Downloads**

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.

Please note that, although you are logged into the AMS, you did not connect to the UKB Showcase as a particular Application. If you wish to access Application-specific data then please return to the AMS and re-connect to the Showcase from an Application-specific section there.

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

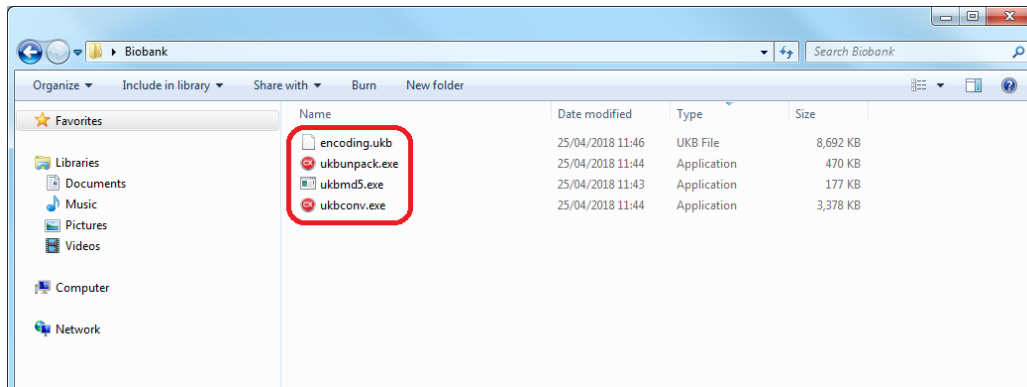
5 File Handlers 4 Miscellaneous Utilities

Name	Purpose	Operating System
encoding.ukb	Encoding dictionaries for use with ukb_conv	all
genotype_map	Mapping dictionary for use with gconv	all
ukb_field	List of available fields	all
ukb_category	List of available categories	all

Improving the health of future generations

Figure 2.1.3: Encoding file

At this point in the process you should have a folder containing four files similar to that shown in Figure 2.1.4.



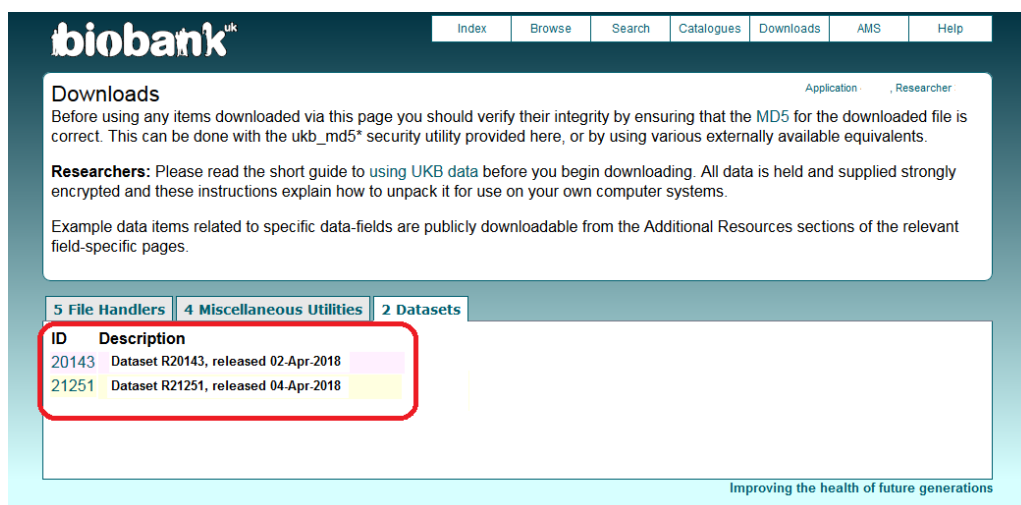
**Figure 2.1.4: Helper programs & encoding file**

## 2.2 Downloading a main dataset from Showcase

To download a dataset, you must first login to the Access Management System (AMS), navigate to the Projects tab and select the relevant application ID. Then click the blue button View/Update, then click on the Data tab at the top right, and then on the “Go to Showcase to refresh or download data” button which will lead to the Downloads page.

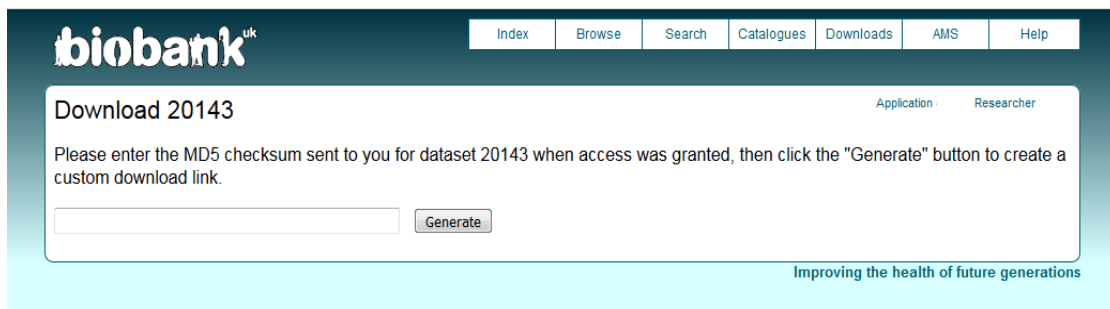
Only the project Principal Investigator (PI) and collaborators with delegate access are able to access the Data Download page. The project PI can assign delegate rights to other collaborators by using the Collaborators tab on the AMS.

Your dataset will be shown in the Dataset tab, as shown in figure 2.2.1:



**Figure 2.2.1: Location of datasets**

Click on the ID for the dataset you wish to download, which will take you to the authentication screen:

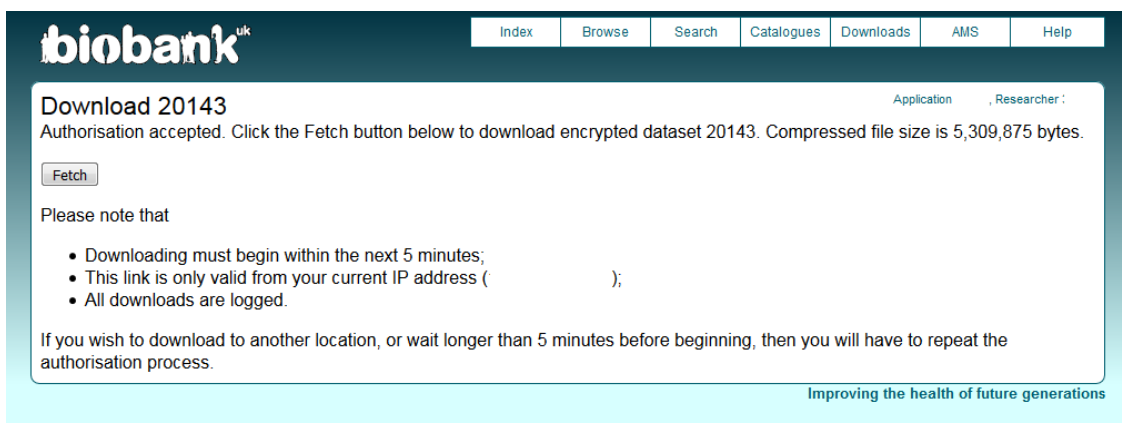


The screenshot shows the Biobank UK website with a navigation bar containing links: Index, Browse, Search, Catalogues, Downloads, AMS, and Help. The main content area is titled "Download 20143" and includes a sub-header "Application Researcher". Below this, a message states: "Please enter the MD5 checksum sent to you for dataset 20143 when access was granted, then click the 'Generate' button to create a custom download link." There is a text input field for the MD5 checksum and a "Generate" button. At the bottom right of the page, the slogan "Improving the health of future generations" is visible.

**Figure 2.2.2: Authentication screen**

Enter the 32-character MD5 checksum (included in the main body of the notification email for the dataset). Then click Generate.

This will open a new page with a link to your dataset as shown in Figure 2.2.3:



The screenshot shows the Biobank UK website with the same navigation bar. The main content area is titled "Download 20143" and includes a sub-header "Application Researcher". Below this, a message states: "Authorisation accepted. Click the Fetch button below to download encrypted dataset 20143. Compressed file size is 5,309,875 bytes." There is a "Fetch" button. Below the button, a section titled "Please note that" contains a bulleted list: "• Downloading must begin within the next 5 minutes;", "• This link is only valid from your current IP address ( );", and "• All downloads are logged." Below the list, a message states: "If you wish to download to another location, or wait longer than 5 minutes before beginning, then you will have to repeat the authorisation process." At the bottom right of the page, the slogan "Improving the health of future generations" is visible.

**Figure 2.2.3: Link to download dataset**

Click the Fetch button to download the encrypted dataset. Then save your dataset in the same file directory as the helper programs.

