Australian UK Biobank Symposium Instruction Day

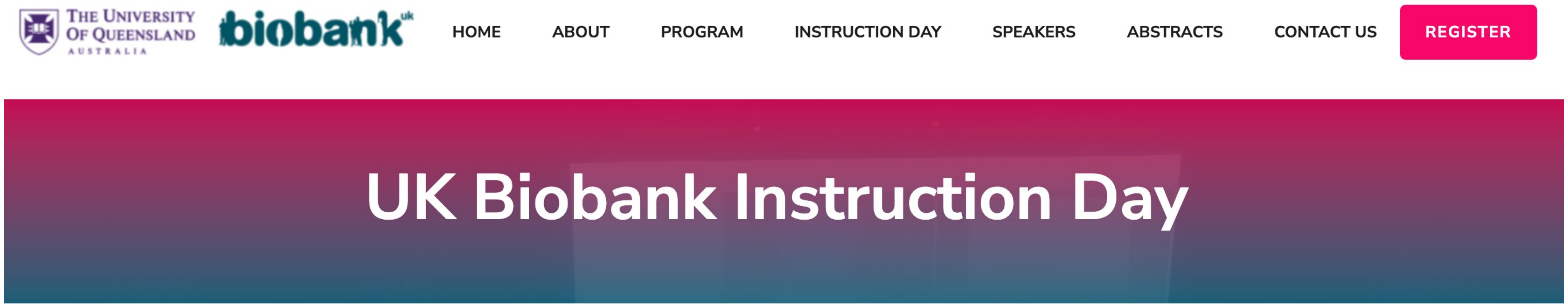
Research Analysis Platform (RAP)

Huanwei Wang
huanwei.wang@uq.edu.au

6/Feb/2024

Download Slides

URL: https://ausukb2024.org/uk-biobank-instruction-day/



UK Biobank Instruction Day

What is RAP?

The UK Biobank Research Analysis Platform (RAP) is an informatics platform that provides **access** to, and **analysis** of, UK Biobank data by its **registered researcher** community.

>390 Community Members

>2,700 Projects

200+ Tools

>500 Active Users

biobank uk

Research Analysis Platform

Enabling scientific discoveries that improve human health

ACCESS PLATFORM

SUPPORT | DOCUMENTATION | COMMUNITY

© 2021 DNAnexus, Inc. All Rights Reserved

DNAnexus aws

3,297 Users

26+ Webinars and Workshops

27.1PB Data

11M Files

463+ Apps

https://dnanexus.gitbook.io/uk-biobank-rap/

Why RAP?

Original dataset (2012): ~20TB

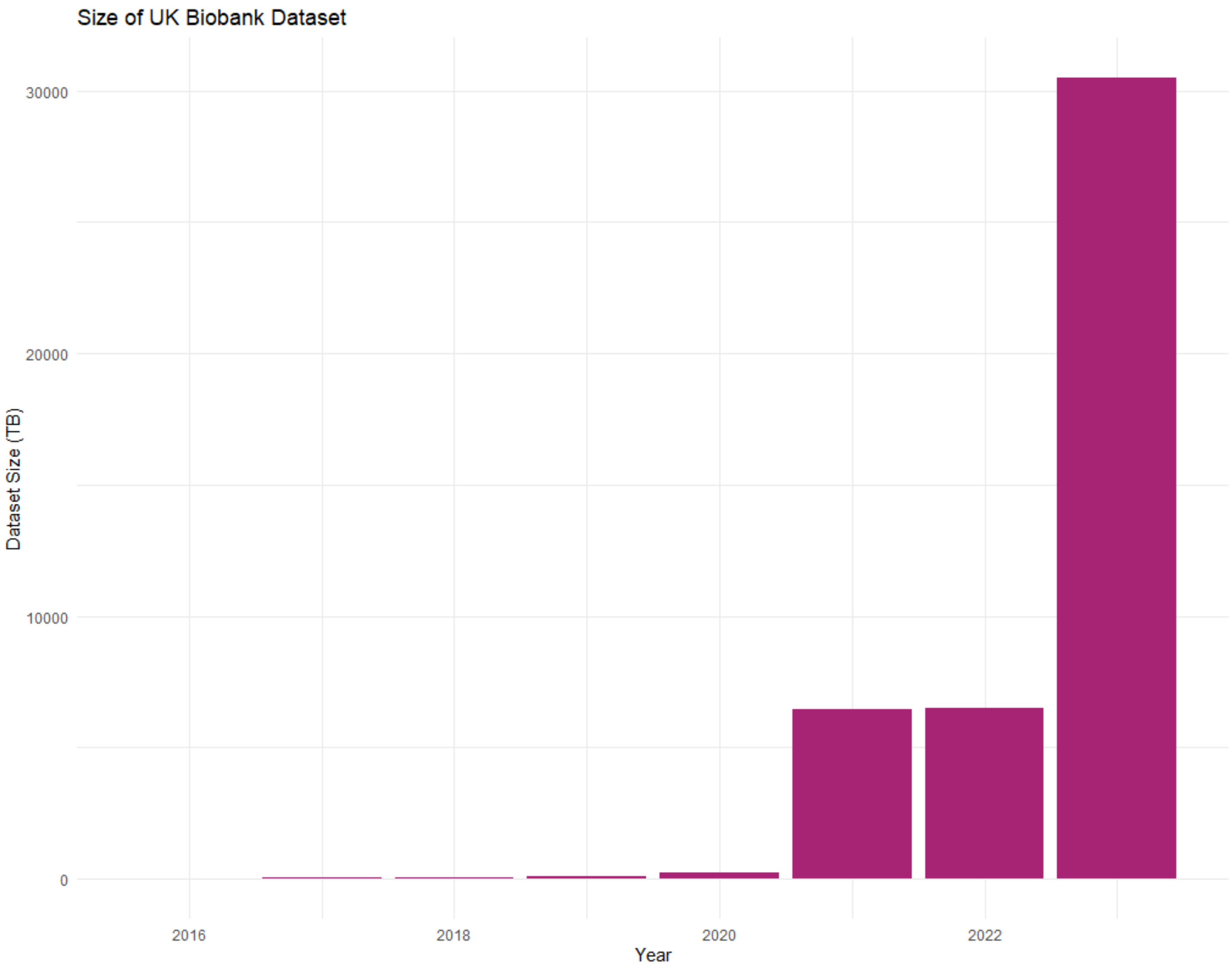
Release of genotyping data (2017): +10TB

First Whole Exome Sequencing release (2019): +45TB

200k WES release (2020): +45TB

200k WGS release (2021): +6PB (+6,000TB)

500k WGS release (2023): to 30PB



Learning objects

- How to set-up your RAP account
- How to explore the UKB data on RAP

- What are the key differences between cloud- and cluster-based computing
- How to find existing apps on RAP
- How to submit a job on UI (user interface)
- How to install dx, the CLI (command line interface)
- How to submit a job using dx

