UK Biobank - Exome Data Release FAQs / December 2020

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, alongside other information which we believe will be of most relevance to researchers.

The document has been substantially revised following the release of 200k exomes in October 2020. Specific FAQs relating to the 50k release made in March 2019 have been retained as an Annex. The 50k exome release is still available for download however is now deprecated and will be removed from further access in due course.

These FAQs will be updated periodically; the most up-to-date version will be available on the UKB website.

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?
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. Which VCF versions are available? Are the VCFs annotated or not? If so, which annotation was used?
17. Are the variants already functionally annotated by a tool like annovar or equivalent?
18. 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?
19. 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?
20. We have downloaded .bim (PLINK format) file, but associated .bed/.fam files are not available.
21. How were variants called? Are there variant-level QC metrics such as read depth and AF?
22. What versions of the human reference genome were used to map the UK Biobank 200k WES data?
23. What are the best practices for analysing the UK Biobank 200k WES variant calls?
24. 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?
25. Are there any plans to release phased versions (haplotypes) of the new UK Biobank exome data as well?

Section 1: General and data access queries

1. What data have been released?

The released data are:

- Joint genotype data – multi-sample PLINK for all 200k WES samples (~1TB);
- Joint genotype data – multi-sample pVCF for all 200k WES samples (~7TB);
- Sample-level variant call data – gVCF and index for each 200k WES sample (~8 TB); and
- Sample-level aligned sequence data – CRAM and index each 200k WES sample (~175 TB).

The data being made available have been processed using a pipeline based on the Functional Equivalence specification that retains the original quality scores (referred to as the OQFE protocol). For more details, please see Table 1 and the UKB 200k preprint: https://www.medrxiv.org/content/10.1101/2020.11.02.20222232v1.

A project level variant call dataset, derived using GLnexus, is being made available in the form of a joint call set file in PLINK (bed/bim) format and also a pVCF file that follows the variant call file specification.

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 data (both PLINK and pVCF file formats when available) via the "gfetch" utility available at: http://biobank.ctsu.ox.ac.uk/showcase/download.cgi?id=600&ty=ut.

For instruction on using the new gfetch utility, researchers should refer to: https://biobank.ndph.ox.ac.uk/ukb/refer.cgi?id=668.

For the PLINK files, 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 instructions above.

Researchers will need to create a basket to request the individual gVCF and CRAM files (fields 23151 to 23154), after the point they become publicly available, before downloading via the "ukbfetch" utility as described at: http://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=644.

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

If you have previously had approval for exome data then you should be able to download the PLINK and pVCF formats and additional participants immediately (though not the individual gVCF and CRAM files) once any embargoes have lifted. If you do not have approval then the gfetch utility will give you an error message explaining this and you should add them to a basket.

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 application, with 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?

The files can be downloaded as a full dataset or via individual file downloads, so you can choose what to download. The PLINK and pVCF data are multi-sample files for all 200k participants generated using the OQFE protocol; for details see http://biobank.ctsu.ox.ac.uk/showcase/label.cgi?id=170. Individual gVCFs 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?**

	OQFE
PLINK	~1 TB
pVCF	~7 TB
gVCF	~8 TB
CRAM	~175 TB

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

   We will consider requests 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 Data 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. The ordering of the joint call set PLINK data will be the same for every researcher. It is the FAM files that will be dynamically generated and specific to an application.

   The pVCF files are being supplied with Application-specific identifiers already bound into them internally.

   Recognising these are sizeable datasets, UK Biobank will be making available a Research Analysis Platform during 2021 to enable the ability to analyse these data in "situ" as an alternative to researchers needing to maintain their own substantial infrastructure. This approach will democratise access and will help to enable the widest possible use of UK Biobank data by approved researchers; further details will be provided in due course.

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; this 200k release represents the next tranche of data; it is anticipated that the data for the remaining 300,000 participants will be released in late 2021 / early 2022.

    C. **Whole Genome Sequencing (WGS) data** will be generated and released for the full UK Biobank cohort. A project is currently underway to perform WGS on all UK Biobank samples. It is anticipated that the first release of WGS samples will be for ~200,000 participants in Q3 2021.

For further details, please see the genetics section on 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 initial 50k sample set prioritised 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. The subsequent samples beyond the initial 50k participants have been randomly selected.

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 (UKB 50k) and S4 (additional samples in UKB 200k) flow cells. Please note that the samples in the UKB 50k release were sequenced with a different IDT v1.0 oligo lot than were the subsequent ~150k samples; all remaining UKB WES samples were processed with the 150K oligo lot.

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 is available for download via the Data Showcase (http://biobank.ndph.ox.ac.uk/ukb/refer.cgi?id=3803). Please note that the UK Biobank 200k 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 processing as described below.

14. Do the CRAM files also contain unmapped reads?

Yes. Original sample FASTQs are losslessly re-creatable (up to read ordering) from the provided OQFE 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 are available and a joint-genotyped set as both PLINK and pVCF files. The dataset is unfiltered dataset to maximize a broad range of analyses and researchers will need to carefully consider an appropriate filtering strategy based on their intended analysis.