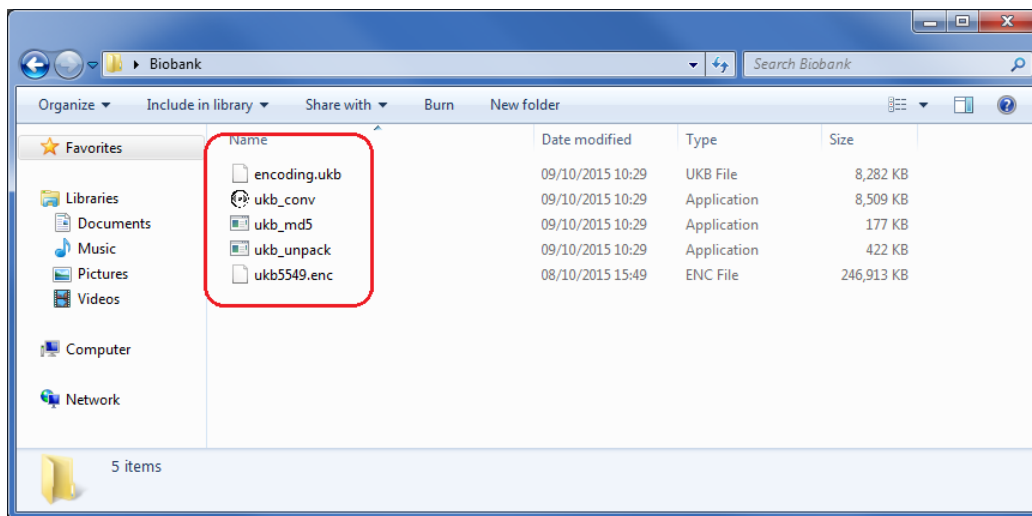
## 2.3 Open a command prompt / terminal

In order to proceed with the download process: validating, decrypting & converting the downloaded file, it is necessary to be able to run the helper programs (see section 2.1) using command line instructions from a command prompt in Windows or a terminal window in Linux.

For guidance on how to work with the command prompt in Window please see Section 8.1. The next few sections assume basic familiarity with command-line interfaces.

## 2.4 Validating the download (ukbmd5)

At this point you should now have five files in your folder, similar to as shown in figure 2.4.1. Note that the number used in the .enc file will be specific to your dataset; it should be the basket's run ID.



**Figure 2.4.1: The helper programs & dataset for run ID 5549**

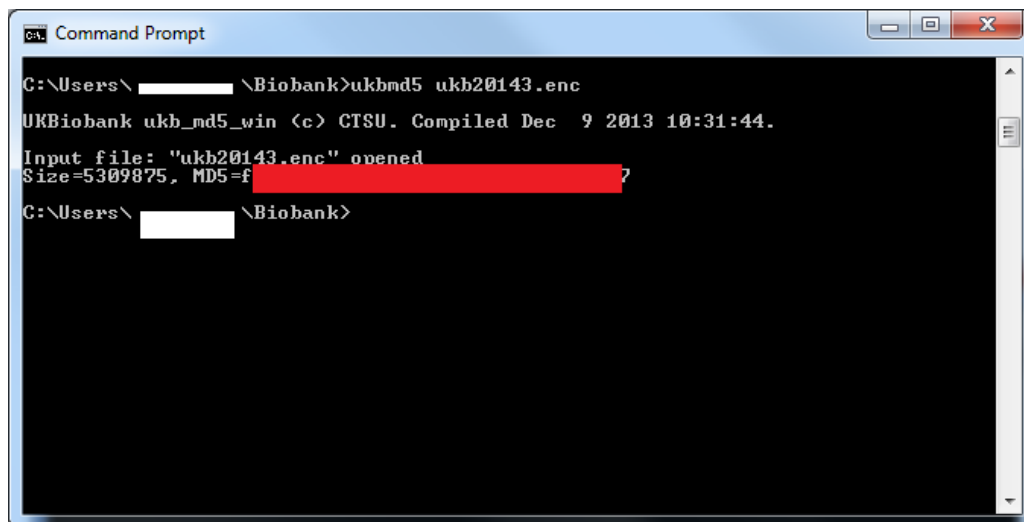
Open a command prompt / terminal and set the folder containing the above five files as your current working directory (by entering the command “cd” followed by the location; e.g. “cd username”).

You can verify the integrity of the files that you have downloaded by typing the command:

**ukbmd5 ukb23456.enc**

replacing **ukb23456.enc** with the name of your dataset file.

You should get output similar to Figure 2.4.2 below:



**Figure 2.4.2: Validation using ukbmd5**

where the red bar will be replaced by the 32-character MD5 checksum that you used to download the data. If the MD5 checksum is different to the one in your notification email it indicates that something has gone wrong in the download. In this case, you should delete the dataset and download it again.

## 2.5 Decrypting the dataset (ukbunpack)

Datasets are supplied in a compressed encrypted format. The ukbunpack program decrypts and uncompresses the downloaded file into a custom UK Biobank format.

To use the program, type the command:

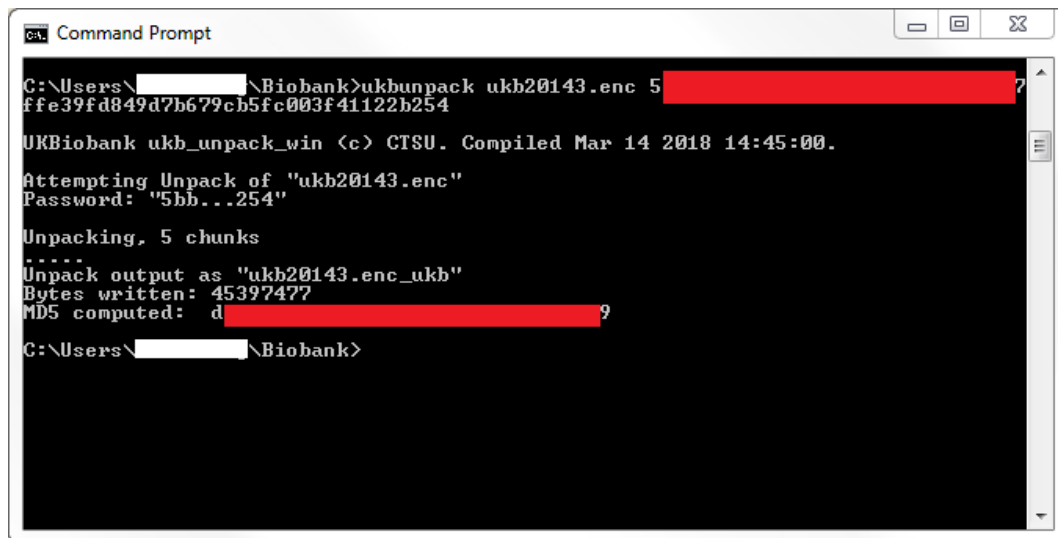
**ukbunpack ukb23456.enc keyValue**

replacing:

- **ukb23456.enc** with the name of your dataset file;
- **keyvalue** with the 64-character password from the second line of the keyfile attachment in your notification email (this will have a name of the form **k56789r23456.key** where 56789 is replaced by your application ID and 23456 by the run ID of the dataset). Note that the keyvalue is not the same as the MD5 checksum. The keyfile may not open directly from an email server (due to the .key extension) but as it is simply a text file it can be downloaded and then opened in a text editor such as Notepad.

**Each .enc file has a different keyvalue** that can be found as an attachment to the notification email for the release of that particular dataset. Although each keyfile is given the same name, the keyfiles are not interchangeable, and the passwords of datasets released for the same project will each be different.

After the command above has been entered, the file will be decrypted (“unpacked”) as shown in figure 2.5.1. This could take a few minutes.



```
C:\Users\ [redacted] \Biobank>ukbunpack ukb20143.enc 5 [redacted]
ffe39fd849d7b679cb5fc003f41122b254

UKBiobank ukb_unpack_win <c> CTSU. Compiled Mar 14 2018 14:45:00.

Attempting Unpack of "ukb20143.enc"
Password: "5bb...254"

Unpacking, 5 chunks
*****
Unpack output as "ukb20143.enc_ukb"
Bytes written: 45397477
MD5 computed: d [redacted]

C:\Users\ [redacted] \Biobank>
```

**Figure 2.5.1: Decrypting / unpacking a dataset**

This process will create a new file in your directory, named: **ukb23456.enc\_ukb**, where 23456 is replaced by the run ID of your dataset.

Note that an alternative way of using ukbunpack is using the command:

**ukbunpack ukb23456.enc k56789r23456.key**

where **ukb23456.enc** is as before and **k56789r23456.key** is the keyfile from the notification email, which must have been placed in the same folder as ukbunpack and your .enc file.

## 2.6 Conversion of the dataset (ukbconv)

The result of the unpacking program is a dataset in a custom UK Biobank format (the .enc\_ukb file shown above). The ukbconv program can be used to convert this into various other formats.

The ukbconv program is run via the command:

**ukbconv ukb23456.enc\_ukb option**

where **ukb23456.enc\_ukb** is the file generated from the previous unpacking step (with 23456 replaced by the run ID of your dataset) and **option** is replaced by one of: docs, csv, txt, r, sas, stata or bulk depending on the output desired.

The various options do the following:

- **docs**: generates a data dictionary for your dataset (see section 2.6.1);
- **csv, txt, r, sas** or **stata**: converts the dataset into a csv file, tab-separated txt file, or a file suitable for one of the statistics packages R, SAS and Stata (see section 2.6.2);
- **bulk**: creates a “bulk” file which is used in conjunction with ukbfetch to download bulk data (see section 3.2.2 for further details). This option is only relevant for the downloading of bulk data items such as MRI images etc.