Set-up an account

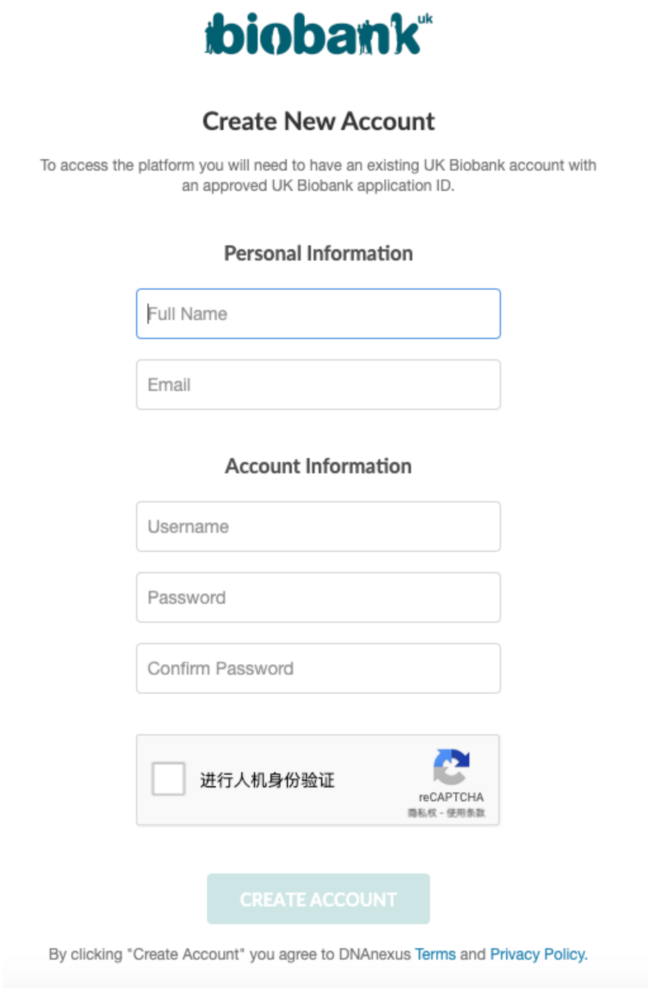
Set-up your RAP account

Step 1: register an account on RAP

Research Analysis Platform

Enabling scientific discoveries that improve human health

© 2024 DNAnexus, Inc. All Rights Reserved



Create New Account

To access the platform you will need to have an existing UK Biobank account with an approved UK Biobank application ID.

Personal Information

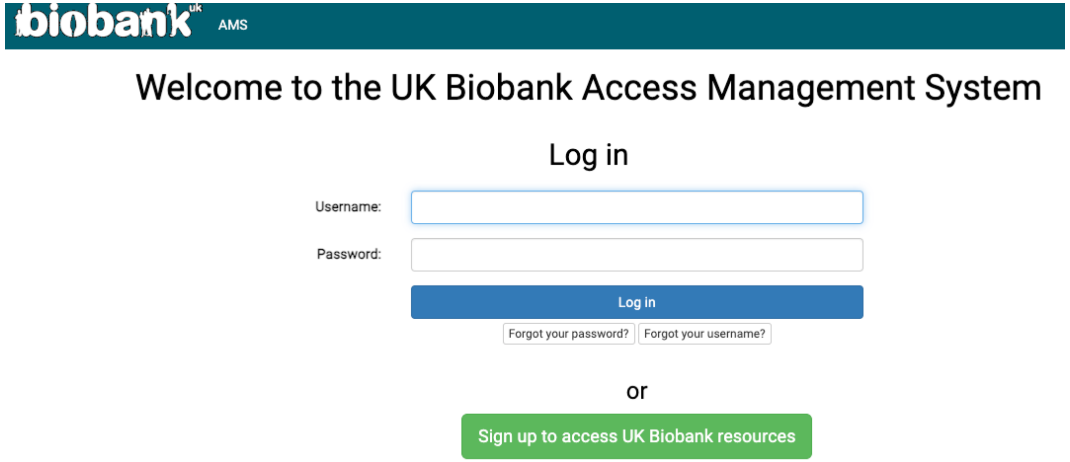
Account Information

By clicking "Create Account" you agree to DNAnexus Terms and Privacy Policy.

https://ukbiobank.dnanexus.com/landing

Set-up your RAP account

Step 2: Connecting RAP account to UKB AMS account



Welcome to the UK Biobank Access Management System

Log in

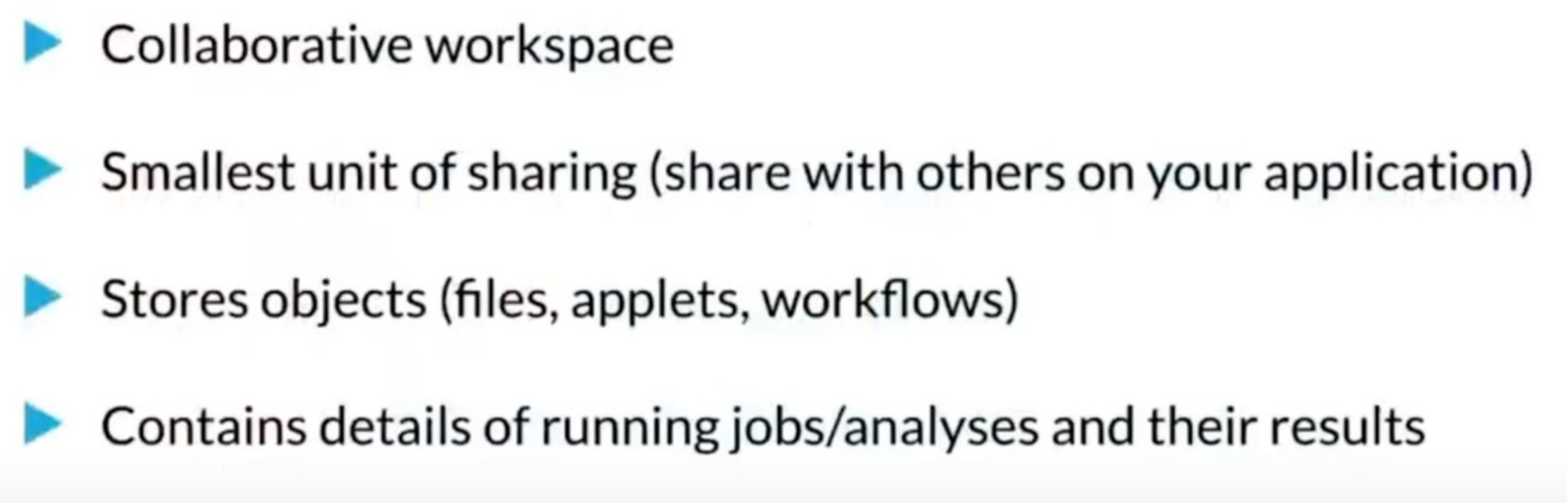
Username:

Password:

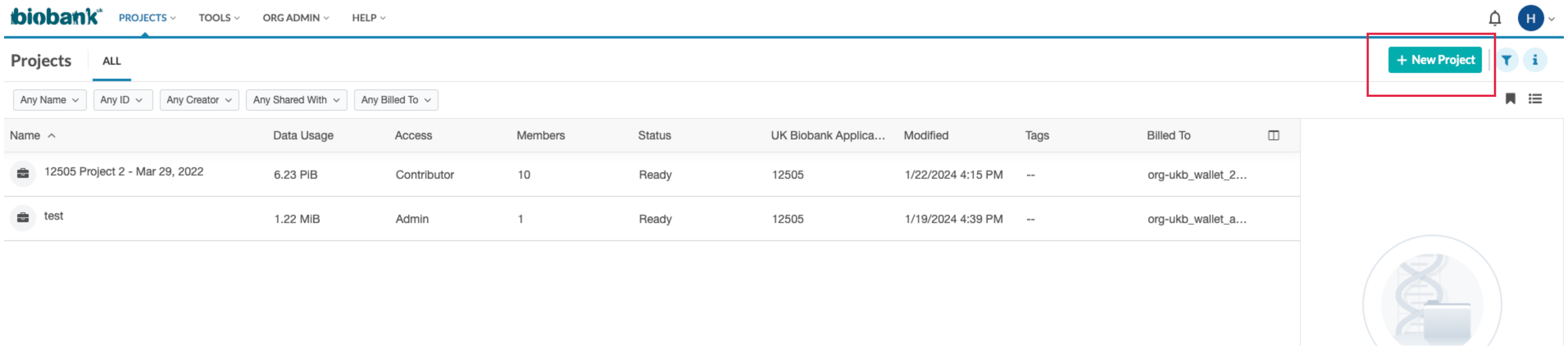
https://dnanexus.gitbook.io/uk-biobank-rap/getting-started/creating-an-account

Project

Create a project



- Collaborative workspace
- Smallest unit of sharing (share with others on your application)
- Stores objects (files, applets, workflows)
- Contains details of running jobs/analyses and their results



