

## UK Biobank - Exome Data Release FAQs July 2019

This document provides further information for researchers relating to the release of exome data for participants in UK Biobank. It has been prepared based on questions which have been received by UK Biobank's Access Team over the past four months, alongside other information which we believe will be of most relevance to researchers: these topics are covered in Sections 1 and 2.

Section 3 of this document describes an error identified in the marking of duplicate sequence reads in the exome data release for the first 50,000 UK Biobank participants. The answers below detail the duplicate read marking of data processed with the SPB pipeline, outline how this error occurred and the impact of the error, and explain the corrective action being taken and a timeline of when the corrected files will be available.

Please note these FAQs should be read in conjunction with the accompanying manuscript (<https://www.biorxiv.org/content/10.1101/572347v1>) to best interpret the data.

These FAQs will be updated periodically with the most up-to-date version available via the UK Biobank website.

### Frequently Asked Questions (FAQs):

#### Section 1: General and data access queries

1. [What data have been released?](#)
2. [How do we access and download these data?](#)
3. [How do we confirm if we have already requested/had approval for the exome dataset?](#)
4. [We do not have approval for genetic data. How do we request approval?](#)
5. [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?](#)
6. [Are these data encrypted?](#)
7. [What is the size of the available data?](#)
8. [Can we still link to a key/institute genetic dataset?](#)
9. [We would like to use exome data for a different project than the one currently registered with UK Biobank, would we need to submit a separate application for that?](#)
10. [Can you explain the number of targeted genotype samples and how many have been sequenced to date?](#)
11. [How does this exome release relate to the other assays on UK Biobank samples?](#)

#### Section 2: Experimental design and data analysis pipeline queries

12. [What sequencing technology is being used for UK Biobank WES?](#)
13. [What capture design is being used for UK Biobank WES?](#)
14. [Do the CRAM files also contain unmapped reads?](#)
15. [In what format are the variant calls available?](#)
16. [Where can I obtain a multi-sample VCF \(pVCF\) file for the UK Biobank 50,000 WES dataset?](#)
17. [Which VCF versions are available? Are the VCFs annotated or not? If so, which annotation was used?](#)
18. [Are the variants already functionally annotated by a tool like annovar or equivalent?](#)
19. [For the GATK-based pipeline, is VSQR already performed or still necessary to perform ourselves?](#)
20. [We have downloaded the joint call set PLINK files and have noticed that in the fam file there are samples with "missing" ID values which are negative. Should these individuals be excluded?](#)
21. [As CRAM files are compressed, it's better to have access to the original BAM files - is there any possibility to get these for each WES individual?](#)
22. [We have downloaded .bim \(PLINK format\) file, but associated .bed/.fam files are not available for](#)

[download.](#)

23. [How were variants called? What cutoffs in read depth and quality were used to make these calls? Which reference genome build alignment was used? Are there accessory variant-level QC metrics such as average read depth and allele frequency?](#)
24. [What versions of the human reference genome were used to map the UK Biobank 50,000 WES data?](#)
25. [What are the best practices for analysing the UK Biobank 50,000 WES variant calls?](#)
26. [Is it possible to download slices of the genome at a time? The VCF file is very large and I initially want to query regions of the genome. Is it possible to download the data for specific regions?](#)
27. [Are there any plans to release phased versions \(haplotypes\) of the new UK Biobank exome data as well?](#)

### **Section 3: Duplicate read marking issue queries**

28. [What does the duplicate read marking issue relate to?](#)
29. [What are duplicate reads?](#)
30. [What is duplicate read marking?](#)
31. [Why is duplicate read marking required?](#)
32. [What went wrong with duplicate read marking in the UK Biobank 50,000 WES data?](#)
33. [What is the impact of this error?](#)
34. [What data are affected by the error in marking duplicate reads?](#)
35. [What action is being taken to correct the error in marking duplicate reads?](#)
36. [How and when will the corrected data files be made available to researchers?](#)

## Section 1: General and data access queries

### 1. What data have been released?

The released data are:

- Project level variant call data – a joint call set file in PLINK (BED/BIM) format (~100 GB);
- Sample level variant call data – gVCF files for 49,960 exomes (~5 TB);
- Sample level aligned sequence data – CRAM files for 49,960 exomes (~50 TB).