In all cases the original .enc\_ukb file remains intact so the converter may be used multiple times to generate different outputs.

### 2.6.1 Creating a data dictionary

The option ‘docs’ creates an HTML document that lists information about the structure of the dataset. The first nine rows of such a file are shown below for illustration:

Column	UDI*	Count <sup>†</sup>	Type	Description
0	eid	502619	Sequence	Encoded anonymised participant ID
1	<a href="#">31-0.0</a>	502619	Categorical (single)	Sex <a href="#">Uses data-coding 9</a>
2	<a href="#">34-0.0</a>	502619	Integer	Year of birth
3	<a href="#">46-0.0</a>	499206	Integer	Hand grip strength (left)
4	<a href="#">46-1.0</a>	20202		
5	<a href="#">46-2.0</a>	23075		
6	<a href="#">47-0.0</a>	499273	Integer	Hand grip strength (right)
7	<a href="#">47-1.0</a>	20217		
8	<a href="#">47-2.0</a>	23070		

\* UDI - the Unique Data Identifier for an item of data within the UK Biobank repository. The format for standard data fields is field\_id-instance\_index.array\_index with genomic SNPs begin prefixed by "affy".

<sup>†</sup> Count - the number of non-empty rows present in this dataset.

See section 2.8 on the structure of a main dataset for an explanation of what is meant by “instance index” and “array index” for a main dataset.

Running the ‘docs’ option also creates a text file called fields.ukb giving a list of all the fields contained in the dataset; this file can be useful when using some of the other options for ukbconv as detailed in the next section. A log file is also created to summarise the results of the conversion process.

## 2.6.2 Converting to a csv or statistics package format

Using the option: csv, txt, r, sas or stata with ukbconv transforms this dataset into various standard formats.

To convert the dataset into a standard format we use the command:

**ukbconv ukb23456.enc\_ukb option**

where **ukb23456.enc\_ukb** is the file generated from the previous unpacking step (with 23456 replaced by the run ID of your dataset), and **option** is replaced by one of: csv, txt, r, sas or stata depending on the output desired.

Assuming the file encodings.ukb (see section 2.1) is contained in the folder where the conversion is taking place the options r, sas and stata will not only convert the data, but replace all categorical Data-Codings with their meanings.

All four options generate the following two files:

- fields.ukb – a simple text file, giving a list of all the Showcase field numbers appearing in the dataset
- ukb23456.log – a log file used to summarise the result of the conversion process, giving the date & time, name of the output file, application identifier, basket identifier, number of variables, and the time required to convert.

In addition, depending on the conversion type, a number of other files will be generated as shown in the table below:



<b>Format</b>	<b>File generated</b>	<b>Description</b>
<b>csv</b>	ukb23456.csv	Comma-separated file output with all fields double-quoted (to account for possible text fields containing commas).  The Data-Codings will be retained rather than replaced by their meanings.
<b>txt</b>	ukb23456.txt	A basic tab-separated text file output. As for csv above, the Data-Codings will be retained rather than replaced by their meanings.
<b>r</b>	ukb23456.tab	This is the actual file containing the data, in a tab-separated format. This file could potentially be imported directly into R, but none of the values will be coded.
	ukb23456.R	This file should be opened and executed in R (or any other R environment, such as RStudio). It contains a list of commands that will import the dataset (as a data.frame named bd) and recode all categorical variables.
<b>sas</b>	ukb23456.sd2	This is the actual file containing the data as a SAS Data Set. This file could potentially be imported directly into SAS, but none of the values will be coded.
	ukb23456.sas	This file is the SAS program that should be opened and executed. It contains a list of commands that will import the dataset (as a dataset named WORK.LABELLED_LFVPWW) and recode all categorical variables.
<b>stata</b>	ukb23456.raw	This is the actual file containing the data. This file could potentially be imported directly into Stata, but none of the values will be coded.
	ukb23456.do	This file should be opened and executed in Stata. It contains a list of commands that will import the dataset and recode all categorical variables.
	ukb23456.dct	A dictionary of values used by ukb23456.do to format and label variables in the imported dataset.

**Table 2.6.2: Conversion formats**

For example, with regards to the `encodings.ukb` file, if the `csv` option is used to convert the dataset then the field corresponding to “Sex” (field 31) will contain 0 and 1 (meaning Female and Male respectively). If the file is converted using the `r` option, then whilst the `.tab` file will still contain 0s and 1s in this column, if the `.R` file is opened and run, the dataset will be displayed with the column for field 31 showing “Female” and “Male”.

Note that if the file `encodings.ukb` is not present when the conversion into R, SAS or Stata format is run, then the conversion will still proceed but without the categorical variables being recoded.

Please note that large datasets may take a considerable amount of time (possibly hours) to convert, depending on the speed of the local system.

### 2.6.3 Optional parameters for `ukbconv`

Various optional parameters can be applied to the conversion, in particular to restrict which columns are included in the output. The full list is shown in the table below:

Flag	Meaning
-s	Specify a single field (only) to include in the output
-i	Specify a subset of fields to include in the output
-x	Specify a subset of fields to exclude from the output
-o	Specify an alternative name for the output file
-e	Specify an alternative file from which to extract encoding information

**Table 2.6.3: Optional parameters for `ukbconv`**

Options are included by adding them to the end of the `ukbconv` command. So for example the command:

```
ukbconv ukb23456.enc_ukb r -s20002
```

would convert the dataset into an R format, keeping only the `eid` column and all columns relating to field 20002. Note that since field 20002 (Non-cancer illness code, self-reported) has numerous different instance and array indices this will produce multiple additional columns (see section 2.8 for an explanation of instance and array indices).

When using the options `-i` or `-x` to select fields to include or exclude respectively from the converted dataset, the option should be immediately followed by the name of a text file which contains the list of fields with one field number per row. To assist with preparing

this file, the converter outputs the file named “field.ukb” each time it is run, and this lists all the available fields associated with the dataset. This can be edited to identify the particular fields which are to be included in or excluded from the subset.

Note that running the converter twice, using the same subset file but with -i and -x on alternate runs, will split the dataset into two complimentary parts.

By default ukbconv will look for the encoding file “encoding.ukb”, as described in the previous section, but by using the -e option a different filename can be used as the source for the Data-Coding definitions.

## 2.7 Decryption and conversion example

A researcher has been notified by email that data for their application 56789 is available for download. The email provides the run ID 23456 for the dataset. The 32-character MD5 Checksum is:

“abcdef0123456789abcdef0123456789”

and the 64-character Password (contained in the second line of the attached text file k56789r23456.key) is:

“a1b2c3d4a1b2c3d4a1b2c3d4a1b2c3d4a1b2c3d4a1b2c3d4a1b2c3d4a1b2c3d4”.

We assume that the three helper programs and encoding file have already been downloaded in accordance with section 2.1 and that the dataset is being downloaded into the same folder.

1. The researcher (either the PI or a collaborator with delegate access) logs on to the Application Management System (AMS), clicks Projects and then clicks on the blue button View/Update for project 56789. They select the Data tab at the top right of the page, and select the option to go to the Showcase download page. From this page they select the Dataset tab. An entry with run ID 23456 should be listed.
2. They click on the (run) ID 23456 for the entry and on the following screen enter the MD5 checksum given above (from the main body of the notification email):

abcdef0123456789abcdef0123456789

into the box and click Generate. This will open the download page; they click Fetch to initiate the download of file `ukb23456.enc` and save it in the same folder as the helper programs and encoding file.

3. To verify that the file has arrived intact they open a command prompt, navigate to the appropriate folder and enter:

```
ukbmd5 ukb23456.enc
```

This displays an MD5 value which matches the MD5 Checksum from the notification email (the one used to download the dataset). If the MD5 checksum had not matched, the researcher would need to repeat the download operation. If there was still no match they would need to contact the Access Management Team (AMT) for further assistance.

4. They next unpack (decrypt) the data by entering into the command prompt:

```
ukbunpack ukb23456.enc a1b2c3d4a1b2c3d4...a1b2c3d4
```

where we have truncated the 64-character keyfile so it fits onto the line above, but the researcher would have needed to include all 64 of the characters.

This will produce a file `ukb23456.enc_ukb`.

5. To create a comma separated variable (csv) version of the data, they enter into the command prompt:

```
ukbconv ukb23456.enc_ukb csv
```

This will produce a file `ukb23456.csv` which can be processed by standard programs.

## 2.8 The structure of a main dataset

Having followed the above steps a researcher will now have a main UK Biobank dataset. We here give some indication of what this would look like, focusing in particular on the meanings of the column headers.

A main dataset will be rectangular with one participant per row, and columns headers giving the Showcase field number that the data in the that column relates to together with the “instance index” and “array index” of that item. Broadly speaking, the instance index is used to distinguish data for a field which was gathered at different times, and the array index is used to distinguish multiple pieces of data for that field which were gathered at the same time.