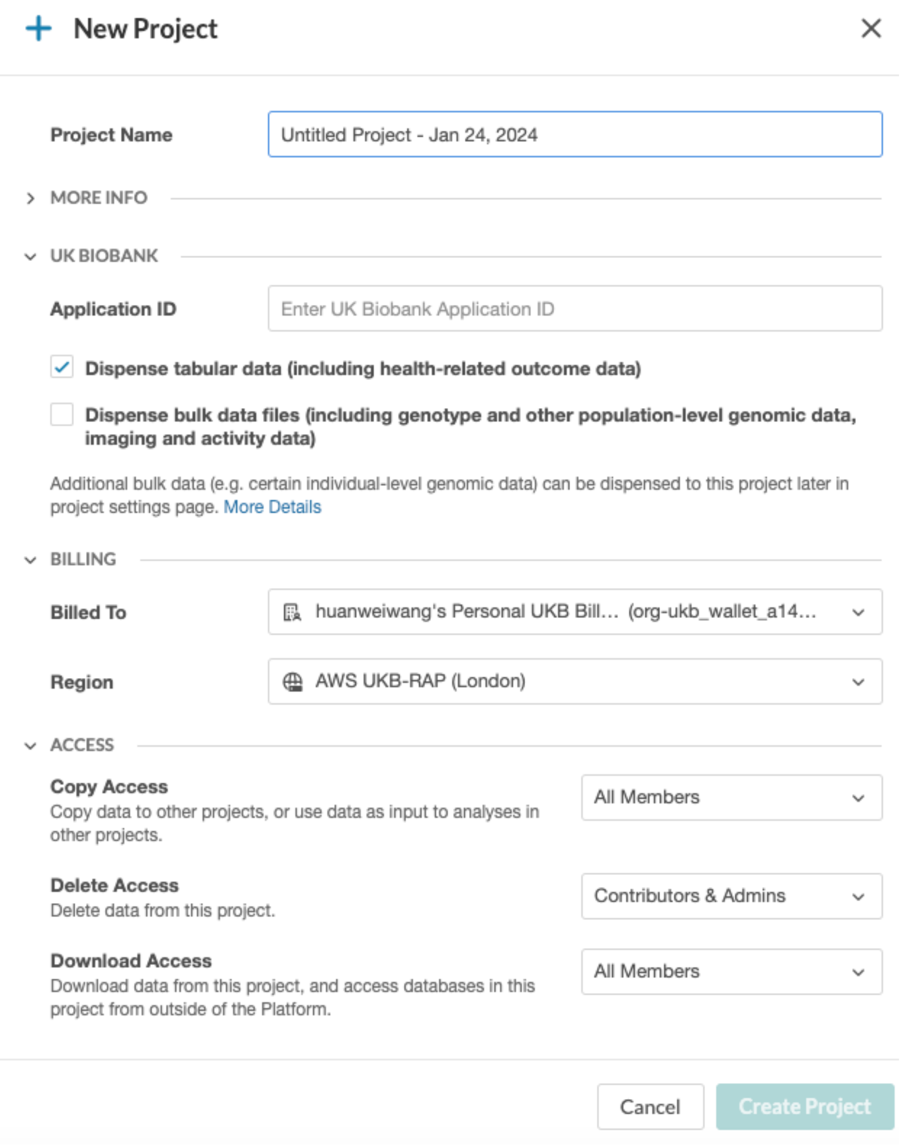
Name	Data Usage	Access	Members	Status	UK Biobank Applica...	Modified	Tags	Billed To
12505 Project 2 - Mar 29, 2022	6.23 PiB	Contributor	10	Ready	12505	1/22/2024 4:15 PM	--	org-ukb_wallet_2...
test	1.22 MiB	Admin	1	Ready	12505	1/19/2024 4:39 PM	--	org-ukb_wallet_a...

https://dnanexus.gitbook.io/uk-biobank-rap/getting-started/creating-a-project
https://www.youtube.com/watch?v=uT_jD1Ey3Fk

Create a project

Application ID

Dispensing data to your new project will take some time (1-24h)



New Project

Project Name Untitled Project - Jan 24, 2024

MORE INFO

UK BIOBANK

Application ID Enter UK Biobank Application ID

☑ **Dispense tabular data (including health-related outcome data)**

☐ **Dispense bulk data files (including genotype and other population-level genomic data, imaging and activity data)**

Additional bulk data (e.g. certain individual-level genomic data) can be dispensed to this project later in project settings page. More Details

BILLING

Billed To huanweiwang's Personal UKB Bill… (org-ukb_wallet_a14…

Region AWS UKB-RAP (London)

ACCESS

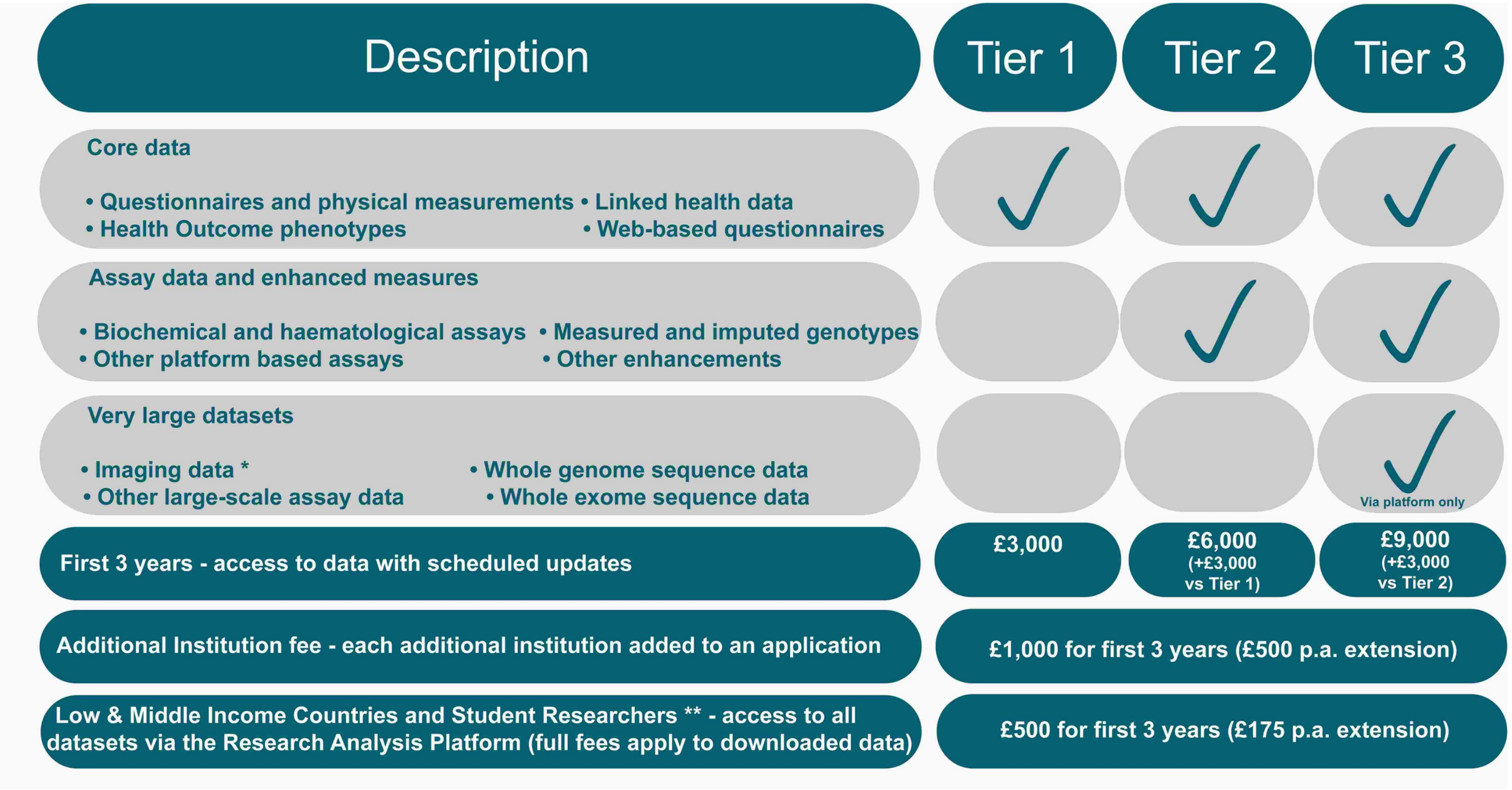
Copy Access All Members
Copy data to other projects, or use data as input to analyses in other projects.

