UK Biobank - Genotyping and Imputation Data Release

May 2017

This document provides further information for the release of genotyping and imputation data for all 500,000 participants in UK Biobank. It has been collated based on questions received by UK Biobank's Access Team alongside information we believe will be of most interest to researchers.

Please note this FAQ does not replace the full documentation that will be provided in due course for researchers to interpret the data and have confidence in the quality assurance processes that have been applied. These FAQs may be updated over time and the current version will be made available via the UK Biobank website.

List of Frequently Asked Questions (FAQs):

1. **Can you confirm if we have already requested/had approval for the full genetic dataset?**
2. **We have approval to use genetic data from 150,000 participants. Will we be granted automatic access to the data when it is available or do we need to update our application?**
3. **We do not have approval for genetic data. How do we request approval?**
4. **We are working on storage solutions to receive the genotyping and imputation dataset. Can you be more specific in terms of the size of the data being made available?**
5. **We are concerned that some researchers may be given a head-start. How will UK Biobank ensure a level playing field?**
6. **How will UK Biobank determine that researchers have had sufficient time to download the genotyping and imputation data?**
7. **How will you be distributing the data and how will you meet the needs of your international scientific users?**
8. **Do you have any plans to build a GWAS results data repository in UK Biobank?**
9. **UK Biobank had previously communicated that these data would be released during 2016. Why has this release been delayed?**
10. **Can we still link to a key genetic dataset?**
11. **What should we do now to get the data?**
12. **Do all the files need to be downloaded at once or can we choose what to/not to download? Will these data be available for download at a later time or is there a time-limit beyond which this will not be possible?**
13. **When will the data be available?**
14. **Where are the UK Biobank datasets located within the EGA?**
15. **How do I access / login to the EGA to download these data?**
16. **Is there information available on how to configure our firewall to connect to the EGA?**
17. **What is the process for downloading these data from the EGA?**
18. **How do I use the EGA's own download streaming client?**
19. **Where do I go for further information relating to the EGA and download instructions?**
20. **When will application-specific files be made available via the UKB Showcase?**
21. **Which files will be available via the UKB Showcase versus the EGA?**
22. **How will updated genetic phenotypes (i.e. principal components, relatedness) be released?**
23. **Will the EGA be the only route for downloading the core genotyping and imputation data?**
24. **How will the Haplotypes be made available?**
25. **How will the CEL files be made available and when will they be released?**
26. **When will the Chr X (and remaining sex chromosomes) be made available?**
27. **Why am I seeing 'bad magic number' or 'bad decrypt' errors when decrypting files?**

1. **Can you confirm if we have already requested/had approval for the full genetic dataset?**

   Anyone with approval for the 150,000 (interim) genotype data release has approval for the full release. There are currently 96 data-fields in total ranging from 22000 – 22325 and you can check which have been requested per application by checking Annex A of your Material Transfer Agreement (and any later Addenda).

2. **We have approval to use genetic data from 150,000 participants. Will we be granted automatic access to the data when it is available or do we need to update our application?**

   Provided your application is approved and contains genetics-related fields, and that you have requested an EGA account to be created on your behalf, then you will be able to access the data when they are available.

3. **We do not have approval for genetic data. How do we request approval?**

   If your application has not been approved for genetics-related fields, you can email the Access Team to request a project scope extension with justification for why you would like to conduct further analyses.

4. **We are working on storage solutions to receive the genotyping and imputation data. Can you be more specific in terms of the size of the data being made available?**

   The uncompressed file size for the data being made available now is around 12TB. The raw genotyping CEL files are a further 14TB however these will be of interest to only a subset of the research community and will be made available later. The table below provides further details of the package, formats and sizes of this release.

Data	UKB File Naming	N	GB
Genotyping Data			
Genotyped SNP index *	ukb_snp_chrN_v2.bim	1-26	0.03
Calls	ukb_cal_chrN_v2.bed	1-26	92
Confidences	ukb_con_chrN_v2.txt	1-26	2900
Intensities	ukb_int_chrN_v2.bin	1-26	2900
CNV B-allele-freq	ukb_baf_chrN_v2.txt	1-26	1500
CNV log ratio	ukb_l2r_chrN_v2.txt	1-26	2300
Sample QC	ukb_sqc_v2.txt	n/a	0.3
Imputation Data			
Imputation SNP index *	ukb_bgi_chrN_v2.bgi	1-22	4
Imputed data	ukb_imp_chrN_v2.bgen	1-22	2100
Minor-allele freq + info scores *	ukb_mfi_chrN_v2.txt	1-22	4
HLA	ukb_hla_v2.txt	n/a	0.4

Notes:

- The family /sample files *.fam (genotyping) and *.sample (imputation) are generated dynamically and will be downloaded directly from the UK Biobank Showcase.
- Items marked * are not person-specific and will also be made available publicly via the UK Biobank Showcase.
- Lettered chromosomes are mapped to numbers as X=23,Y=24, XY=25, MT=26.
- The compressed file sizes held at the EGA are ~4TB (genotyping) and ~2TB (imputation).