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

gVCFs are VCF v4.2 and are not annotated.

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

The variants are not annotated.

18. **We have downloaded the joint call set PLINK files and have noticed that in the FAM file there are samples with "missing" ID values. 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.

19. **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 OQFE CRAMs.

20. **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 gfetch utility, as is the .fam file using the –m flag.

21. **How were variants called? Are there accessory variant-level QC metrics such as average read depth and allele frequency?**

    Single-sample variants were called from OQFE CRAMs with DeepVariant 0.0.10 employing a retrained model and are provided as single-sample gVCFs. All gVCFs were aggregated with GLnexus 1.2.6 using the default joint-genotyping parameters for DeepVariant referenced in the manuscript below.
    https://www.biorxiv.org/content/10.1101/2020.02.10.942086v2.

    To ensure that the UKB 200k data support a broad range of analyses, no variant- or sample-level filters were pre-applied to the pVCF or PLINK files. The publically released pVCF is the direct output of GLnexus, from which the PLINK files are generated. The pVCF contains allele-read depths and genotype qualities for all genotypes and to which analysis-specific filters can be applied. Examples of such filtering are described in the UKB 200k preprint:
    https://www.medrxiv.org/content/10.1101/2020.11.02.20222232v1.

22. **What versions of the human reference genome were used to map the UK Biobank 200k WES data?**

    The UK Biobank WES 200k release includes CRAM and gVCF files processed using the OQFE protocol described in Table 1. The OQFE protocol maps to a full GRCh38 reference version including all alternative contigs in an alt-aware manner. The full GRCh38 reference is linked below and contained within the OQFE docker (https://hub.docker.com/r/dnanexus/oqfe), which was used to process all 200k WES samples.

    GRCh38 Reference Genome files:
    ftp://ftp.ncbi.nlm.nih.gov/1000genomes/ftp/technical/reference/GRCh38_reference_genome/

23. **What are the best practices for analysing UKB 200k WES variant calls?**

    The UK Biobank 200k 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. The OQFE variant file sets (gVCFs, PLINKs, pVCFs) report variants from both the target capture regions 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 OQFE protocol generates CRAM files that contain the original quality scores from the raw read data (FASTQs). Thus, original FASTQs can be extracted from the OQFE CRAMs. To ensure unbiased mapping, all CRAMs should be name sorted prior to FASTQ extraction.
c. Initial findings from the UKB 200,000 WES data as reported in the manuscript (https://www.medrxiv.org/content/10.1101/2020.11.02.20222232v1) were derived from the OQFE data set after additional filters were applied. These additional filters are described in the methods section of the manuscript.
d. As noted previously, the samples in the initial 50k release were sequenced on S2 flow cells and with a different IDT v1.0 oligo lot than were the remaining samples in the 200k release, which were sequenced on S4 flow cells. Inclusion of this information as a covariate in downstream analyses is recommended

24. 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.

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

No plans at present.

Table 1: OQFE protocol

FE Requirements	OQFE protocol https://hub.docker.com/r/dnanexus/oqfe			
	Program	**Version**	**Command Options**	**OQFE Update Notes**
align reads: GRCh38DH with .alt file, BWA mem v0.7.15 -Y -K 100000000	bwa mem	0.7.17	-K 100000000 -Y	The UKB 50k FE data did not include the .alt file. The UKB 200k OQFE samples were mapped with this file present.
Retain the minimal set of tags (RG, MQ, MC and SA)	samblaster	0.1.24	--addMateTags -a	
	sambamba sort	0.6.4	-n	
	sambamba merge	0.6.4		
mark duplicates: Picard v2.4.1 or above	picard MarkDuplicates	2.21.2	ASSUME_SORT_ORDER= queryname	Resolves a known issue concerning which reads in a duplicate set are marked as a duplicate, which can affect the number of supplementary duplicates.
	sambamba sort	0.6.4		
BQSR				Excluded in OQFE
apply BQSR: 4-bin				Excluded in OQFE
convert to CRAM: PG records; RG: PL, PU, SM, LB; tags: RG, MQ, MC, SA, original query names	samtools view	1.9	-C	

Annex - Frequently Asked Questions (FAQs) relating to the previous 50k exome release in March 2019

This Annex provides FAQs specific to the 50k release in March 2019 relating to the description of an error (communicated to researchers in August 2019) identified in the marking of duplicate sequence reads in the SPB version of the exome data release for the first 50,000 UK Biobank participants. It also details a further error (communicated to researchers in December 2019) identified in the mapping of reads to alternative contig regions in the FE version of exome data release for the first 50,000 UK Biobank participants.

Both these errors have been corrected as part of making available the 200k exomes and the information provided here is solely for reference for researchers who may have worked with the 50k exome dataset.

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.