These will display differently depending on the format that the dataset has been converted to (see table 2.8.2 at the end of this section). The example we give in table 2.8.1 below shows a small portion of a sample dataset as it would appear in .csv format opened in Excel:

eid	53-0.0	53-1.0	53-2.0	20002-0.0	20002-0.1	20002-1.0	20002-1.1	20002-2.0	20002-2.1	...
1256847	11/04/2007		03/01/2017	1077				1077	1075	
8645816	29/10/2009									
4652658	15/08/2009									
2328974	12/07/2008	09/03/2013				1002				
3315794	22/02/2010	01/12/2012	19/11/2018	1111		1111		1111	1065	
9497726	25/02/2006									
4582852	06/06/2008			1222	1265					
...										

**Table 2.8.1: A portion of a sample main dataset**

The eid is the encoded participant identifier for the project in question. The remaining column headers are in the format F-I.A where F is the field number, I is the instance index and A is the array index.

Two fields are shown in the sample dataset: [Field 53](#) (Date of attending assessment centre) and [Field 20002](#) (Non-cancer illness code, self-reported). In each case there are three “instances” of the variable (the first number after the -). Using the “Instance” tab on the fields pages on Showcase we see that these correspond the visit type: 0 for the initial (baseline) visit, 1 for the repeat assessment and 2 for the first imaging assessment.

The columns 53-0.0, 53-1.0 and 53-2.0 then hold the dates each participant attended that particular type of assessment centre. In the above, all participants attended a baseline assessment centre (this would always be the case), but only two attended the repeat assessment, one of whom also attended an imaging centre. The first participant attended an imaging centre, but did not attend the repeat assessment.

At each assessment centre visit a participant can self-report illnesses, which are coded in Field 20002. The illnesses are coded using Coding 6, as indicated on the Field 20002 page on Showcase. Clicking on the “6” of “Coding 6” on that page allows us to see a list giving the meanings of the codes given above.

For example: looking at the participant with eid 3315794 we see that at each of their three assessment centre visits they self-reported having asthma (code 1111). As the “first” condition reported this is assigned to have array index 0 (the final number in 20002-0.0 etc). At their imaging assessment visit (instance 2) they also report hypertension (code 1065), and this being the second reported condition at that visit it is assigned to array index 1, i.e. in field 20002-2.1.

Note that in reality field 20002 has array indices running from 0 to 33 (indicating at least one participant self-reported 34 illness codes), and so the real dataset would be considerably wider than that shown above even with only these two fields in it.

Note also that due to the nature of field 20002 being a self-report field (i.e. reported at an assessment centre), we can only have data for a particular instance index for field 20002 if that same instance index in field 53 has a value. For example, since the participant with eid 4582852 only attended baseline assessment they can only have values for field 20002 with instance index equal to 0.

The instance index is not exclusively used to refer to the assessment centre visit. For example, the “Diet by 24-hour recall” fields (see [Category 100090](#)) use instance 0 to refer to the baseline assessment centre (as above), but then instances 1 to 4 refer to the four on-line cycles of this questionnaire. As another example, reports from the cancer register (see [Category 100092](#)) are given a new instance index for each additional type of cancer reported.

As indicated above, the column headers appear slightly differently depending on which package you are using. The various output formats display the headers as follows:

File type	Column header	Notes	Example
csv	F-I.A		31-0.0
R	f.F.I.A	with f. preceding all fields	f.31.0.0
SAS	a_F_I_A	a indicates the type of variable, e.g. a will be n for numerical fields and s for string fields.	n_31_0_0
Stata	a_F_I_A	As for SAS.	n_31_0_0

**Table 2.8.2: Column header formats**

where, as previously, F represents the field number, I the instance index and A the array index

## 3 Bulk data

This section deals with accessing bulk data, such as imaging data (e.g. brain MRIs), accelerometer and ECG data, i.e. fields for which each item is a complex/compound dataset in itself.

These are accessed using a command line utility `ukbfetch`. This can be downloaded from the File Handlers tab on the [Download section](#) of Showcase. Both a Windows and Linux version of `ukbfetch` are available.

The `ukbconv` program will also usually be needed to generate a “bulk file” allowing the download of multiple bulk items at once (see section 3.2.2).

If you have a bulk data field in your project basket, there will be a column for it in your main dataset, however only the field ID will be present rather than the actual contents of the bulk data. The purpose is to indicate which participants have that bulk field available.

Note that `ukbfetch` creates a temporary file during the download, and then checks the MD5 checksum of the resulting file against its expected value. If the checksums do not agree then the download will fail. There is hence no separate validation step needed.

The sizes of some of the bulk field files are given for reference in Section 8.3 in the Appendix.

### 3.1 Connectivity and authentication

The bulk repository consists of a pair of mirrored systems each connected to the UK JANET network by independent links. The system names are:

- `biota.ndph.ox.ac.uk`
- `chest.ndph.ox.ac.uk`

To access bulk data your computer must be able to make http (Port 80) connections to at least one, and preferably both, of these systems. Please note that navigating to the above websites is not part of the download process; you simply need to ensure that your computer is able to connect to them. For most researchers this will not be a problem; however, please see section 8.2.6 for a way of checking this if you believe this may be an issue on your system. It is not possible to use a proxy server when using the `ukbfetch` utility.



In order to use `ukbfetch` it is necessary for you to authenticate yourself to the system. To do this you will need the authentication “keyfile” containing the 64-character password which was attached to the email notifying you that your data was ready to download (called `k56789r23456.key` where 56789 is replaced by your application ID and 23456 the run ID of the data extract). This is a simple text file containing your Application ID on the first line and the 64-character decryption password for that dataset on the second line.

The authentication keyfile should be saved in the folder where you will be running `ukbfetch`. The utility expects by default that the authentication keyfile has been renamed as `“.ukbkey”` (i.e. this is its full name with no other file extension). However, it is still possible to run the utility with the keyfile named differently by using the `-a` option (see section 3.2 for further details).

## 3.2 Using `ukbfetch`

The following two sections give general instructions for accessing Bulk data using the `ukbfetch` utility. Further details are given in [UKB Resource 644](#).

### 3.2.1 Downloading a single bulk item

We assume for illustration that a participant with eid 2143432 has data for the bulk [Field 20252](#) (T1 structural brain images - NIFTI). In a main dataset this will be indicated by the cell corresponding to the row with eid 2143432 and the column 20252-2.0 having the value 20252 in it (we are assuming the particular Field-Instance-Array format for a `.csv` file here; see section 2.8 for more information about this).

Note here that the instance index is 2 because the field was collected at a first imaging centre (instance 2) and the array index is 0 because only a single item of data was collected for this field at that centre.

To download the brain image for this participant we would use the command:

```
ukbfetch -e2143432 -d20252_2_0
```

assuming that the authentication keyfile has been renamed as `.ukbkey` (and placed in the same folder as `ukbfetch`). If the keyfile is instead called `k56789r23456.key` (for example) then the command would be:

```
ukbfetch -e2143432 -d20252_2_0 -ak56789r23456.key
```

Note that there must be no spaces between the flags (`-e`, `-a` etc) and the following arguments.

### 3.2.2 Creating and using a bulk file

To download many bulk fields at once, `ukbconv` can be used to generate a “bulk file” which lists participant eids and field numbers (including instance & array indices) for which that bulk field exists.

For example, let us assume we want to download all the T1 structural brain images for all participants at once.

Firstly, to generate the bulk file we run the command:

```
ukbconv ukb23456.enc_ukb bulk -s20252
```

where `ukb23456.enc_ukb` is our unpacked (but not converted) main dataset (see sections 2.5 & 2.6), and 23456 would be replaced by the run ID corresponding to your dataset.

The above command would output a file called **ukb23456.bulk** the first few lines of which would look something like:

```
3422567 20252_2_0
5321753 20252_2_0
2457842 20252_2_0
```

i.e. a simple list with each row the eid of a participant and the Field\_Instance\_Array of the relevant data.

Note that we cannot specify particular instance and array indices in the `ukbconv` call as the `-s` flag does not have this functionality. If this is a problem the bulk file can be edited using an appropriate software package to keep only the particular instances/arrays required.

Note that the `-i` flag for `ukbconv` can replace the `-s` flag to select a group of fields rather than a single one as in the example (see section 2.6.3).

Next, using our bulk file we can now use the command:

```
ukbfetch -bukb23456.bulk
```

to download every file for Field 20252. Once again there should be no space between the `-b` flag and the filename.

We can limit the number of files we download at once using ukbfetch by using the -s and -m flags. There is a limit of 50,000 files per ukbfetch call and so this will sometimes be an essential element of the process.

The flag -s gives the starting row of our bulk file to work from, and the -m flag sets how many rows from the bulk file we process.

For example, we could download 5000 files at a time for the above field by running the following commands one by one:

```
ukbfetch -bukb23456.bulk -s1 -m5000
ukbfetch -bukb23456.bulk -s5001 -m5000
ukbfetch -bukb23456.bulk -s10001 -m5000
ukbfetch -bukb23456.bulk -s15001 -m5000
ukbfetch -bukb23456.bulk -s20001 -m5000
```

Assuming that there are less than 25000 participants with this field, which is true at the time of writing, this would download all files for field 20252.

These commands could also be added to a batch file / shell script and run in one go. In this case there is -o flag which can be used to specify a different name for the logfile for each call of the ukbfetch utility.

Further details for using ukbfetch are given in [UKB Resource 644](#).

## 4 Genetics data

This section deals with accessing the genetics data. There are a variety of different types of genetics data available through UK Biobank, and different methods are used for downloading the different types.