All these data are available in 2 forms: one generated using Regeneron's own pipeline (SPB) (<https://www.biorxiv.org/content/10.1101/572347v1>), and one generated using a Functionally Equivalent (FE) pipeline (<https://www.ncbi.nlm.nih.gov/pubmed/30279509>).

### 2. How do we access and download these data?

Researchers named on approved applications with permission to access exome data will be able to download the joint call set PLINK data via the ukbgene utility:

<http://biobank.ctsu.ox.ac.uk/showcase/download.cgi?id=665&ty=ut>.

Instructions for the specific types are given at:

<http://biobank.ctsu.ox.ac.uk/showcase/field.cgi?id=23170> and

<http://biobank.ctsu.ox.ac.uk/showcase/field.cgi?id=23160>.

You will also need a project specific mapping file (fam file) to link these data to the non-genetic phenotypes: instructions for downloading the fam files are given in the Notes tabs on the pages for Data Fields [23160](#) and [23170](#).

You will need the ukbgene program to download the combined joint call set PLINK-format data and the ukbfetch program to download the individual VCF and CRAM files.

### 3. How do we confirm if we have already requested/had approval for the exome dataset?

You can check which genetic data has been approved per project by checking Annex A of the Material Transfer Agreement (and any later updated Annex A).

If you do not have approval for the exome dataset, see Question 4 below.

### 4. We do not have approval for genetic data. How do we request approval?

If your project has not been approved for genetics-related fields, you can submit a Change Request to extend the scope of your research application adding justification for why you would like to conduct further analyses.

### 5. 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, so you can choose what to download. The PLINK data is a multi-sample file for all 50,000 participants and is available in two flavours – one generated using Regeneron's SPB pipeline, and one generated using the FE pipeline; for details see <http://biobank.ctsu.ox.ac.uk/showcase/label.cgi?id=170>. Individual VCFs and CRAMs are available for each sample and researchers can choose to download (or not) as they wish using the standard tools. There is no time limit other than we are releasing them in a phased manner.

**6. Are these data encrypted?**

No, there are no encryption wrappers with these data.

**7. What is the size of the available data?**

	SPB	FE
PLINK	~100 GB	~100 GB
gVCF	~5 TB	~5 TB
CRAM	~50 TB	~50 TB

**8. Can we still link to a key/institute genetic dataset?**

Yes. We will continue to support the ability for institutes to hold a key genetic dataset that is shared between multiple applications, so that an Institute does not need to hold multiple copies.

The UK Biobank Showcase is being used to hold and distribute the exome data. Researchers will need to create the fam files that map the ordering of the exome data to their pseudonymised IDs using Showcase; instructions for this are available for both the SPB and FE datasets (see [FAQ #2](#)). The ordering of the joint call set PLINK data will be the same for every researcher. It is the family files that will be dynamically generated and specific to an application.

**9. We would like to use exome data for a different project than the one currently registered with UK Biobank; would we need to submit a separate application for that?**

Yes, each use of UK Biobank data must be approved through an application and have its own Material Transfer Agreement in place.

**10. Can you explain the number of targeted genotype samples and how many have been sequenced to date?**

There are 3 sources of genomic data that are (or will be) included in the UK Biobank Resource:

- A. Genome-wide genotype data** for all 500,000 UK Biobank participants generated using the Affymetrix UK BiLEVE Axiom array (initial 50,000 participants) and the Affymetrix UK Biobank Axiom Array (remaining 450,000 participants), along with an imputed dataset of over 90 million SNPs. Genotype data for the full cohort was released in July 2017.
- B. Whole Exome Sequencing (WES) data** will be generated and released for the full UK Biobank cohort. A WES dataset for the first 50,000 UK Biobank participants was released in March 2019; the next tranche of data for a further 100,000 participants is due to be released in Q1 2020; the data for the remaining 350,000 participants will be released in 2021.
- C. Whole Genome Sequencing (WGS) data** will be generated and released for the full UK Biobank cohort. A Vanguard WGS project is currently underway to perform WGS on 50,000 UK Biobank samples. There is currently no timeline for WGS data availability.