5. **We are concerned that some researchers may be given a head-start. How will UK Biobank ensure a level playing field?**

   UK Biobank operates a policy of no preferential access for any group. We are making the genotyping and imputation data files available for approved researchers in an encrypted format in order that the data can be downloaded over a period of a few weeks (and to provide time to resolve any issues that any researchers have with getting access to these data).

   At the end of this period, the de-encryption keys and the mapping files that allow the data to be used will be made available to all researchers who have an approved project underway at the same time.

6. **How will UK Biobank determine that researchers have had sufficient time to download the genotyping and imputation data?**

   UK Biobank is working with the European Genome Archive to help distribute these data. The download mechanism allows us to track which users have downloaded complete datasets. We will monitor download traffic and use this to inform when de-encryption keys are released.

   We recognise that not all research groups will want these data straightaway, and provided there has been no significant contention or other technical issues, we will release the keys once the majority of users whom we know want immediate access have had the opportunity to download and/or where we see download traffic reducing.

7. **How will you be distributing the data and how will you meet the needs of your international scientific users?**

   UK Biobank is working closely with the European Genome Archive. Data will be distributed from three physical sites – two located in the UK and the third located in Barcelona. We will be monitoring download traffic during the period and will have some flexibility to balance workload across the three sites.

   The majority of users will be taking advantage of the EGA's own download streamer. The use of FTP and Aspera client may also be offered as an alternative.

8. **Do you have any plans to build a GWAS results data repository in UK Biobank?**

   All the GWAS datasets returned will be made available to registered researchers through the UKB **Showcase**.

9. **UK Biobank had previously communicated that these data would be released during 2016. Why has this release been delayed?**

   There has been significant effort expended to ensure the quality of the genotyping and imputation data. This has inevitably taken time to get right and is not impervious to the usual delays experienced by projects of this nature. We are keenly aware that any delay can result in frustration to our researchers and have worked as fast as possible to get these data released.

10. **Can we still link to a key genetic dataset?**

    Yes. We will continue to support the ability for institutes to hold a key genetic dataset that is shared between multiple Access applications and that removes the need for them to store multiple copies.

    The EGA is being used to hold and distribute the majority of the genotyping and imputation files. Researchers will still need to utilise the UK Biobank **Showcase** to receive the family/sample files that map the ordering of the genotyping and imputation data to their pseudonymised IDs. The ordering of the genotyping and imputation dataset will be the same for every researcher. It is the family files that will be dynamically generated and specific to an application.

11. **What should we do now to get the data?**

Please engage with your IT teams (or similar) to ensure the necessary infrastructure is in place (storage and bandwidth) at your registered institute.

If you have not already responded to the communications sent out by the Access Team, please get in touch now to ensure an EGA account is created for you. This will be needed if you intend to download an electronic copy of the genotyping and imputed dataset.

Send an email to **access@ukbiobank.ac.uk** providing the following information:

- Application Number(s):
- First name / surname:
- Institute:
- Your Personal Institute Email address used to register with UK Biobank (to be used for EGA account 'user id'):

Some researchers may already have an EGA account and if this is the case, permissions to access the UK Biobank dataset will simply be added to your existing account – as long as the email address provided above is the same.

12. **Do all the files need to be downloaded at once or can we choose what to/not to download? Will these data be available for download at a later time or is there a time-limit beyond which this will not be possible?**

The files can be downloaded as a full dataset or via individual file downloads, where the researcher can choose what to/not to download. There are two datasets held at the EGA – the first for the genotyping data and the second for the imputation data.

We are working with the EGA to distribute these data in a short space of time. There will be the ongoing ability for researchers to download all the genotyping and imputation data from the UK Biobank **Showcase** directly.

13. **When will the data be available?**

Given we will be transferring more than 2PB of data in aggregate in a short space of time, there have been aspects that have remained difficult for us to control directly and some flexibility has been required.