Delete Access Contributors & Admins
Delete data from this project.

Download Access All Members
Download data from this project, and access databases in this project from outside of the Platform.

Dispensing data

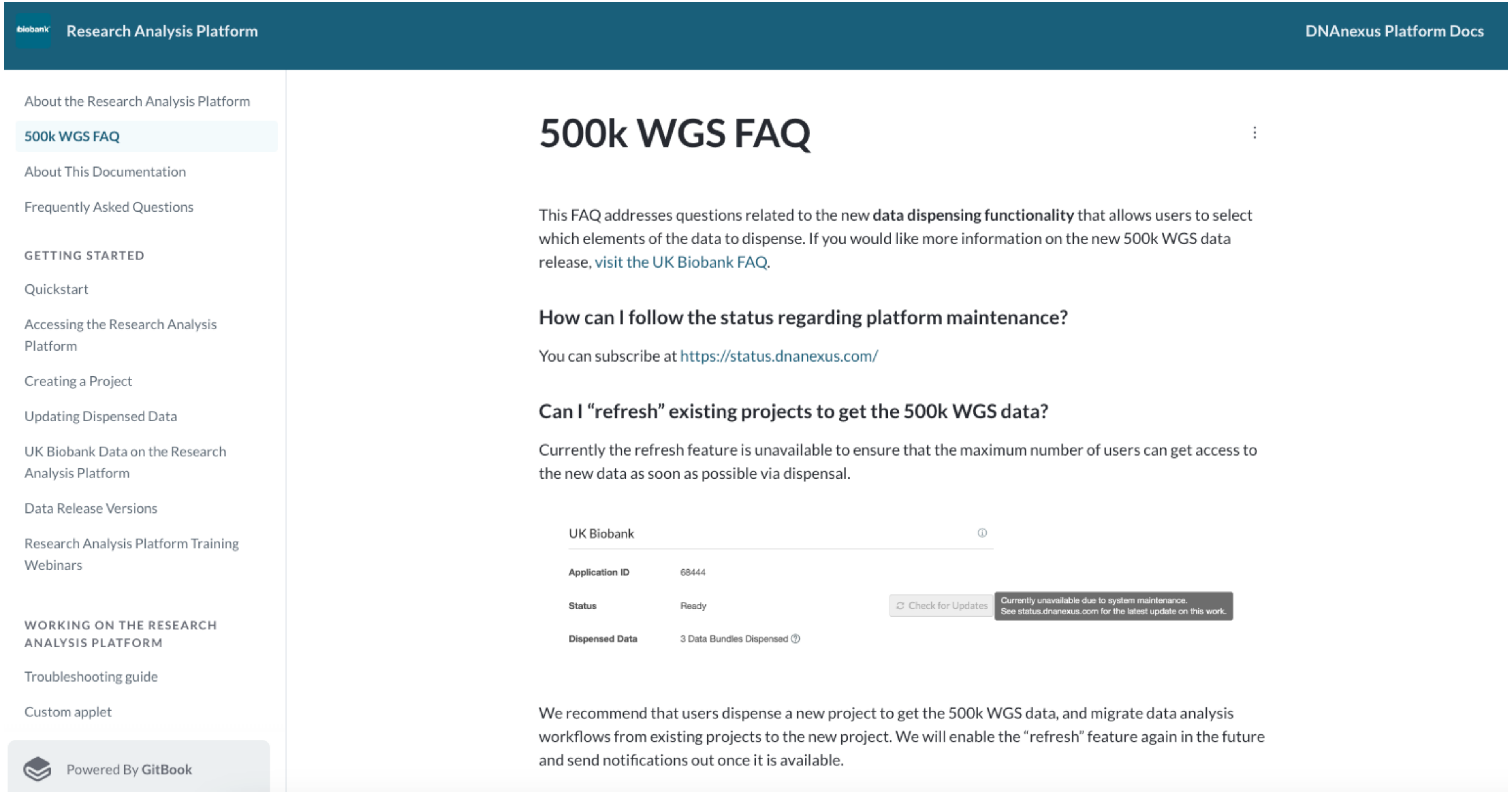
UKB AMS – View – Section B



Description	Tier 1	Tier 2	Tier 3
Core data • Questionnaires and physical measurements • Health Outcome phenotypes • Linked health data • Web-based questionnaires	✓	✓	✓
Assay data and enhanced measures • Biochemical and haematological assays • Other platform based assays • Measured and imputed genotypes • Other enhancements		✓	✓
Very large datasets • Imaging data * • Other large-scale assay data • Whole genome sequence data • Whole exome sequence data			✓ Via platform only
First 3 years - access to data with scheduled updates	£3,000	£6,000 (+£3,000 vs Tier 1)	£9,000 (+£3,000 vs Tier 2)
Additional Institution fee - each additional institution added to an application	£1,000 for first 3 years (£500 p.a. extension)		
Low & Middle Income Countries and Student Researchers ** - access to all datasets via the Research Analysis Platform (full fees apply to downloaded data)	£500 for first 3 years (£175 p.a. extension)		

https://www.ukbiobank.ac.uk/enable-your-research/costs

Dispense 500k WGS



500k WGS FAQ

This FAQ addresses questions related to the new **data dispensing functionality** that allows users to select which elements of the data to dispense. If you would like more information on the new 500k WGS data release, visit the UK Biobank FAQ.

How can I follow the status regarding platform maintenance?

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

Can I “refresh” existing projects to get the 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.

Application ID 68444

Status Ready

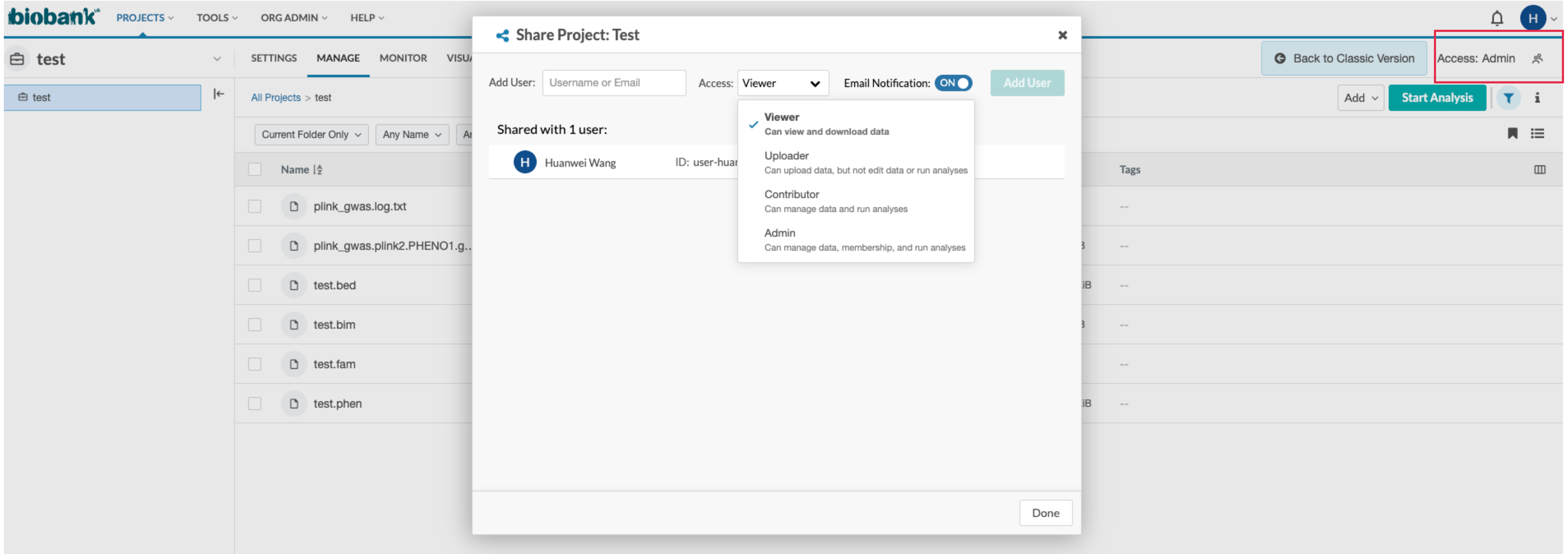
Dispensed Data 3 Data Bundles Dispensed

Currently unavailable due to system maintenance.
See status.dnanexus.com for the latest update on this work.