For further details please see our website which will be updated periodically: <https://www.ukbiobank.ac.uk/scientists-3/genetic-data/>

**11. How does this exome release relate to the other assays on UK Biobank samples?**

The sample set prioritises individuals with MRI imaging data (from the UK Biobank Imaging Study), enhanced baseline measurements, hospital episode statistics (HES), and/or linked primary care records. One disease area was selected for enrichment, including individuals with admission to hospital with a primary diagnosis of asthma.

## Section 2: Experimental design and data analysis pipeline queries

### 12. What sequencing technology is being used for UK Biobank WES?

Exomes were captured using the IDT xGen Exome Research Panel v1.0 including supplemental probes. Multiplexed samples were sequenced with dual-indexed 75 x 75 bp paired-end reads on the Illumina NovaSeq 6000 platform using S2 flow cells.

### 13. What capture design is being used for UK Biobank WES?

The UK Biobank whole-exome sequencing basic design targets 39 Mbp of the human genome. The GRCh38 coordinates of the targeted regions are provided in the bed file "xgen\_plus\_spikein.b38.bed" which is available for download here: <http://biobank.ctsu.ox.ac.uk/showcase/refer.cgi?id=3801>. Please note that the UK Biobank 50,000 variant call sets include variants in both the target regions and 100 bp flanking regions upstream and downstream of each capture target. While these flanking-region calls may be informative for certain analyses, only the targeted capture regions are required to meet all sequencing quality standards such as unique read coverage. All variants in both the flanking and target regions are subject to the same variant quality control metrics (e.g. depth and allele balance) as described in the manuscript (<https://www.biorxiv.org/content/10.1101/572347v1>).

### 14. Do the CRAM files also contain unmapped reads?

Yes. Original sample FASTQs are losslessly recreatable (up to read ordering) from the provided SPB CRAMs, which contain every read regardless of whether it maps and all original quality scores. Please note that CRAMs should be name sorted or randomized prior to extracting a FASTQ to ensure uncorrelated read sets for subsequent parallelized mapping (e.g. BWA).

### 15. In what format are the variant calls available?

Sample level gVCFs and joint call set level PLINK files are available (released in March 2019).

### 16. Where can I obtain a multi-sample VCF (pVCF) file for the UK Biobank 50,000 WES dataset?

The pVCFs for the UK Biobank 50,000 WES data sets (SBP and FE) will be made available along with the CRAMs, gVCFs and PLINK files in September 2019. All files will be derived from reprocessed data that addresses the duplicate marking issue (to be described in section 3 of this document).

### 17. Which VCF versions are available? Are the VCFs annotated or not? If so, which annotation was used?

VCFs are VCF 4.2 and are not annotated.

### 18. Are the variants already functionally annotated by a tool like annovar or equivalent?

The variants are not annotated.

### 19. For the GATK-based pipeline, is VSQR already performed or still necessary to perform ourselves?

VQSR is not performed. The FE pipeline uses GATK for single-sample variant calling (i.e. CRAM to gVCF) and GLnexus (<https://www.biorxiv.org/content/10.1101/343970v1>) for population aggregation (gVCFs to joint call set PLINK).

### 20. We have downloaded the joint call set PLINK files and have noticed that in the fam file there are samples with "missing" ID values which are negative. Should these individuals be excluded?

Yes, a negative person ID in the fam file means that the corresponding participant has withdrawn consent and should therefore be excluded.

### 21. As CRAM files are compressed, it's better to have access to the original BAM files - is there any

**possibility to get these for each WES individual?**

No, as they are not required. The CRAM files have utilised lossless compression and the original FASTQ can be fully reconstructed (including all instrument generated quality scores) from the SPB CRAMs.

**22. We have downloaded .bim (PLINK format) file, but associated .bed/.fam files are not available for download.**

The bim files are an openly downloadable file from the UK Biobank website as they do not contain participant data. The bed file (which is the PLINK file of call data itself) is downloadable using the `ukbgene` utility, as is the .fam file using the `-m` flag.

