UK Biobank 500k Whole Genome Sequencing Release FAQs

November 2023

This document provides guidance to researchers using the whole genome sequencing (WGS) data released on 30th November 2023 from half a million UK Biobank volunteers. The guidance has been prepared based on: questions from researchers who used the interim release of 200,000 participants' genomic data; researchers anticipating the 500k WGS release; and new users of the UK Biobank Research Analysis Platform (UKB-RAP).

Access to all sequencing data is available via the UKB-RAP cloud platform (https://ukbiobank.dnanexus.com). Researchers who have an existing Tier 3 application, or a reduced fee Student or Low- or Middle-income country (LMIC) application, do not need to apply to gain access to the data on release. New researchers, or those on a different tier of data access, should follow the guidance here: https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access or contact the Access team (access@ukbiobank.ac.uk).

Section 1: General and data access queries

1. What data have been released?

Two versions of the data have been released, one produced using BWA-MEM/GATK pipelines, as used for the initial 200,000 participant dataset, and a second dataset produced using Illumina DRAGEN v3.7.8.

For the BWA-MEM/GATK data, the released data consists of:

- Individual level CRAM and CRAM index files (Field 23372)
- Individual level gVCF and gVCF index files (Field 23370)
- Joint variant called data produced using GraphTyper2 (Field 23374)
- Genotype concordance metrics (Fields 23378, 23379, 23380, and 23381)
- Measures of sample contamination (Fields 23377, 23383, and 23384)
- Base quality score recalibration (BQSR) (Field 23376)
- Supplementary files produced during data quality control (Field 23382)

For the DRAGEN data, the released data consists of:

- Individual level CRAM and CRAM index files (Field 24048)
- Individual level gVCF and gVCF index files (Field 24051)
- Individual level VCF and VCF index files (Field 24053)
- Joint variant called data produced using DRAGEN (Field 24310)
- Individual measures of copy number variation (CNV) (Fields 24056 and 24058)
- Identified short tandem repeats (STRs) (Fields 24062 and 24064)
- Structural variant data (SV) (Fields 24059 and 24061)
- Genotype calls for CYP2D6 (Field 24065)
- Supplementary and diagnostic information (Fields 24050 and 24055)

Additionally, tabular data containing quality control metrics from the sequencing process are available in Category 187.

2. How do we access these data?

The whole genome sequencing data (together with exome data and all future genetic datasets) will only be accessible via the UK Biobank Research Analysis Platform (UKB-RAP). The UKB-RAP is accessible via https://ukbiobank.dnanexus.com/.

The UKB-RAP user guide can be accessed here. The user guide, which includes a video demonstrating key platform functionality, will continue to be updated. Frequently Asked Questions can be found here, and a community forum is also available to provide additional help and support.

On sign-up to the platform, you will receive £40 credit (sufficient, for example, to run approximately 100 hours of analyses, including example Genome Wide Association Studies (GWAS) and Polygenic Risk Score (PRS) analyses using genotype data) towards the cost of any compute or data storage used. Once the free credit has been consumed, researchers will need to provide billing details to perform subsequent analyses. (Note: the core UK Biobank dataset is held at no cost to the researcher, and it is only compute or data storage used in support of analyses that are chargeable.)

To access the WGS data, you will need to have entered into a new Material Transfer Agreement (MTA) with UK Biobank together with payment of the applicable access application fee. Please see UK Biobank website FAQs for details https://www.ukbiobank.ac.uk/enable-your-research/costs/transitional-arrangements-and-faqs. If you require further information, please contact the UK Biobank's Access Management Team.

3. How do we confirm if we already have access to this data?

Researchers who currently have an approved application at Tier 3, or a reduced fee Student or LMIC application on the new MTA, can access the new WGS data. There is no requirement to submit an additional data request for researchers who are on these tiers.

4. Can these data be accessed downloaded through Data Showcase?

The whole genome sequencing data can only be used within the UKB-RAP, and researchers are not permitted to download the data either through UKB-RAP or from Data Showcase.