Note that genetics fields will have a corresponding column in your main dataset, but in the same way as for bulk fields only the field ID will be present rather than the actual contents of the bulk data. The purpose is to indicate which participants have that genetics field available.

### 4.1 Genotype fields

Some genotype data fields appear in the main dataset, some can be downloaded using the gfetch utility (for multi-person files), some need to be downloaded using the ukbfetch utility (for single-person files), and some can be downloaded from the European Genome-phenome Archive (EGA). Some types of genetics data can be downloaded using more than one of these methods.

Most of the relevant information is shown on the page [Category 263 \(Genotypes\)](#) and in [UKB Resource 668](#). Access via the EGA requires an account to be set up through the Access Management Team (AMT)

Some information about the method of download is also given on the Notes tab of the relevant field. For example, the CEL files, [Field 22002](#), need to be downloaded using ukbfetch using the same methods as described in section 3.

Further information about the Genotyping data, including the size of the files, and using the EGA is available in the "Genotyping and Imputation FAQs" on the following page: [UKB Genetic data](#).

The data in the Genotype BED and BGEN files appear in a common order for all researchers. In order to match your participant eids to the data (which is done by position) it is necessary to use gfetch to download appropriate FAM and sample files.

## 4.2 Using gfetch

The gfetch utility can be used to download some parts of the genotype data, as described in [Category 263 \(Genotypes\)](#) and in [UKB Resource 668](#), in particular it is used to create the FAM and Sample files for a project to match the project eids, by position, to the BED and BGEN files.

Only a Linux version of gfetch is available (see section 8.2.5 for a way to proceed if you do not have Linux available).

Note that gfetch creates a temporary file during the download, and then checks the MD5 checksum of the resulting file against its expected value. If the checksums do not agree then the download will fail. There is hence no separate validation step needed.

### 4.2.1 Connectivity & authentication

The bulk repository consists of a pair of mirrored systems each connected to the UK JANET network by independent links. The system names are:

- biota.ndph.ox.ac.uk
- chest.ndph.ox.ac.uk

To access bulk data your computer must be able to make http (Port 80) connections to at least one, and preferably both, of these systems. Please note that navigating to the above websites is not part of the download process; you simply need to ensure that your computer is able to connect to them. For most researchers this will not be a problem; however, please see section 8.2.6 for a way of checking this if you believe this may be an issue on your system. It is not possible to use a proxy server when using the gfetch utility.

In order to use gfetch it is necessary for you to authenticate yourself to the system. To do this you will need the “keyfile” containing the 64-character password which was attached to the email notifying you that your data was ready to download (called k56789r23456.key where 56789 is replaced by your application ID and 23456 by the run ID of the data extract). This is a simple text file containing your Application ID on the first line and the 64-character decryption password for that dataset as the second line.

The authentication keyfile should be saved in the folder where you will be running gfetch. The utility expects the authentication keyfile to be renamed as “.ukbkey” (i.e. this is its full name with no other file extension). However, it is still possible to run the utility with the keyfile named differently by using the -a option (see section 4.1.3).

### 4.2.2 A gfetch example

A researcher has gained access to the genotype calls by including [Field 22418](#) (Genotype calls) in their project basket, which has subsequently been approved. They wish to download the chromosome .bed file and its associated .fam file (i.e. the link file giving the order that their project eids appear in the .bed file).

They have downloaded gfetch from [Download 600](#) by running the wget command given on a Linux terminal. To make gfetch an executable file they have then run:

```
chmod 755 gfetch
```

They have also saved their authentication keyfile k56789r23456.key from their notification email (where 56789 is their application ID) into the same folder as gfetch, and renamed it as .ukbkey (this being the full filename).

To download the Genotype call .bed file for Chromosome 5, they enter the command:

```
gfetch 22418 -c5
```

To download the associated FAM file they use the command:

```
gfetch 22418 -c5 -m
```

Note that sometimes ./gfetch needs to be used in place of gfetch because of the way a Linux system is set up (see section 8.2.1). If the researcher had not renamed their keyfile, and left it with the filename k56789r23456.key, then they would have had to replace the above commands with:

```
gfetch 22418 -c5 -ak56789r23456.key
```

and

```
gfetch 22418 -c5 -m -ak56789r23456.key
```

Further information about the various options available with gfetch are given in UKB [Resource 668](#).

## 4.3 Exome sequences

A description of the Exome sequence fields are contained in [Category 170 \(Exome sequences\)](#) on Showcase.

The population level exome files in PLINK & pVCF formats (Fields [23155](#) & [23156](#) respectively) are downloaded using gfetch as described in the Notes of those fields.

The VCF & CRAM files are downloaded using ukbfetch as described in section 3.2.

Further information about the Exome sequence data is contained in the "Exome Sequencing FAQs" on the following page: [UKB Genetic data](#).

## 5 Record-level data

### 5.1 Record-level data on the Data Portal

The record-level data is available from the record repository accessed via the Data Portal. The record repository is divided into a number of interconnected database tables covering: hospital inpatient data, death data, primary care (GP) data & COVID-19 test results.

Information about the tables available for each of these data types can be found in the Resources for their relevant sections on Showcase:

- [Category 2000](#) for hospital inpatient data;
- [Category 100093](#) for death data;
- [Category 3000](#) for primary care data (covering approximately 45% of the cohort);
- [Category 999](#) & the [COVID-19 data](#) page of the Showcase Essential Information for test results and additional primary care data available for research relating to COVID-19.

### 5.2 Gaining access to the Data Portal

Access to each table on the Data Portal is granted to a research project on a table-by-table basis by including a specific data-field in a project basket.

For example, including [Field 41259](#) in a basket will give access to the main HESIN (hospital inpatient) table. Similar fields can be found in each of the categories above.

The main dataset will include a column for each such field but the values shown in that column will be a count of the number of rows that each participant has in the corresponding table.

Once access to a table on the Data Portal has been approved, the Data Portal can be accessed by:

1. Logging in to the Access Management System;
2. Navigating to the relevant project (click Projects then View/Update);
3. Selecting the Data tab;
4. Clicking on the "Go to Showcase to refresh or download data" button;
5. This will lead to the Downloads page where, if approved for record data, there will be a "Data Portal" tab;
6. Clicking on the Connect button will open up the portal.

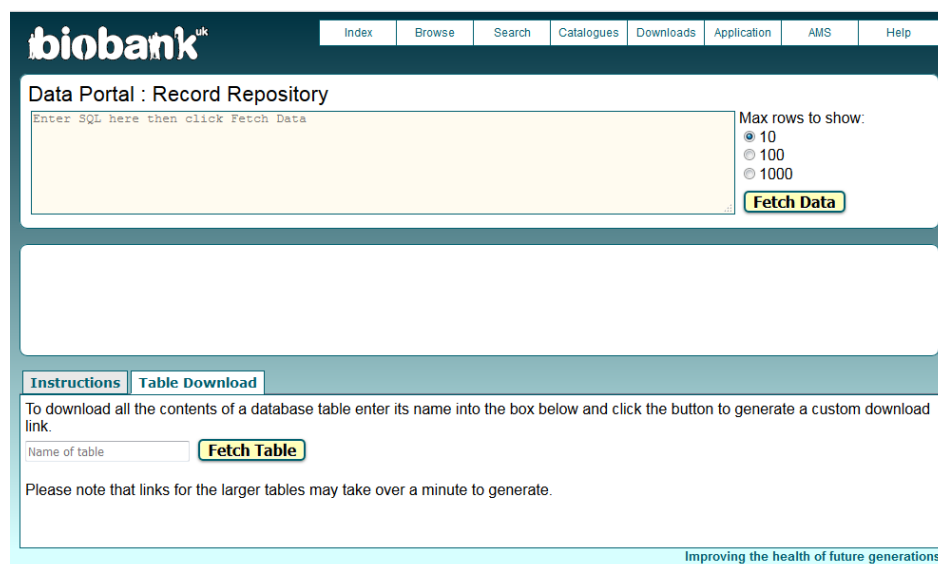


Only the project Principal Investigator (PI) and collaborators with delegate rights are able to access the Data Portal. The project PI can assign delegate rights to other collaborators by using the Collaborators tab on the AMS.

### 5.3 Downloading tables from the Data Portal

Once a researcher has accessed the data portal they can download each complete table as shown below, or query the data prior to downloading it (see section 5.4).

To download a complete table click on the 'Table Download' tab in the bottom panel, enter the name of the table you wish to download (e.g. hesin\_diag) and click on the 'Fetch Table' button as shown in Figure 5.3.1.



The screenshot shows the Biobank Data Portal interface. At the top is a navigation bar with links: Index, Browse, Search, Catalogues, Downloads, Application, AMS, and Help. Below this is a header section titled 'Data Portal : Record Repository'. It contains a text input field with the placeholder 'Enter SQL here then click Fetch Data' and a 'Fetch Data' button. To the right of the input field is a 'Max rows to show:' section with radio buttons for 10 (selected), 100, and 1000. Below the header is a large empty white box. At the bottom is a panel with two tabs: 'Instructions' and 'Table Download'. The 'Table Download' tab is active and contains the following text: 'To download all the contents of a database table enter its name into the box below and click the button to generate a custom download link.' Below this text is a text input field labeled 'Name of table' and a 'Fetch Table' button. At the bottom of the panel, there is a note: 'Please note that links for the larger tables may take over a minute to generate.' The Biobank logo is in the top left corner, and the tagline 'Improving the health of future generations' is at the bottom right.