**23. How were variants called? What cutoffs in read depth and quality were used to make these calls? Which reference genome build alignment was used? Are there accessory variant-level QC metrics such as average read depth and allele frequency?**

Please read the summary methods material within the paper on BioRxiv titled ‘Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank’: <https://www.biorxiv.org/content/10.1101/572347v1>

**24. What versions of the human reference genome were used to map the UK Biobank 50,000 WES data?**

The UK Biobank WES 50,000 release includes two sets of CRAM and gVCF files, one for each of the SPB and FE pipelines (described here: <http://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=170>). The SPB pipeline maps to a “flat” version of the GRCh38 human reference genome, and the FE pipeline maps to a full GRCh38 reference version including all alternative contigs. The commands listed below detail how to download and process each reference version from public resources for CRAM decompression.

**SPB:**

wget

[ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA\\_000001405.15\\_GRCh38/seqs\\_for\\_alignment\\_pipelines.ucsc\\_ids/GCA\\_000001405.15\\_GRCh38\\_no\\_alt\\_plus\\_hs38d1\\_analysis\\_set.fna.gz](ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_plus_hs38d1_analysis_set.fna.gz)

```
zcat GCA_000001405.15_GRCh38_no_alt_plus_hs38d1_analysis_set.fna.gz | sed 's/^>chr/>/g' > GCA_000001405.15_GRCh38_no_alt_plus_hs38d1_analysis_set.fna
```

samtools faidx

```
GCA_000001405.15_GRCh38_no_alt_plus_hs38d1_analysis_set.fna
```

**FE:**

wget [ftp://ftp-](ftp://ftp-trace.ncbi.nlm.nih.gov/1000genomes/ftp/technical/reference/GRCh38_reference_genome/GRCh38_full_analysis_set_plus_decoy_hla.fa)

[trace.ncbi.nlm.nih.gov/1000genomes/ftp/technical/reference/GRCh38\\_reference\\_genome/GRCh38\\_full\\_analysis\\_set\\_plus\\_decoy\\_hla.fa](trace.ncbi.nlm.nih.gov/1000genomes/ftp/technical/reference/GRCh38_reference_genome/GRCh38_full_analysis_set_plus_decoy_hla.fa)

```
sed -i 's/^>chr/>/g' GRCh38_full_analysis_set_plus_decoy_hla.fa
```

```
samtools faidx GRCh38_full_analysis_set_plus_decoy_hla.fa
```

Pre-processed versions of each reference are available for download via the following links.

**SPB:**

<http://biobank.ctsu.ox.ac.uk/showcase/refer.cgi?id=838>

FE:

<http://biobank.ctsu.ox.ac.uk/showcase/refer.cgi?id=1000>

## 25. What are the best practices for analysing UKB 50,000 WES variant calls?

The UK Biobank 50,000 WES data release includes multiple data sets and file formats to facilitate a broad range of downstream analyses. The following details should inform researchers' decisions on how to best analyse the sample set.

- a. Both the SPB and FE variant file sets (gVCFs, PLINKs, pVCFs) report variants from both the target capture regions (see [FAQ #13](#)) and 100 bp flanking regions. All sequencing quality control metrics are assessed only from the target regions, ensuring consistent quality and variant calling over these bases. Downstream variant analyses should be region-aware and distinguish between calls in the target regions and those in the buffer regions.
- b. The SPB pipeline maps the raw reads to a "flat" version of the GRCh38 reference genome that does not include alternative contigs. The FE pipeline maps the reads to the "full" GRCh38 reference genome, including all alternative contigs (see [FAQ #24](#) for more details). Given that reads will map differently to each reference genome, the variants from each pipeline are expected to differ, especially in those regions of the "flat" reference from which the alternative contigs are derived. A bed file providing coordinates for the ~0.1% of the target regions from which at least one alternative contig is derived is available for download via the following link: <http://biobank.ctsu.ox.ac.uk/showcase/refer.cgi?id=3802>
- c. The SPB pipeline generates CRAM files that contain the original quality scores from the raw read data (FASTQs). Thus, original FASTQs can be extracted from the SPB CRAMs. To ensure unbiased mapping, all CRAMs should be name sorted prior to FASTQ extraction. The FE CRAMs contain GATK Recalibrated base qualities, as prescribed by the Functionally Equivalent pipeline specification (<https://www.ncbi.nlm.nih.gov/pubmed/30279509>).
- d. Initial findings from the UKB 50,000 WES data as reported in the manuscript (<https://www.biorxiv.org/content/10.1101/572347v1>) were derived from the SPB data set after additional filters were applied. These additional filters are described in the methods section of the manuscript.