5. Can I still access the 200k participant whole genome sequencing data?

The individual-level data, such as CRAM and VCF files, from the first 200k participant WGS release has been merged into the enduring 500k release fields. The joint variant calls and phased datasets produced for the interim 200k release are still accessible to researchers in Category 271 and on UKB-RAP. These fields will be deprecated in due course, but UK Biobank will notify researchers in advance.

6. Will PLINK or BGEN versions be released?

PLINK 2.0 and BGEN versions for both the DRAGEN and GraphTyper joint variant calls are planned to be released in 2024.

Section 2: Queries related to use of the UKB-RAP

1. How can I follow the status regarding platform maintenance?

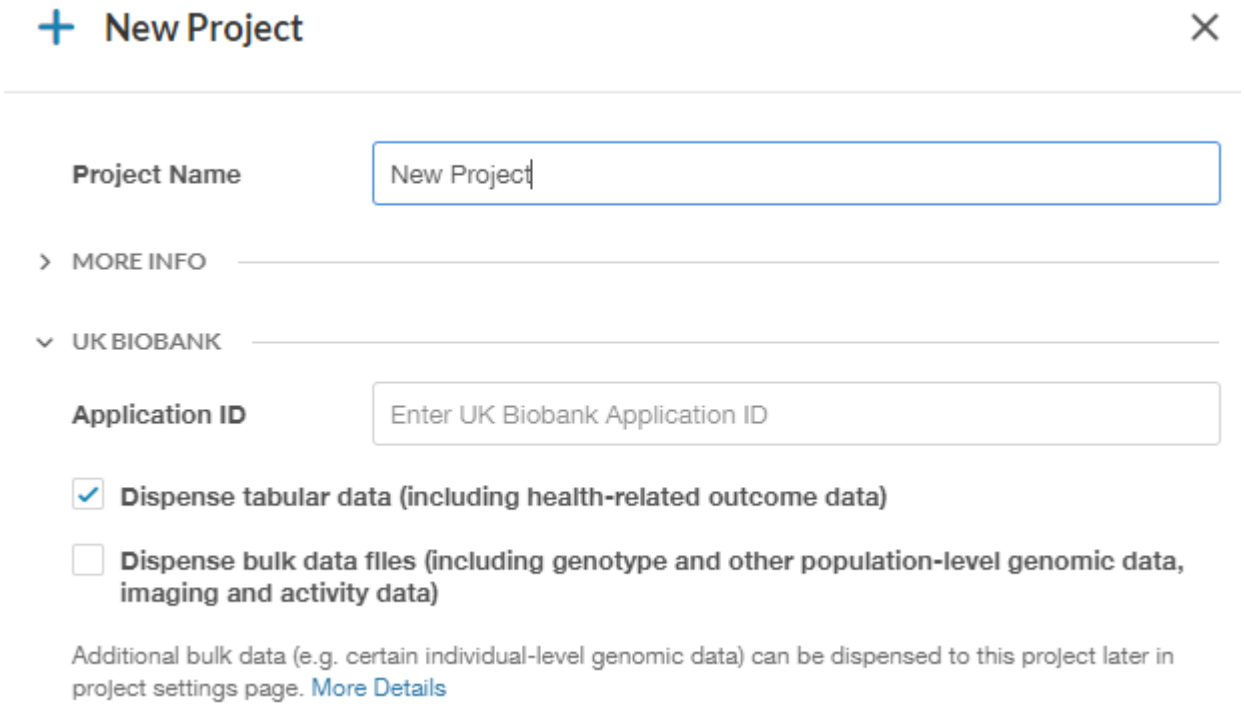
You can subscribe at https://status.dnanexus.com/

2. How many projects can I dispense data to?

We recommend that each research application dispense data to only one project to be considerate for of other researchers who would like to access the data.

3. How do I select the data I'd like to dispense?

Within UKB-RAP you can selectively dispense within the project workspace creation process. Users will now see a new section with the different data types available to dispense. For a faster dispense time, select only the data you'll need.