**Figure 5.3.1: Table download tab**

This will generate a custom download link that you can paste into a web browser and a wget command for those using a linux system. The resulting dataset will be provided as a tab separated text file (.txt). Please note it can take some time to download the complete tables.

### 5.4 Using SQL to query the tables

An alternative to downloading whole tables is to use SQL statements to do simple explorations or select data of interest prior to download.

SQL (Structured Query Language) is the control language used to manage and manipulate information within most modern relational databases. If you do not know SQL already then there are a number of free tutorials available on the web.

Each major database uses a slightly different dialect to that of other vendors, however most common commands are identical across them. The UK Biobank system uses the Ingres platform to host its relational databases. A reference manual is available online and can be located by an internet search for “Ingres 10.2 SQL Reference Guide”.

Some examples of SQL statements that can be used to investigate the record-level data are given in each Resource particular to that type of data located in the Categories given in Section 5.1 above.

## 6 Returned datasets

“Returns” are datasets returned by researchers who have used UKB data in their research. Some returned datasets are incorporated into the main resource, but those that have not been need to be downloaded using the ukblink utility.

The ukblink utility can be downloaded from the File Handlers tab on the [Download section](#) of Showcase. Both a Windows and Linux version of ukblink are available.

### 6.1 Authentication

In order to provide authentication for the download you will need to have your authentication keyfile in the same folder as ukblink (this is the attachment to your notification email with a name like k56789r23456.key). This is the same requirement as for ukbfetch and gfetch (see, for example, section 3.1 for more details).

### 6.2 Using ukblink

We use [Return 1362](#) as an example. We assume that we have been granted access to this dataset, that we have downloaded the ukblink utility (and if using Linux, made it executable; see section 8.2.1) and moved our keyfile into the same folder.

To download it we use the command:

```
ukblink -r1362
```

assuming our keyfile has been renamed as .ukbkey. Otherwise we use:

```
ukblink -r1362 -ak56789r23456.key
```

assuming the keyfile still has its original filename. (In Linux we may need to replace ukblink by ./ukblink, see section 8.2.1.)

Note that files will download as generic .dat files. More recent Returns are in fact all .zip files and renaming them as such should allow standard unzipping programs to be run. Older files may either be .zip or .7z files. A list of which of the older Returns has which type of zipped file format is included in the Appendix (section 8.4).

Some returned datasets provide participant-level data, and for these the ukblink utility also allows the creation of a bridge to connect your project eids with those used in the Return.

Return 1362 is an example of a Return that includes participant-level data (this can be seen from its [Showcase page](#) in the “Personal” row). In order to download the bridge we need to know the Application that this Return was generated as part of. This can be determined from the first line of its Showcase page where we can see that it was part of Application 2964.

Hence, to generate the appropriate bridge file we use the command:

**ukblink -b2964**

(adding **-ak56789r23456.key** if appropriate).

Further details for accessing Returns using the ukblink utility are given in [UKB Resource 655](#).

## 7 Bridges

### 7.1 Linking to Genetic data

Given the size of the genetics data, some projects will be given approval to link to a institution-held copy of the data rather than each project being required to have a separate copy. Any project accessing genetics data, even through a dataset downloaded by a different project, must have the relevant genetic fields included in an approved basket for their own project.

The genetics data appears in a common order for different projects, and the appropriate link file (FAM or sample file) for a project then provides the order in which the participants appear in the data.

All that is necessary for a new project to link to a genetics dataset downloaded by another project is for them to generate the appropriate link file using gfetch, so as to determine the order that their eids appear in the genetics data.

Note that if the 'owner' of a genetics dataset, i.e. the project who originally downloaded it, is approached to share a genetics dataset they should confirm with UK Biobank (at [access@ukbiobank.ac.uk](mailto:access@ukbiobank.ac.uk)) that the appropriate approvals are in place before allowing access to the data. They should also ensure that they have seen the fully executed MTA for the other project, with genetic data included.

Note that approval to reuse a genetics dataset in this way does not permit projects to share processed data with each other directly, or to construct a bridge to share other elements of UK Biobank data. This would constitute a breach of the project's Material Transfer Agreement (MTA).

### 7.2 Bridge files for bulk fields

In some instances UK Biobank will release bridging files to link two separate UK Biobank applications together, in order for bulk images and other bulk fields to be shared between projects.

UK Biobank is currently reviewing its procedures with regards to bridging files and will be providing updated information in due course.

## 8 Appendix

### 8.1 Using a command prompt in Windows

If you are using Windows:

- **Windows XP:** go to Start > All Programs > Accessories > Command Prompt
- **Windows Vista:** go to Start > type cmd in the Search bar, and click on Command Prompt once it has appeared
- **Windows 7:** go to Start > All Programs > Accessories > Command Prompt
- **Windows 8:** go to Start > type cmd in the Search bar, and click on Command Prompt once it has appeared
- **Windows 10:** go to the Search icon > type cmd in the Search bar, and click on Command Prompt once it has appeared

For any version of Windows, if the Command Prompt does not appear by following the steps above, please press the following combination of keys: Windows+R. (The Windows key is located between the Ctrl and the Alt keys on your keyboard). This will open a small window named "Run". Type cmd in the "Open:" space, then click OK. This will open a Command Prompt window.

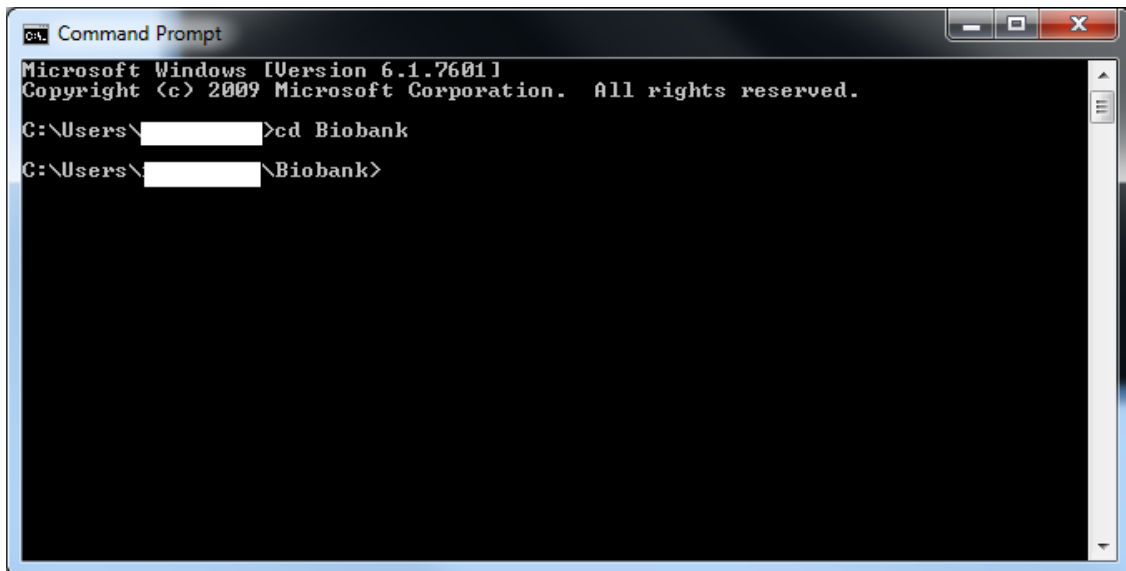
Once it is opened, the Command Prompt window should display only a bit of text at the top, and then a blinking cursor preceded by a directory address on your computer (by default, this should be **C:\Users\YourUserName**), as shown in Figure 8.1.1.



Figure 8.1.1: The command prompt window

The next step is to navigate to the directory in which you previously downloaded all the helper files, the encoding file and your dataset. To do this, type **cd** followed by the path that you wish to navigate to, from the current folder.

In our example, we downloaded the files in a directory named Biobank, which is located in the home directory for the user. All we need to do is type **cd Biobank** and press Enter to navigate to the Biobank directory, as shown in Figure 8.1.2.