We recommend that users dispense a new project to get the 500k WGS data, 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 out once it is available.

https://dnanexus.gitbook.io/uk-biobank-rap/500k-wgs-faq

Share project



Share Project: Test

Add User: Username or Email

Access: Viewer

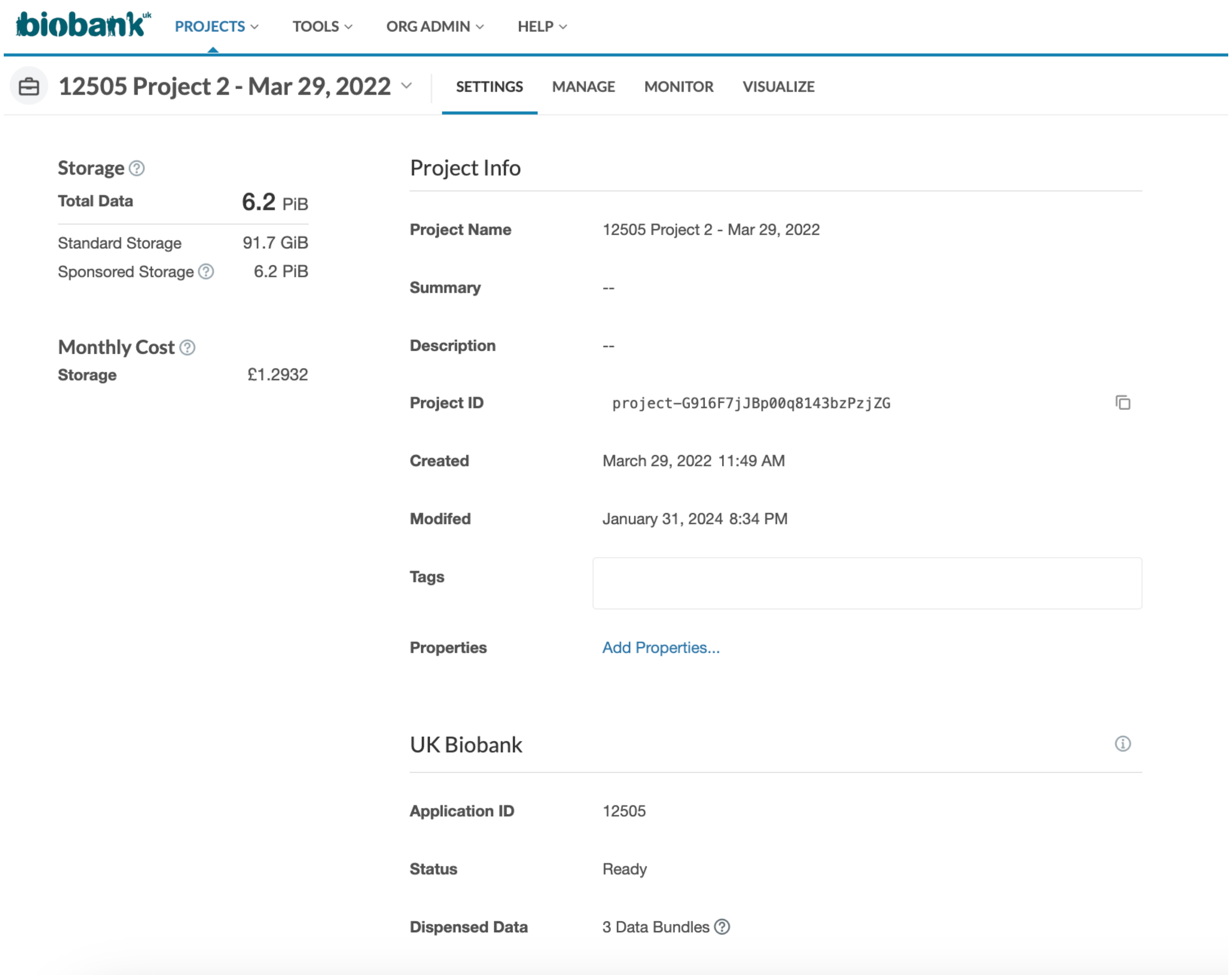
Email Notification: ON

Shared with 1 user:

Huanwei Wang ID: user-huar

- ✓ Viewer — Can view and download data
- Uploader — Can upload data, but not edit data or run analyses
- Contributor — Can manage data and run analyses
- Admin — Can manage data, membership, and run analyses

Project "settings"



12505 Project 2 - Mar 29, 2022

Storage

Total Data	6.2 PiB
Standard Storage	91.7 GiB
Sponsored Storage	6.2 PiB

Monthly Cost

Storage	£1.2932

Project Info

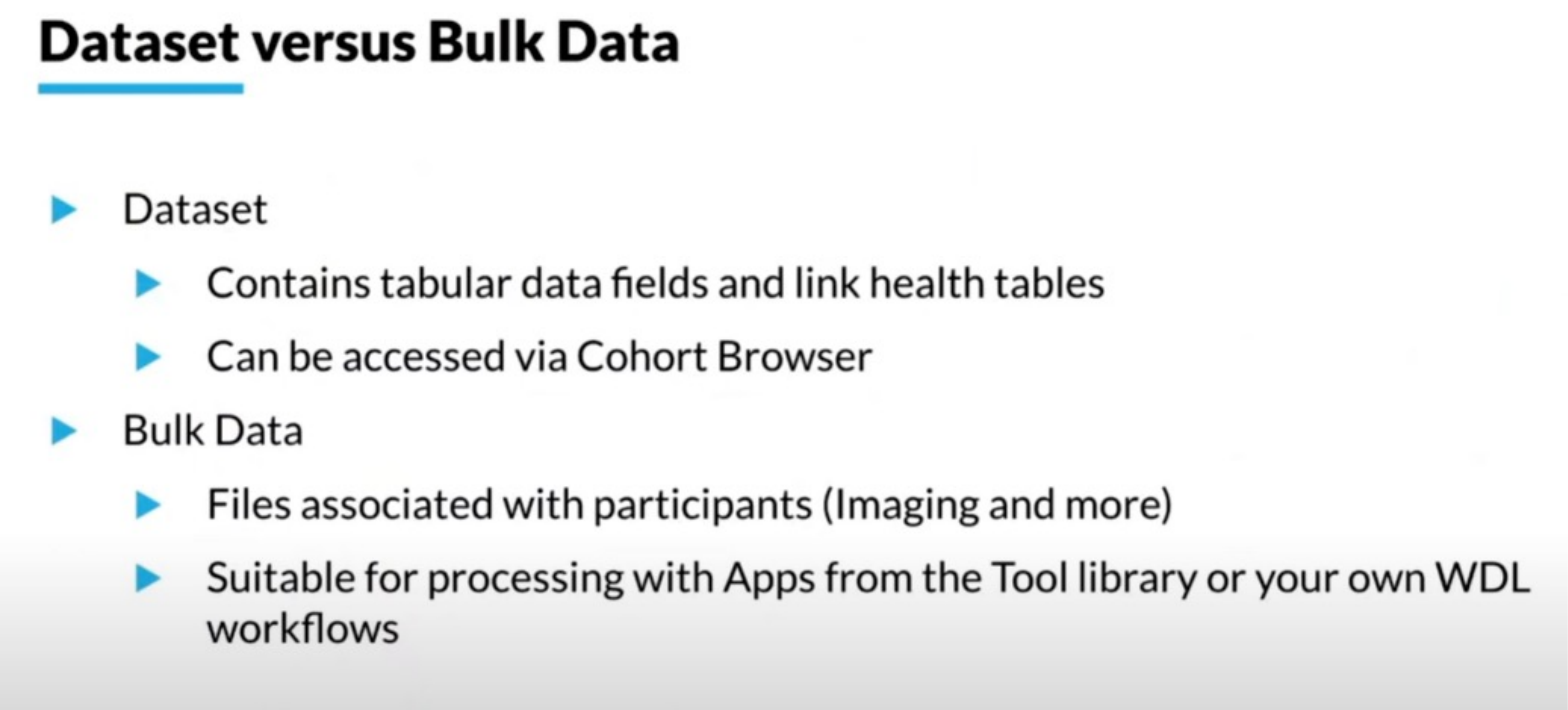
Project Name	12505 Project 2 - Mar 29, 2022
Summary	--
Description	--
Project ID	`project-G916F7jJBp00q8143bzPzjZG`
Created	March 29, 2022 11:49 AM
Modifed	January 31, 2024 8:34 PM
Tags	
Properties	Add Properties...