Section 3: Duplicate read marking issue

26. What does the duplicate read marking issue relate to?
27. What are duplicate reads?
28. What is duplicate read marking?
29. Why is duplicate read marking required?
30. What went wrong with duplicate read marking in the UK Biobank 50,000 WES data?
31. What is the impact of this error?
32. What data are affected by the error in marking duplicate reads?
33. What action is being taken to correct the error in marking duplicate reads?
34. How and when will the corrected data files be made available to researchers?
35. Will UK Biobank still provide access to the original SPB dataset (even though these issues have been identified)?

Section 4: Read mapping to alternative contigs issue

36. What does the read mapping to alternative contigs issue relate to?
37. What is alt-aware read mapping?
38. What went wrong with alt-aware read mapping in the UK Biobank 50,000 WES data?
39. What is the impact of this error?
40. What data are affected by the error in read mapping to alternative contigs?
41. What action is being taken to correct the error in mapping reads to alternative contigs?
42. How and when will the corrected data files be made available to researchers?

Section 3: Duplicate read marking issue

26. What does the duplicate read marking issue relate to?

In July 2019 an issue was identified within the 50k SPB pipeline exome data, in which duplicate sequence reads were not correctly marked.

27. 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.

28. 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.

29. 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.

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

The SPB mapping protocol applied to the UK Biobank 50k 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.

31. What is the impact of this error?

This under-marking 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.

32. 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.

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

The Regeneron Genetics Center has reprocessed the data from the initial UK Biobank 50k WES cohort release using a corrected SPB pipeline to generate corrected data files (CRAMs and gVCFs).

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

UK Biobank released a corrected 50k SPB dataset via the Data Showcase in February 2020 (see Category 170, data-fields 23176 – 23179).

35. Will UK Biobank still provide access to the original SPB dataset (even though these issues have been identified)?

The original 50k SPB exome dataset has not been removed from the resource, however it has been withdrawn from general availability and further download due to the issues identified. For researchers who have published papers using these data, there is a mechanism in place where UK Biobank will make these data available upon request (for the purposes of replication studies and similar).

Section 4: Read mapping to alternative contigs issue

36. What does the read mapping to alternative contigs issue relate to?

An issue was identified (and communicated to researchers in December 2019) with the 50k exome dataset generated by the FE pipeline, as it has been incorrectly mapped in a non-alt aware manner.

37. What is alt-aware read mapping?

As specified by the Functionally Equivalent protocol (PMID: 30279509), FE crams are mapped with BWA to the full GRCh38 reference genome, which includes alternative contigs. These alternative contigs are derived from regions of the primary assembly (i.e. autosomal and sex chromosomes) and described in the ".alt" file included with the reference genome. The BWA mapping algorithm uses this reference alt file to accurately map reads to both the primary assembly and the alternative contigs (https://github.com/lh3/bwa/blob/master/README-alt.md). Specifically, inclusion of the reference alt file ensures that map quality scores are not penalized for a read mapping to both the primary assembly and a known alternative contig. The FE specification requires alt-aware mapping via inclusion of the reference alt file.

38. What went wrong with alt-aware read mapping in the UK Biobank 50,000 WES FE data?

The UKB 50k WES FE CRAMs were incorrectly mapped to a version of the GRCh38 reference that does not include the reference alt file. Reads that map to both the primary assembly and an alternative contig will have deflated and often zero map qualities. Low-map-quality reads are generally ignored by variant callers and thus this error can result in an under-calling of variants in affected regions. Thus, all UKB 50k FE WES data (CRAMs, gVCFs, and PLINKs) are affected.

39. What is the impact of this error?

The capture design for the UKB WES data comprises 204,829 targets (39.20 MBp), 6,784 of which (1.27 MBp) overlap the alt-derived regions described in the reference alt file. An additional 770 targets (0.26 MBp) were observed to have changes in either the number of reads mapped to the target or the average read-mapping quality in test samples when compared between alt-aware and non-alt-aware mappings.

40. **What data are affected by the error in read mapping to alternative contigs?**
    To facilitate existing and continued analysis of the existing FE WES UKB data while we work to provide the corrected data set, a BED file has been provided describing WES target regions impacted by the lack of alt-aware mapping (xgen_plus_spikein_b38_alt_affected.bed). This annotated BED file contains these 7,554 targets affected by the non-alt-aware mapping. It is recommended that these regions be excluded from any analysis of the FE data and that all researchers consider how this mapping error might impact any results derived from the FE data.

41. **What action is being taken to correct the error in mapping reads to alternative contigs?**
    The Regeneron Genetics Center has reprocessed the data as part of the preparation for 200k data release and the new dataset has been made available from October 2020 (see start of document).

42. **How and when will the corrected data files be made available to researchers?**
    UK Biobank's position has been that rather than try to re-issue the 50k dataset, it would be more beneficial for researchers to have UK Biobank concentrate its efforts on the release of the first 200k exomes. The 200k dataset is now available and incorporates a re-release of the corrected 50k dataset.

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.