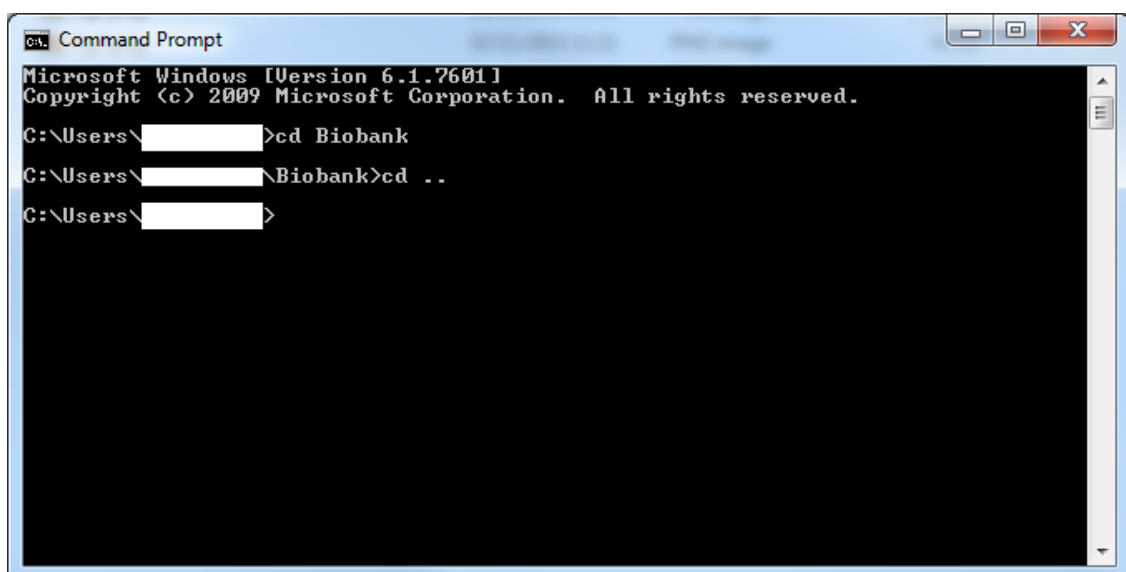


```
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\>cd Biobank
C:\Users\>\Biobank>
```

**Figure 8.1.2: Changing the directory**

Note that you can also use **cd** followed by two dots (**cd ..**) to go back to the parent directory, as shown in Figure 8.1.3.

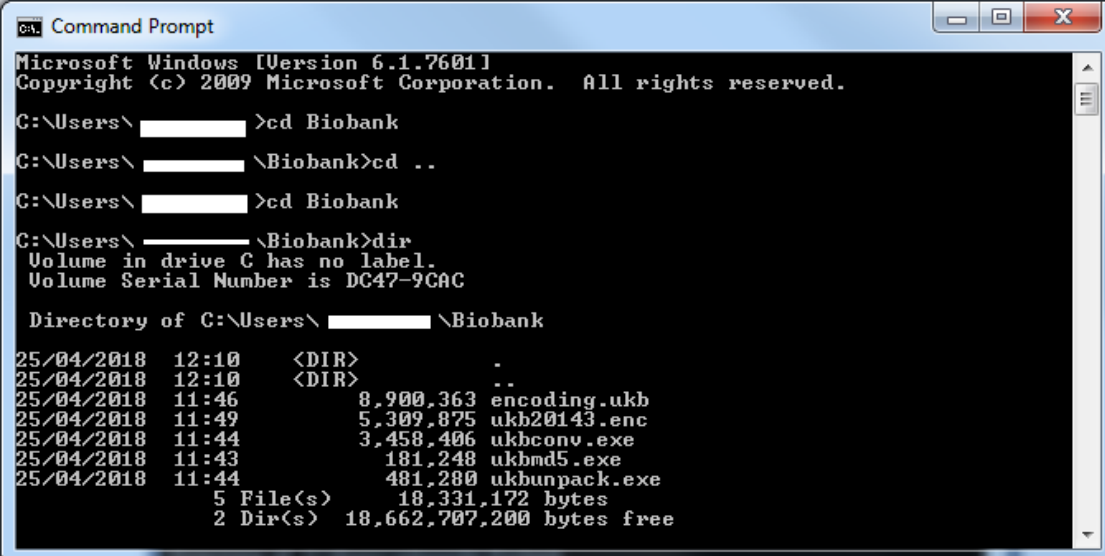


```
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\>cd Biobank
C:\Users\>\Biobank>cd ..
C:\Users\>
```

**Figure 8.1.3: Moving up a directory**

Use the **cd** command to navigate to the chosen directory. Once you are in the right directory, you can use the **dir** command to list all the files in the current directory (Figure 8.1.4). This allows you to check that you are indeed in the right place: the **dir** command should display the name of the 5 files that you previously downloaded.



```
ca. Command Prompt
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\ [redacted] >cd Biobank
C:\Users\ [redacted] \Biobank>cd ..
C:\Users\ [redacted] >cd Biobank
C:\Users\ [redacted] \Biobank>dir
Volume in drive C has no label.
Volume Serial Number is DC47-9CAC

Directory of C:\Users\ [redacted] \Biobank

25/04/2018  12:10    <DIR>          .
25/04/2018  12:10    <DIR>          ..
25/04/2018  11:46             8,900,363  encoding.ukb
25/04/2018  11:49             5,309,875  ukb20143.enc
25/04/2018  11:44             3,458,406  ukbconv.exe
25/04/2018  11:43             181,248  ukbmd5.exe
25/04/2018  11:44             481,280  ukbunpack.exe
                5 File(s)              18,331,172 bytes
                2 Dir(s)  18,662,707,200 bytes free
```

Figure 8.1.4: Displaying the contents of a directory

## 8.2 Issues with helper files & utilities

### 8.2.1 General

- If you are trying to run ukbunpack, ukbconv etc in a Windows environment and receive an “Access is denied” error, then it is likely you do not have permissions set to run executable files which are unknown to the system. You may need to log on as an Administrator or contact your local IT support for assistance.
- If you are working in a Linux environment having downloaded a utility such as ukfetch, it will not by default be recognised as being executable. To fix this use the command:

**chmod 755 ukbfetch**

You may also find that your system cannot locate “ukbfetch” because it does not search the current working directory when looking for executable files. The easiest way around this is to prefix the command as follows:

**./ukbfetch <other parameters>**



to indicate to your system that ukbfetch is located in the current directory (designated by a dot . ).

- A “malloc” error, meaning a “memory allocation error”, is encountered when your computer runs out of working memory during the download. This is particularly prone to happening when the Windows versions of the helper files / utilities are being used, and so if this happens we recommend you use the Linux versions instead.

### 8.2.2 ukbunpack

- When attempting to unpack the dataset, if you receive the error:

*FAIL: Unpack : failed to get uncompressed data - uncompression failed*

you are probably using the wrong 64-character Password. Please note that the main dataset can only be unpacked using the Password from the keyfile k56789r23456.key contained in the notification email for that particular dataset, i.e. for the dataset released as run 23456. You cannot reuse the keyfile Password from a different data release on the same project.

### 8.2.3 ukbconv

- While using the ukbconv utility, some researchers, depending on the variables in their dataset, may see the following error message appear in the command-line terminal:

*Rosetta error: ROSETTA Error: member "eXXX" not found  
Validity error: ROSETTA Error: member "eXXX" not found  
(XXX can be any integer)*

This bug is being investigated at the moment, but this message does not affect the conversion process in any way, and has no consequence on the data being extracted. Researchers can directly open the files generated by ukbconv without worrying about these errors.

### 8.2.4 ukbfetch

- If you are running ukbfetch with a bulk file and are receiving an error indicating that it cannot find data for a particular eid/field combination, then this might be because you created a bulk file (using ukbconv) containing fields which are not accessed using ukbfetch. For instance, if ukbconv is run with the bulk option without specifying a particular field (or set of fields), it will include genomics fields that need to be downloaded using gfetch in amongst the bulk fields. Hence, most fields appearing in

the bulk file will fail to download because it is not possible to access their data in this way.

### 8.2.5 gfetch

- When running gfetch you may find that the data appears to be “fetched” properly, but then cannot be “written”, causing the process to abort. This is most likely due to the large size of some of the genetics data (particularly the imputed data) overwhelming the local storage available during the download. We recommend contacting your local IT support to deal with this issue.
- If you only have a Windows computer available, it is possible to set up a Linux shell to run within it from which you can run gfetch. Googling “running linux on windows” or similar will provide links describing how to do this.

### 8.2.6 ukbfetch / ukblink / gfetch

- If you are uncertain whether your IT system will allow you to access the websites:
  - biota.ndph.ox.ac.uk
  - chest.ndph.ox.ac.uk

needed for bulk and some genetics data, then the command line utility ping can be used to check the connection. From a Windows command line the command:

```
ping biota.ndph.ox.ac.uk
```

will send four signals to the website and report if a reply is received. In Linux the command:

```
ping biota.ndph.ox.ac.uk -w4
```

has the same effect (the `-w` flag is to limit the number of signals sent, which otherwise will continue until Ctrl-C is entered).

- The keyfile (received as the attachment to your notification email) needs to be in the same directory as the utility. Both utilities by default expect it to have been renamed as `.ukbkey` (note that this as its full name, with no other file extension). This can cause problems in Windows, and hides the file in Linux (`ls -a` will show such “hidden files”). If you prefer to give a different name to the keyfile, then `ukbfetch`, `ukblink` & `gfetch` can still be run, but you will need to add `-ak56789r23456.key` to the end of your command where `k56789r23456.key` is replaced by the name of your keyfile.

- If you get the error:

*Invalid authentication file*

*File names must be 1-64 characters long*

it is because you have put a space between the -a and the keyfile name.

- When using a Linux system, if you receive an error along the lines of:

*`GLIBC\_2.14' not found (required by ukbfetch)*

it means that your local Linux libraries are not compatible with our standard versions of the utility ukbfetch (in this example), ukblink or gfetch. In each case it is possible to create a version of the utility that will run on your system. See [Resource 645](#) (for ukbfetch), [Resource 656](#) (for ukblink) or [Resource 669](#) (for gfetch), for further details.