UK Biobank

Application ID	12505
Status	Ready
Dispensed Data	3 Data Bundles

UKB data on RAP

Data access



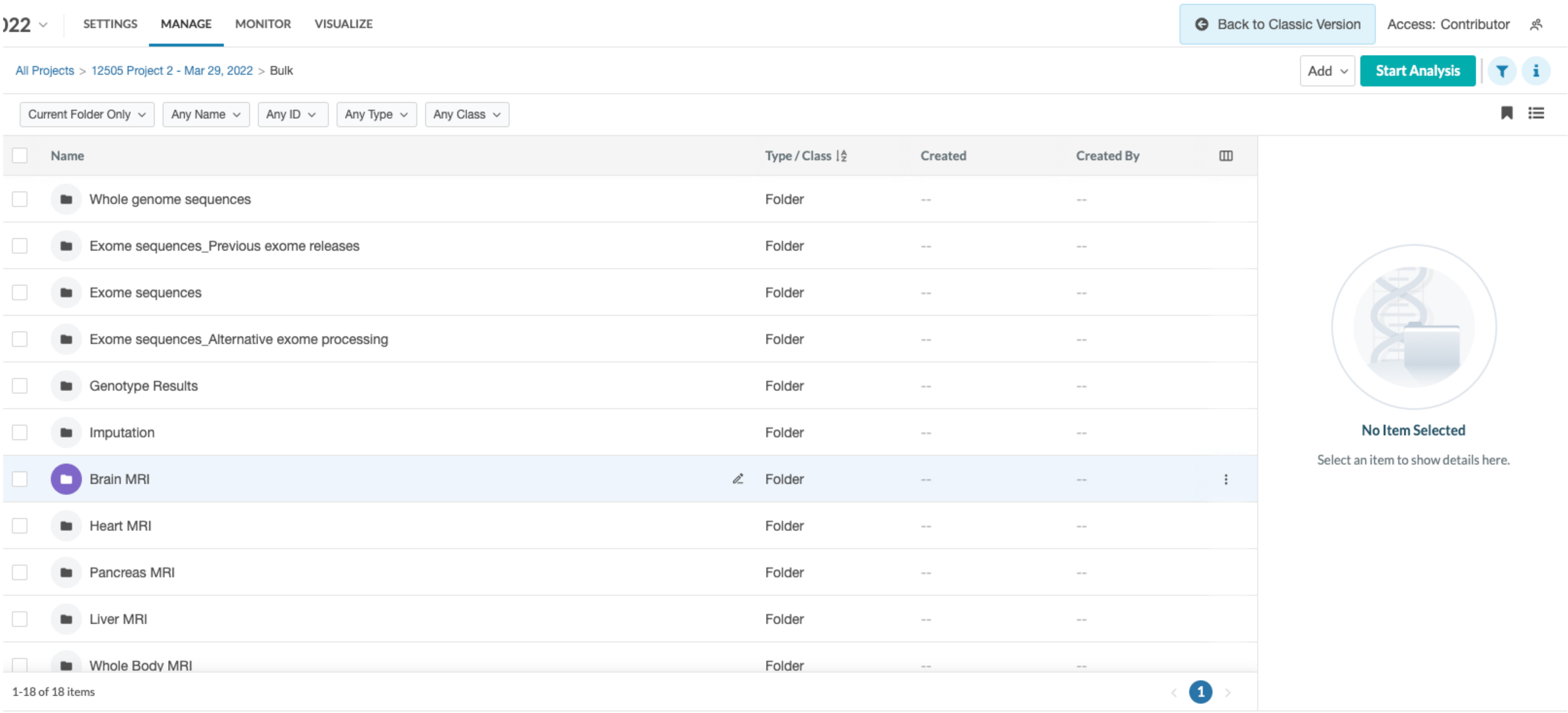
Dataset versus Bulk Data

- Dataset
  - Contains tabular data fields and link health tables
  - Can be accessed via Cohort Browser
- Bulk Data
  - Files associated with participants (Imaging and more)
  - Suitable for processing with Apps from the Tool library or your own WDL workflows

https://dnanexus.gitbook.io/uk-biobank-rap/getting-started/working-with-ukb-data

Bulk data

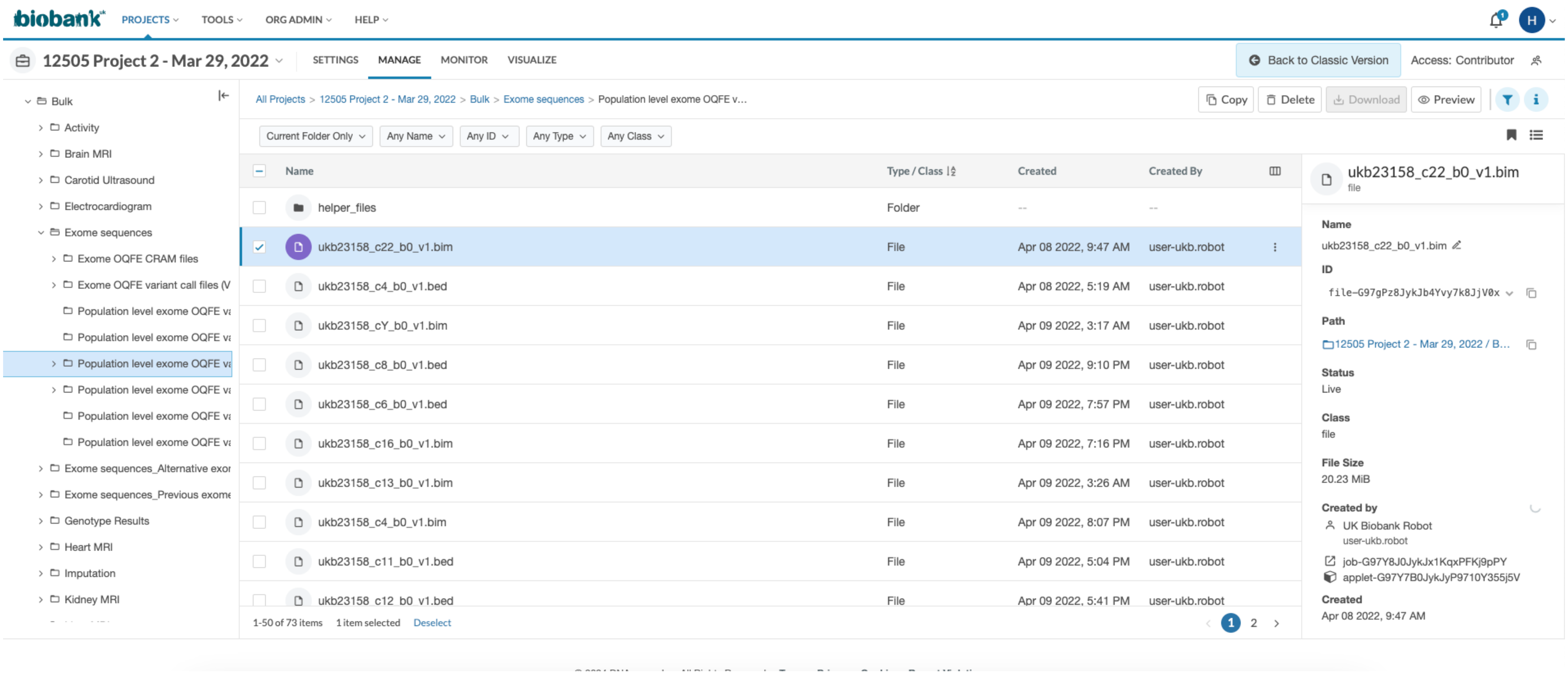
Bulk folder



Name	Type / Class	Created	Created By
Whole genome sequences	Folder	--	--
Exome sequences_Previous exome releases	Folder	--	--
Exome sequences	Folder	--	--
Exome sequences_Alternative exome processing	Folder	--	--
Genotype Results	Folder	--	--
Imputation	Folder	--	--
Brain MRI	Folder	--	--
Heart MRI	Folder	--	--
Pancreas MRI	Folder	--	--
Liver MRI	Folder	--	--
Whole Body MRI	Folder	--	--