The data are now available for download. Account credentials will be sent from Tues 30th May.

We would hope to be in a position to provide de-encryption keys before the end of June, but will provide further communication on this during the download period.

14. **Where are the UK Biobank datasets located within the EGA?**

UK Biobank has uploaded its genotyping and imputation datasets to the EGA under Study ID EGAS00001002399. Please be aware that the Study ID is not yet publicly visible on the EGA directory. Researchers who have requested an EGA account via UK Biobank (or asked that appropriate permissions be added to an existing account) will have been granted access to the datasets held within this study.

There are two datasets under this Study ID:

- EGAD00010001225 (imputation), and
- EGAD00010001226 (directly genotyped)

There are 67 files contained within the imputation dataset and 157 files contained within the genotyped dataset. The list of files (in tab-delimited form) comprising EGA File Accession ID, filename and file size are provided here: **imputed data** and **genotyped data**. Please see Q4 for further details on the file naming standard.

15. **How do I access / login to the EGA to download these data?**

The EGA will start to issue access accounts from Tuesday 30th May. Given the number of accounts to be created, this may take a few days. You will receive an email from the **EGA Helpdesk** with subject "*An administrator created an account for you at the European Genome-phenome Archive (EGA)*".

If you have not received any login details by Friday 2nd June please check your junk email in the first instance. If they have not been received, please contact the **EGA Helpdesk** including UKBIO in the subject title and address it for the attention of Giselle Kerry who will be our point of contact at the EGA during this period.

The email will contain a unique one-time link that will allow you to login and set your password after which you will be able to log in at **https://www.ebi.ac.uk/ega/user** using:

username: your email addressed provided to UK Biobank
password: your password set using the one-time link

When changing your password using the one-time link, you will likely receive a further email from the **EGA Helpdesk** advising "*EGA account: Notification of request to change password*" and that the account password will be activated within 24 hours.

This will be followed by a further email advising "*Account details for xxx@xxx.xxx at European Genome-phenome Archive (approved)*" and that you can now access your EGA account using your new password.

16. **Is there information available on how to configure our firewall to connect to the EGA?**

For more details on access to the EGA, please review the 'Learn about the EGA' material to be found at **https://www.ebi.ac.uk/ega/**

If using the EGA download streamer client, Java 1.7+ is required and access must be available to the Internet on standard ports 80 (http) and 443 (https). For UDT there must also be UDP port 80 open. The EGA maintains a load balancer on 'ega.ebi.ac.uk' which resolves to IP address 193.62.193.9.

17. **What is the process for downloading these data from the EGA?**

Once your EGA account credentials have been activated, you will be able to start to download the genotyping and imputation datasets. Due to the number of researchers requesting access, there are three download mirrors. Researchers will be allocated a preferred download mirror by the EGA.

The majority of researchers will be issued credentials to download these data using the EGA's own download streaming client. In certain cases, particularly for researchers located internationally, the EGA will issue credentials to use Aspera as an alternative download mechanism. Further information with links will be provided as appropriate.

18. **How do I use the EGA's own download streaming client?**

The EGA have compiled excellent documentation on how to use their download streamer and which can be found here: **https://ega-archive.org/using-ega-download-client**

Briefly, the EGA download client can be used either in interactive or command-line mode. Once logged in, issuing the command 'datasets' should return the two datasets available:

```
EGA > datasets
  EGAD00010001225
  EGAD00010001226
```

Issuing the command 'files dataset EGADnnnnnnnnnn' will display the files contained within the dataset:

```
EGA > files dataset EGAD00010001225
Files in EGAD00010001225:
  /001/ukb_bgi_chr11_v2.bgi.gz.enc.cip  91868672  EGAF00001594923  available
  /001/ukb_imp_chr11_v2.bgen.gz.enc.cip  102567821408  EGAF00001594925  available
…
```

The general process when using the EGA download streaming client is to create a download request ticket by using the 'request' command:

```
EGA > request file EGAF00001595142 password1234 request_EGAF00001595142
Requesting....
Resulting Request:
  request_EGAF00001595142       (1 new request(s)).
```