## 26. Is it possible to download slices of the genome at a time? The VCF file is very large and I initially want to query regions of the genome. Is it possible to download the data for specific regions?

No, this functionality is not currently available.

## 27. Are there any plans to release phased versions (haplotypes) of the new UK Biobank exome data as well?

No plans at present.

## Section 3: Duplicate read marking queries

### 28. What does the duplicate read marking issue relate to?

In July 2019 an issue was identified within the SPB pipeline exome data, in which duplicate sequence reads were not correctly marked.

### 29. What are duplicate reads?

Duplicate reads are multiple reads (or read pairs) that originate from the same template sequence during library preparation of a given sample. They arise upstream of sequencing from processes such as PCR. Duplicate reads are easily identified as any set of reads (or read pairs) with the same reference alignment, such that each read in a given pair has the same start and end positions as the corresponding read in another pair.

### 30. What is duplicate read marking?

Duplicate read marking is a standard step in the primary analysis of DNA sequencing data. After mapping sequence reads to a reference genome, duplicate reads are “marked” using software such as [Picard](#). Duplicate read marking does not remove any sequence reads from the file but does distinguish duplicate reads from non-duplicates, leaving one representative read pair from each duplicate set unmarked.

### 31. Why is duplicate read marking required?

Duplicate marking is required to identify upstream sequencing duplicates that can affect variant calling accuracy. If duplicates are not marked, genomic regions can be over-represented in a sequencing dataset. Duplicate reads are generally ignored by downstream analyses as they do not represent independent observations of the underlying genomic sequence and may introduce bias into a sequencing dataset.

### 32. What went wrong with duplicate read marking in the UK Biobank 50,000 WES data?

The SPB mapping protocol applied to the UK Biobank 50,000 WES reads was previously designed for single-library loadings to individual lanes on the Illumina HiSeq platform. In this HiSeq protocol, each individual flow cell lane contained one WES library derived from a given sample, so marking of duplicate reads within each flow cell lane (per-lane duplicate marking) would identify all duplicate reads from each library.

All UKB 50,000 WES samples were sequenced on the NovaSeq platform, which distributes the sample library across multiple lanes of the flow cell (two lanes for the S2 flow cell), but the SPB mapping protocol was not updated for this platform. While duplicate reads were correctly marked within each flow cell lane, the inter-lane duplicate marking step required for processing NovaSeq data was not added to the SPB mapping protocol. Therefore, for the 50,000 WES data, all duplicates within each flow cell lane have been correctly marked, but duplicates across lanes (maximum of one duplicate per unique read pair) have not been marked.

### 33. What is the impact of this error?

This undermarking of duplicate reads causes the unique-read coverage reported for each sample to be inflated and can create variant errors. False positive variant calls can arise when unmarked duplicate reads carry a variant allele, and false negative calls can arise when the unmarked duplicates carry the reference allele.

**34. What data are affected by the error in marking duplicate reads?**

The issue is limited to the exome data that have been processed using the SPB pipeline; those data produced using the FE pipeline are not affected.

**35. What action is being taken to correct the error in marking duplicate reads?**

The Regeneron Genetics Center is reprocessing the data from the initial UK Biobank 50,000 WES cohort release using a corrected SPB pipeline to generate corrected data files (CRAMs, gVCFs and PLINKs).

**36. How and when will the corrected data files be made available to researchers?**

Corrected data files (CRAMs, gVCFs and PLINKs) will be made available via the UK Biobank Data Showcase. It is expected that this re-release will occur in September 2019. We will send more details about the data re-release, which will also include release of pVCF data files, and timelines in due course.

For residual questions not answered, please use the UKB-GENETICS mailing list. This has been created for researchers who wish to share ideas/queries about the UK Biobank genetic data and can be accessed here: <https://jiscmail.ac.uk/cgi-bin/webadmin?A0=ukb-genetics>.