https://dnanexus.gitbook.io/uk-biobank-rap/getting-started/working-with-ukb-data

File operation



☐	Name	Type / Class	Created	Created By
☐	helper_files	Folder	--	--
☑	ukb23158_c22_b0_v1.bim	File	Apr 08 2022, 9:47 AM	user-ukb.robot
☐	ukb23158_c4_b0_v1.bed	File	Apr 08 2022, 5:19 AM	user-ukb.robot
☐	ukb23158_cY_b0_v1.bim	File	Apr 09 2022, 3:17 AM	user-ukb.robot
☐	ukb23158_c8_b0_v1.bed	File	Apr 09 2022, 9:10 PM	user-ukb.robot
☐	ukb23158_c6_b0_v1.bed	File	Apr 09 2022, 7:57 PM	user-ukb.robot
☐	ukb23158_c16_b0_v1.bim	File	Apr 09 2022, 7:16 PM	user-ukb.robot
☐	ukb23158_c13_b0_v1.bim	File	Apr 09 2022, 3:26 AM	user-ukb.robot
☐	ukb23158_c4_b0_v1.bim	File	Apr 09 2022, 8:07 PM	user-ukb.robot
☐	ukb23158_c11_b0_v1.bed	File	Apr 09 2022, 5:04 PM	user-ukb.robot
☐	ukb23158_c12_b0_v1.bed	File	Apr 09 2022, 5:41 PM	user-ukb.robot

1-50 of 73 items   1 item selected   Deselect

< 1 2 >

ukb23158_c22_b0_v1.bim
file

Name
ukb23158_c22_b0_v1.bim

ID
file-G97gPz8JykJb4Yvy7k8JjV0x

Path
12505 Project 2 - Mar 29, 2022 / B...

Status
Live

Class
file

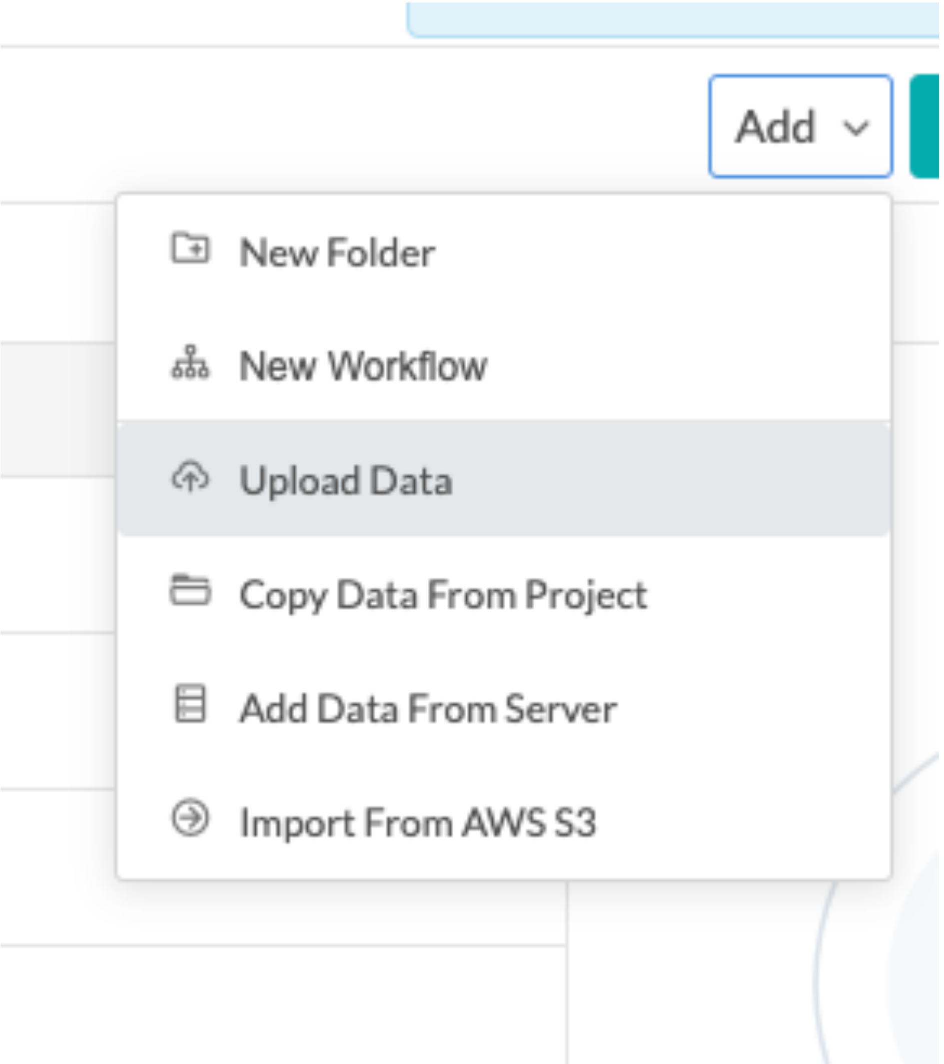
File Size
20.23 MiB

Created by
UK Biobank Robot
user-ukb.robot
job-G97Y8J0JykJx1KqxPFKj9pPY
applet-G97Y7B0JykJyP9710Y355j5V