With a download request ticket created, the download can be initiated using 'download':

```
EGA > download request_EGAF00001595142
Files to download in this request: 1
Start Download Process: 5 (max:15) parallel threads
Iteration 1: 1 files.
Starting download: /002/ukb_sqc_v2.txt.gz.enc.cip  (0/1)
Download Active: 1 jobs submitted, 0 completed.
Completed download: /002/ukb_sqc_v2.txt.gz.enc.cip: true  (0/1)
Completed Download: /002/ukb_sqc_v2.txt.gz.enc.cip
Completed Download Target:   \EGA\_002_ukb_sqc_v2.txt.gz.enc.cip
Rate: xx.x MB/s
Iteration Done! (0 / 1)
Download Attempt Completed. 1 of 1 tickets downloaded successfully.
Post-Download: No outstanding Tickets
Download Iteration Completed. Retry loop for 0 unsuccessful downloads.
```

Once the file has been successfully download, the EGA transfer encryption wrapper needs to be removed using the decryptkeep command:

```
EGA > decryptkeep _002_ukb_sqc_v2.txt.gz.enc.cip password1234
Decrypting 1 file(s).
Done!
```

This will produce a *.gz.enc file (.cip encryption removed) that can subsequently be decrypted once the UK Biobank de-encryption keys are made available.

This process is repeated for each file, though it is possible to create a download request ticket for the full dataset, and this is likely the preferred route that researchers will want to take.

19. **Where do I go for further information relating to the EGA and download instructions?**

Further information on how to use your EGA account can be found at:
https://www.ebi.ac.uk/ega/about/your_EGA_account

Information on how to use the EGA download client and downloading in general can be found at **https://www.ebi.ac.uk/ega/about/your_EGA_account/download_streaming_client** together with a quick start guide that can be found at **https://ega-archive.org/downloader-quickguide**

20. **When will application-specific files be made available via the UKB Showcase?**

The UKB **Showcase** will be updated a few days after we distribute the UKB decryption key. This will allow researchers to fetch application-specific files e.g. the family file that maps an application's pseudo identifiers to the samples/ordering within the genotyping and imputation data.

21. **Which files will be available via the UKB Showcase versus the EGA?**

The following table confirms the naming convention for each file type together with whether a file is being made available via either the UKB **Showcase**, the EGA or both.

Data	UKB File Naming	EGA	Showcase*
Calls BED	ukb_cal_chrN_vZ.bed	✓	✓
Calls BIM	ukb_snp_chrN_vZ.bim	✓	✓
Calls FAM	ukbA_cal_chrN_vZ_sP.fam		✓
Marker-QC	ukb_snp_qc.txt		✓
Sample-QC	ukb_sqc_v2.txt	✓	✓
Relatedness	ukbA_rel_sP.txt		✓
Imputation BGEN	ukb_imp_chrN_vZ.bgen	✓	✓
Imputation BGI	ukb_bgi_chrN_vZ.bgi	✓	✓
Imputation MAF+info	ukb_mfi_chrN_vZ.txt	✓	✓
Imputation sample	ukbA_imp_chrN_vZ_sP.sample		✓
Haplotypes BGEN	ukb_hap_chrN_vZ.bgen		✓
Haplotypes BGI	ukb_hbg_chrN_vZ.bgi		✓
HLA Imputation	ukb_hla_vZ.txt	✓	✓
Intensity	ukb_int_chrN_vZ.bin	✓	✓
Confidences	ukb_con_chrN_vZ.txt	✓	✓
CNV log2r	ukb_l2r_chrN_vZ.txt	✓	✓
CNV baf	ukb_baf_chrN_vZ.txt	✓	✓
SNP-posterior	ukb_snp_posterior_chrN.bin		✓
Batch	ukb_snp_posterior.batch		✓

N = chromosome = 1,...,22,X,Y,XY,MT
A = application ID (integer)
Z = version of dataset, currently 2 for all files
P = number of linked samples (i.e. currently consenting participants) in linked dataset

** files will be available from Showcase once it is updated for the genotyping release, which will be a few days after the UKB decryption key is distributed*

22. **How will updated genetic phenotypes (i.e. principal components, relatedness etc) be released?**

These files will be made available as part of the UKB **Showcase** update release. Researchers will need to download these files using the ukbgene utility (which will supercede gfetch). These files are not currently available via the EGA.

23. **Will the EGA be the only route for downloading the core genotyping and imputation data?**

Once the UKB decryption key has been released and after the UKB Showcase has been updated, all genotyping and imputation files will be made public via the UKB Showcase in the usual manner.

As described above, researchers will need to fetch application-specific files via the UKB Showcase as these will not be made available via any other route. Researchers can then choose whether they download the participant/sample-independent data files from the EGA or via the UKB Showcase.