4. What data should I select for dispensal?

If you are interested in accessing the updated phenotypic, health care and proteomics data, select **structured tabular data**. This option is selected by default but can be unselected if the data is not necessary for your project.

If you are interested in accessing the updated imaging data or the population-level WGS pVCF data, select **unstructured bulk data files**. This option will dispense population-level WGS pVCF data (600,000 files), but not individual-level WGS data such as CRAM or gVCF files. This was decided in order to streamline the new project experience for all users. If your research requires access to the individual-level WGS data (18 million files), return back to the project once the initial dispensing is completed and request an additional dispensing of these data files. Due to the size of the dispensal we recommend waiting until the demand for the WGS has reduced. Please email ukbiobank-support@dnanexus.com if you would like to have an estimate of your expected waiting time.

5. How long will the dispensal process take?

Each dispense request will take about 4-8 hours. However, due to the considerable number of people interested in 500k WGS data and the size of this data, you might experience a long waiting time. **Please do not dispense more than one project.** Please email ukbiobank-support@dnanexus.com if you would like to have an estimate of your expected waiting time.

6. I created a project, but it is stuck at "0%".

Your request to dispense data may be queued behind other users' requests. The system will service your request in the order it was received. Please email ukbiobank-support@dnanexus.com if you have concerns.

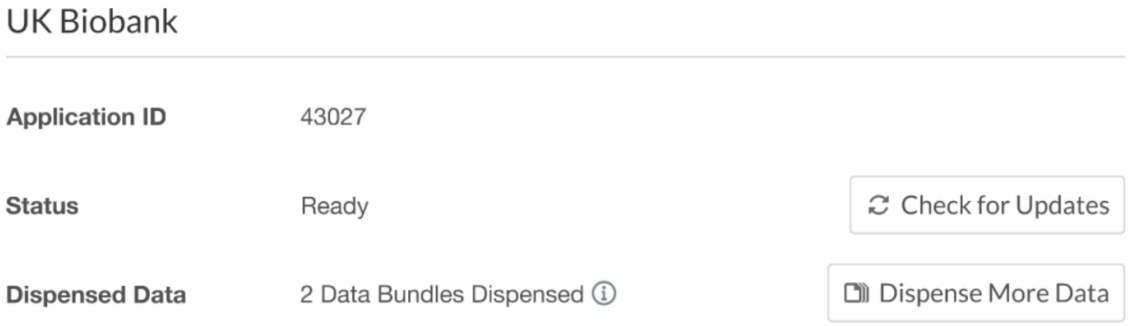
7. What data should I select for dispensal?

If you are interested in accessing the updated phenotypic, health care and proteomics data, select structured tabular data. This option is selected by default but can be unselected if the data is not necessary for your project.

If you are interested in accessing the updated imaging data or the population-level WGS pVCF data, select unstructured bulk data files. This option will dispense population-level WGS pVCF data (600,000 files) but not individual-level WGS data such as CRAM or gVCF files. This ability to select data was created to streamline the project dispensal experience for all users. If your research requires access to the individual-level WGS data, return to the project once the initial dispensing is completed and request an additional dispensal of these data files.

8. How do I dispense the individual-level data?

Due to the size of the data we recommend waiting until demand for the population-level WGS data has reduced. If your research requires access to the individual-level WGS data, you will have to request an additional dispensal after your first request has completed. You can make the request in your project settings and select the "Dispense More Data" button.



9. Could I "refresh" existing projects to get 500k WGS data?

Currently the refresh feature is unavailable to ensure that the maximum number of users can get access to the new data as soon as possible via dispensal. We recommend that users dispense a new project and migrate data analysis workflows from existing projects to the new project. We will enable the refresh feature again in the future and send notifications once it is available.

Section 3: Experimental design and data analysis pipeline queries

1. What sequencing technology has been used for UK Biobank WGS?

The samples have been sequenced using Illumina NovaSeq 6000 instruments, with paired-end sequencing performed using S4 flowcells (v1.0 chemistry).