### 8.3 Sizes of bulk fields

The following table gives the approximate size, per participant, of a number of the bulk fields available:

Field	Name	Estimated size per participant (MB)
20158	DXA images	2
20201	Dixon technique for internal fat - DICOM	71
20202	Pancreatic fat - DICOM	9
20203	Liver images	1
20204	OCRM experimental sequence - DICOM	3
20206	Measurements of pancreas volume - DICOM	2
20207	Scout images for heart MRI - DICOM	7
20208	Long axis heart images - DICOM	9
20209	Short axis heart images - DICOM	81
20210	Aortic distensibility images - DICOM	6
20211	Cine tagging images - DICOM	5
20212	Left ventricular outflow tract images - DICOM	4
20213	Blood flow images - DICOM	5
20214	Experimental shMOLLI sequence images - DICOM	4
20215	Scout images for brain scans - DICOM	5
20217	Functional brain images - task - DICOM	244
20218	Multiband diffusion brain images - DICOM	128
20224	Phoenix - DICOM	<1
20225	Functional brain images - resting - DICOM	360
20249	Functional brain images - task - NIFTI	453
20250	Multiband diffusion brain images - NIFTI	1047
20251	Susceptibility weighted brain images - NIFTI	33
20252	T1 structural brain images - NIFTI	51
20253	T2 FLAIR structural brain images - NIFTI	34
25747	Eprime advisor file	<1
25748	Eprime txt file	<1
25749	Eprime ed2 file	<1
25750	rfMRI full correlation matrix, dimension 25	<1
25751	rfMRI full correlation matrix, dimension 100	<1
25752	rfMRI partial correlation matrix, dimension 25	<1
25753	rfMRI partial correlation matrix, dimension 100	<1

## 8.4 File types of returned datasets

The following table gives the zipped format used for older returned datasets. If a return is not on this table then it will be a newer file in .zip format. The file downloaded by ukblink should be renamed to the correct file type, and standard utilities used to unzip the file.

Return id	Title	Extension
124	Derived variables from application 735/ 15716 - myopia variables	7z
146	5 year mortality predictors in 498 103 UK Biobank participants: a prospective population-based study	zip
147	Built Environment Data for Bristol	zip
164	Suitability of UK BIOBANK Retinal Images for Automatic Analysis of morphometric properties of the vasculature	zip
210	Built Environment Data - Newcastle and Middlesbrough	7z
263	Variants near CHRNA3/5 and APOE have age- and sex-related effects on human lifespan.	zip
265	The effect of functional hearing and hearing aid usage on verbal reasoning in a large community-dwelling population	zip
362	Built Environment Data for Birmingham and Nottingham	zip
363	Built Environment Data for Oxford	zip
403	New reference values for body composition by bioelectrical impedance analysis in the general population: results from the UK Biobank	7z
408	Parental diabetes and birthweight in 236 030 individuals in the UK Biobank Study	7z
421	Chronic widespread bodily pain is increased among individuals with history of fracture: findings from UK Biobank	7z
423	Do smoking habits differ between women and men in contemporary Western populations? Evidence from half a million people in the UK Biobank study.	zip
424	Characteristics of rheumatoid arthritis and its association with major comorbid conditions: cross-sectional study of 502 649 UK Biobank participants.	7z

463	Heaviness, health and happiness: a cross-sectional study of 163066 UK Biobank participants	7z
464	Psychiatry Gender differences in the association between adiposity and probable major depression: a cross-sectional study of 140,564 UK Biobank participants	7z
473	The effect of functional hearing loss and age on long- and short-term visuospatial memory: evidence from the UK Biobank resource	7z
474	Better visuospatial working memory in adults who report profound deafness compared to those with normal or poor hearing: Data from the UK Biobank resource	7z
501	Cognitive function and lifetime features of depression and bipolar disorder in a large population sample: Cross-sectional study of 143,828 UK Biobank participants	7z
504	Low birth weight and features of neuroticism and mood disorder in 83545 participants of the UK Biobank cohort	7z
508	Prevalence and Characteristics of Probable Major Depression and Bipolar Disorder within UK Biobank: Cross-Sectional Study of 172,751 Participants.	7z
509	Associations between single and multiple cardiometabolic diseases and cognitive abilities in 474 129 UK Biobank participants.	7z
511	Adiposity among 132,479 UK Biobank participants; contribution of sugar intake vs other macronutrients	7z
513	Cognitive Test Scores in UK Biobank: Data Reduction in 480,416 Participants and Longitudinal Stability in 20,346 Participants.	7z
526	Change in commute mode and body mass index: prospective longitudinal evidence from UK Biobank	7z
527	Active commuting and obesity in mid-life: cross-sectional, observational evidence from UK Biobank	7z
529	Lifestyle factors and prostate-specific antigen (PSA) testing in UK Biobank: Implications for epidemiological research	7z
534	Ethnic differences in sleep duration and morning-evening type in a population	7z

535	Smoking, screen-based sedentary behaviour, and diet associated with habitual sleep duration and chronotype: data from the UK Biobank	7z
536	Interactive effects of sleep duration and morning/ evening preference on cardiovascular risk factors	7z
542	Multiple novel gene-by-environment interactions modify the effect of FTO variants on body mass index	7z
547	The influence of social interaction and physical health on the association between hearing and depression with age and gender	7z
584	Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk	zip
702	Case-control association mapping by proxy using family history of disease	zip
717	Genome-wide association study identifies 74 loci associated with educational attainment	zip
718	Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analysis	zip
723	Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence	zip
726	Linkage disequilibrium - dependent architecture of human complex traits shows action of negative selection	zip
735	Red blood cell distribution width: Genetic evidence for aging pathways in 116,666 volunteers	zip
736	Mixed model association for biobank-scale data sets.	zip
739	Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets	zip
744	Genome-wide association study reveals ten loci associated with chronotype in the UK Biobank.	zip
745	Genome-wide association analyses of sleep disturbance traits identify new loci and highlight shared genetics with neuropsychiatric and metabolic traits	zip

749	Genome-wide association study of alcohol consumption and genetic overlap with other health-related traits in UK Biobank (N=122117)	zip
752	Genome-wide associations for birth weight and correlations with adult disease	zip
760	Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N=112151)	zip
762	Molecular genetic aetiology of general cognitive function is enriched in evolutionarily conserved regions	zip
776	Rare coding variants pinpoint genes that control human hematological traits	zip
777	An erythroid-specific ATP2B4 enhancer mediates red blood cell hydration and malaria susceptibility	zip
783	Cognitive performance among carriers of pathogenic copy number variants: Analysis of 152,000 UK Biobank subjects	7z
792	The 'Cognitive footprint' of psychiatric and neurological conditions: cross-sectional study in the UK Biobank Cohort	7z
793	Visualization of cancer and cardiovascular disease co-occurrence with network methods	7z
796	Psychological distress, neuroticism, and cause-specific mortality: early prospective evidence from UK Biobank	7z
981	Volumetric measurements of body composition derived from abdominal MRI - application 23889	zip
1362	Derived variable from cardiac MRI	zip
1363	Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort	zip
1364	Built Environment data for Edinburgh and Glasgow	zip
1365	Built Environment data for Greater London Authority	zip
1366	Built Environment data for Liverpool, Manchester and Bury	zip
1367	Built Environment data for Leeds and Sheffield	zip
1368	Built Environment data for Stoke	zip
1369	Built Environment data for Wales	zip



1455	Genetic evidence that lower circulating FSH levels lengthen menstrual cycle, increase age at menopause and impact female reproductive health	zip
1456	Events in Early Life are Associated with Female Reproductive Ageing: A UK Biobank Study	zip
1458	Vitreoretinal interface abnormalities in middle-aged adults with visual impairment in the UK Biobank study: prevalence, impact on visual acuity and associations	zip
1461	Monocular and binocular visual impairment in the UK Biobank study: prevalence, associations and diagnoses	zip
1465	Cost-effectiveness of the polypill versus risk assessment for prevention of cardiovascular disease	zip
1468	Sex differences in body anthropometry and composition in individuals with and without diabetes in UK Biobank	zip
1469	Women's reproductive health factors and body adiposity: findings from the UK Biobank	zip
1470	Differences in morning-evening type and sleep duration between black and white adults: Results from a propensity-matched UK Biobank sample	zip
1472	Calcium and Vitamin D supplementation are not associated with risk of incident ischemic cardiac events or death: Findings from the UK Biobank Cohort	zip
1475	Number of offspring and cardiovascular disease risk in men and women	zip
1476	Chronic multisite pain in major depression and bipolar disorder: cross-sectional study of 149, 611 participants in UK Biobank	zip
1480	Associations between active commuting and incident cardiovascular disease, cancer and mortality: prospective cohort study	zip
1491	Human CCL3L1 copy number variation, gene expression, and the role of the CCL3L1-CCR5 axis in lung function	zip
1502	Long-term intra-individual reproducibility of heart rate dynamics during exercise and recovery in the UK Biobank cohort	zip
1504	Bone mineral density and risk of type 2 diabetes and coronary heart disease: A Mendelian randomization study	zip

1522	Genetic prediction of male pattern baldness	zip
1541	Self-Reported Facial Pain in UK Biobank Study: Prevalence and Associated Factors	zip