24. **How will the Haplotypes be made available?**

These files will be made available as part of the UKB Showcase update release. Researchers will need to download these files using the ukbgene utility (which will supercede gfetch) in the usual manner. These files are not currently available via the EGA.

25. **How will the CEL files be made available and when will they be released?**

These files will be made available as part of the UKB Showcase update release. Researchers will need to download these files using the ukbfetch utility in the usual manner. These files are not currently available via the EGA.

26. **When will the Chr X (and remaining sex chromosomes) be made available?**

These data are still going through final quality assurance and a release date will be communicated in the near future.

27. **Why am I seeing 'bad magic number' or 'bad decrypt' errors when decrypting files?**

The error message 'bad magic number' relates to trying to decrypt a 'bad' file and which is likely for one of two reasons:

1. Trying to use openssl and the UKB decryption key to decrypt an EGA-directly downloaded file e.g. a file with extension *.gz.enc.cip and which will fail with 'bad magic number'
2. Trying to decrypt a file that has been corrupted somewhere in the process – either as part of the EGA download or EGA decryption process.

In either case, provided there is a clear understanding of the download and encryption / decryption process and that ***md5 checksums have been checked at each step of the process*** then this issue should not occur.

Example

Using the Chr 1 call files as a worked example for clarity (EGAF00001595060):

Step 1: download the file from the EGA to provide a file with extension *.gz.enc.cip and where the 'cip' denotes that it is encrypted with an EGA encryption wrapper. The filename will be of similar format to ukb_cal_chr1_v2.bed.gz.enc.cip

Step 2: ensure the md5 checksum for this downloaded file matches the correct checksum before proceeding further i.e. 9f4c1051bdb14f66ba36a21ddf399644

Step 3: remove the EGA encryption wrapper using the EGA Download Client either by using the command line or via the interactive shell to provide a file with extension *.gz.enc and where the 'enc' denotes that it is encrypted with the UKB wrapper. The filename will be of similar format to ukb_cal_chr1_v2.bed.gz.enc

NOTE: If you downloaded the file directly using the EGA Download Client, the EGA decryption password is the one that you provided when creating the download request at the EGA, whereas if you downloaded the file using FTP or Aspera the decryption password is the one provided to you by the EGA. If you downloaded via Aspera or FTP and do not have the decryption password, please contact the **EGA Helpdesk** including UKBIO in the subject title and address it for the attention of Giselle Kerry who will be our point of contact at the EGA during this period.

Please be aware that if you supply the wrong decryption password, the decrypted file will not work and if using the command line option the original *.gz.enc.cip file will be deleted.

```
java -jar EgaDemoClient.jar -p <username> <password> -dc
ukb_cal_chr1_v2.bed.gz.enc.cip -dck <decryption_key>
```

Preferably, a safer option is to use the EGA Download Client interactive shell and make use of the decryptkeep command as explained previously in these FAQ and as per the instructions available from the EGA.

```
decryptkeep ukb_cal_chr1_v2.bed.gz.enc.cip <decryption_key>
```

Step 4: ensure the md5 checksum for this EGA-decrypted file matches the correct checksum before proceeding further i.e. 1bd4ef12c5b988cdecb74498246467db

Step 5: remove the UKB encryption wrapper using the openssl command and the UKB decryption key to provide a compressed gz output file e.g. ukb_cal_chr1_v2.bed.gz

```
openssl enc -in ukb_cal_chr1_v2.bed.gz.enc -out
ukb_cal_chr1_v2.bed.gz -d -pass file:keyfile -aes-256-cbc
```

and where keyfile contains the UKB provided decryption key - dfd64cf9d3a3ee3a792d7351bbbe8d2510f7762aded3d31e5bee757f6e3b6904

If you receive a 'bad decrypt' error at this point it is likely that your keyfile is not right for some reason and you can test this by specifying the decryption key above on the command line directly (rather than in a file).

Step 6: unzip this file to recover the original e.g. ukb_cal_chr1_v2.bed

```
gunzip < ukb_cal_chr1_v2.bed.gz > ukb_cal_chr1_v2.bed
```

Step 7: ensure the md5 checksum for this original file matches the correct checksum i.e. b46d4fefc8e5ea69de19f8e47de42b5d

and then repeat for all other files required.

A consolidated list of checksums for each file (EGA –encrypted, UKB-encrypted and original file) can be found **here** for the genotyping dataset and **here** for the imputation dataset.