2. Do the CRAM files also contain unmapped reads?

Yes. Original sample FASTQs can be re-created from the lossless CRAMs, which contain every read regardless of whether they map and all original quality scores. Please note that CRAMs should be name sorted or randomised prior to extracting a FASTQ to ensure uncorrelated read sets for subsequent parallelised mapping (e.g. BWA).

3. Why are there two versions of the dataset?

During the WGS Main Phase programme, the industry consortium chose to further process the individual level data using the Illumina DRAGEN v3.7.8 pipeline. As well as taking advantage of the potential improvements offered by the mapping and calling algorithms within DRAGEN, other large-scale population genomics initiatives have sought to standardise on the DRAGEN v3.7.8 pipeline with the aim of simplifying cross-cohort analyses.

The canonical outputs of the WGS programme should be seen to be the DRAGEN version of these data. Given DRAGEN remains an upcoming standard increasingly used within the genetics community, and that many researchers will currently use BWA-MEM/GATK as their preferred version, we have chosen to make both versions of the data available at this time to ensure that it is accessible to as many researchers as possible. Please note that the BWA-MEM/GATK CRAMs may be deprecated in the future, a decision will be communicatedto researchers in early 2024..

4. Why are the EIDs in the header of the gVCF and CRAM different to the filename?

The EID in the filename is pseudonymised to match your application EIDs. These EIDs are consistent across your project space, for all bulk and tabular data. Please disregard any sample IDs within the gVCF, VCF and CRAM files. Further information is provided here: https://dnanexus.gitbook.io/uk-biobank-rap/frequently-asked-questions#are-the-headers-of-gvcf-or-cram-files-pseudonymized

5. Why are the data no longer within the fields used for the 200k release?

For individual-level data fields, the data made available in November 2021 as part of the 200k release have been merged into enduring fields for this final release. This allows researchers to access the entirety of these file types within a single field, streamlining analyses. The population-level files, such as the 150k and 200k joint variant calls and phased datasets, are still available to researchers in Category 271.

The relationship between the 200k and new 500k fields is as follows:

200k Field ID	Old Field Name	New 500k Field ID	New Field Name
2319323194	Whole genome CRAM files Whole genome CRAM indices	23372	Whole genome GATK CRAM files and indices [500k release]
23197	BQSR - GATK BaseRecalibrator	23376	BQSR - GATK BaseRecalibrator [500k release]
2319123192	Whole genome variant call files (VCFs) Whole genome variant calls indices	23370	Whole genome GATK variant call files (VCFs) and indices [500k release]
23348	Sample Contamination (verifyBamID) - depthSM	23383	Sample Contamination (verifyBamID) - depthSM [500k release]
23346	Genotype Concordance	23381	Genotype Concordance [500k release]

23349	Sample Contamination (verifyBamID) - selfSM	23384	Sample Contamination (verifyBamID) - selfSM [500k release]
23347	Concatenated QC Metrics	23382	Concatenated QC Metrics [500k release]
23345	Genotype Concordance - Summary Metrics (Picard)	23380	Genotype Concordance - Summary Metrics (Picard) [500k release]
23319	Genotype Concordance - Detail Metrics	23379	Genotype Concordance - Detail Metrics [500k release]
23199	Genotype Concordance - Contingency Metrics	23378	Genotype Concordance - Contingency Metrics [500k release]
23198	Sample Contamination (ReadHaps)	23377	Sample Contamination (ReadHaps) [500k release]

6. How can I tell whether a participant was sequenced by Wellcome Sanger Institute (WSI) or deCODE Genetics?

The sequencing provider can be determined using Field 32051. Samples sequenced in the Vanguard Pilot and Vanguard Phase were sequenced by WSI, in addition to those sequenced by WSI as part of the main phase of the sequencing project.

7. Will an updated version of the genotyping array data be released on GRCh38 to align with the 500k WGS data?

Not at this time. The array data is still in hg37. However, this documentation page provides example code that can be used for LiftOver from GRCh37 to GRCh38.

8. Are there plans to release phased versions (haplotypes) of the new UK Biobank WGS data?

Yes, we planned to release phased versions of the DRAGEN joint variant call in 2024.