Created
Apr 08 2022, 9:47 AM

https://dnanexus.gitbook.io/uk-biobank-rap/getting-started/working-with-ukb-data

Upload data on UI

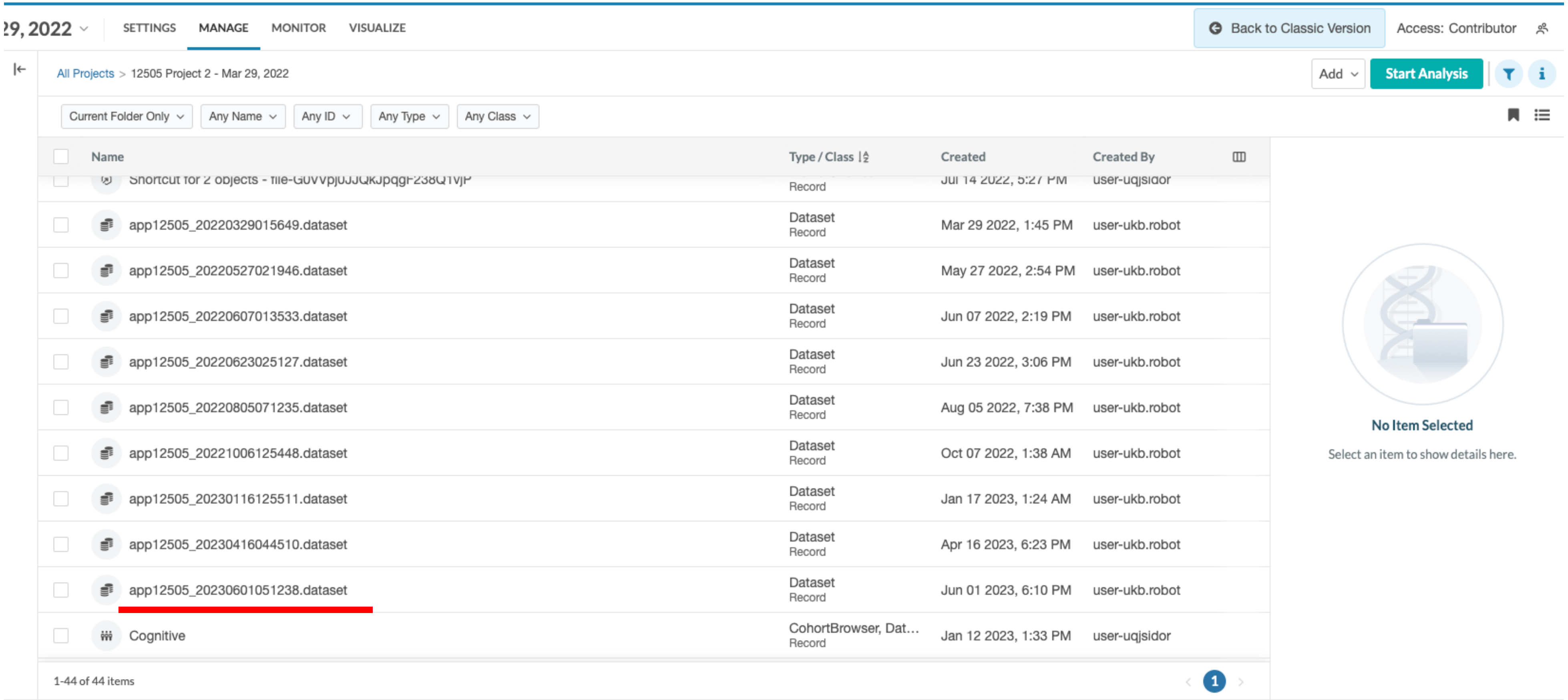


Dataset

Dataset name

Tabular data-fields and linked health data are stored in a SQL database

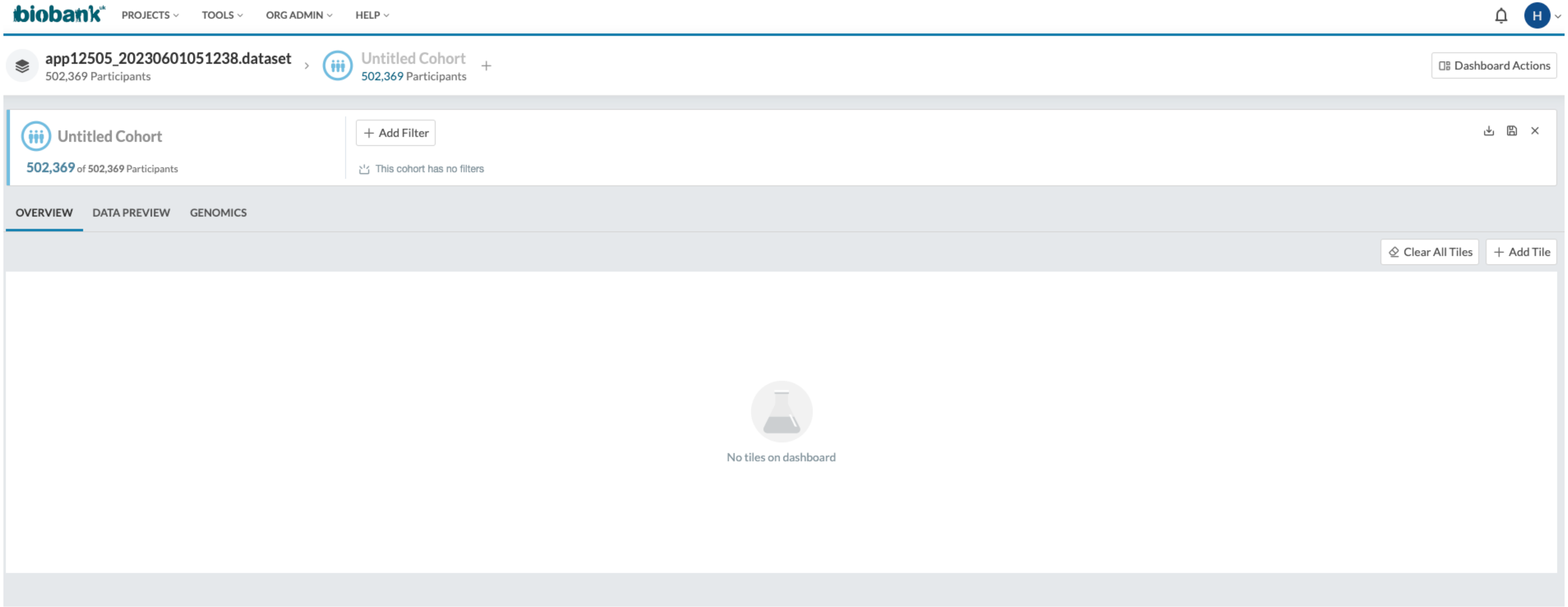
Dataset name: app<APPLICATION-ID>_<CREATION-TIME>



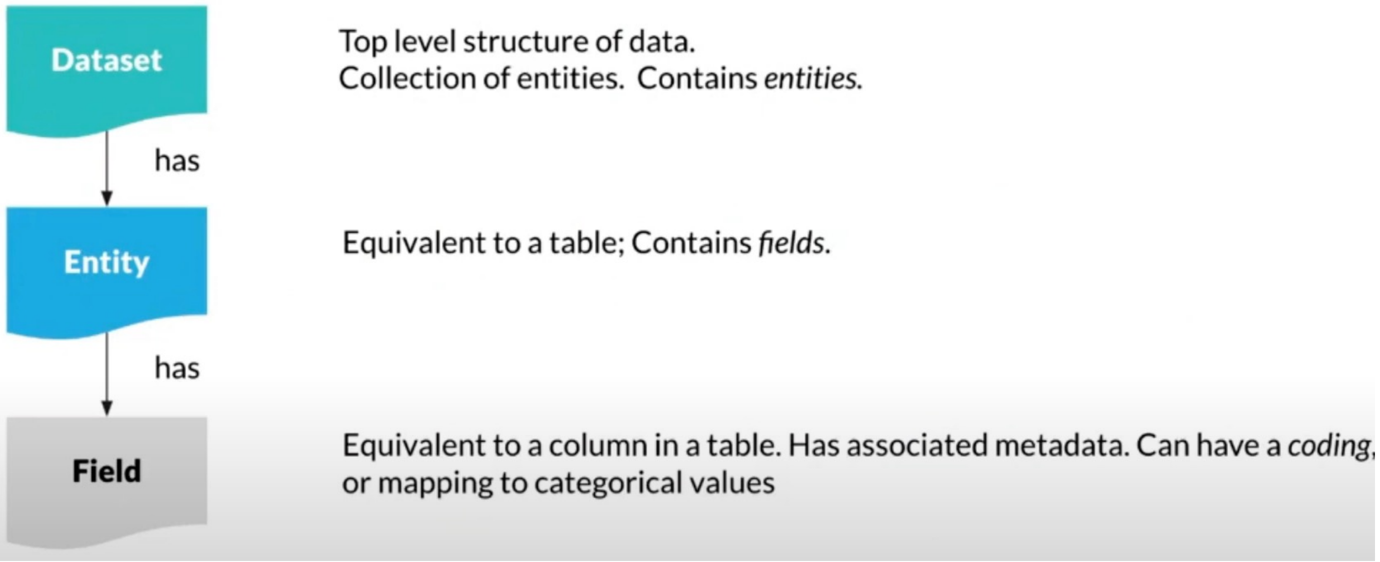
Name	Type / Class	Created	Created By
Shortcut for 2 objects - file-GUVVpj0JJQKJpqgF238Q1VjP	Record	Jul 14 2022, 5:27 PM	user-uqjsidor
app12505_20220329015649.dataset	Dataset Record	Mar 29 2022, 1:45 PM	user-ukb.robot
app12505_20220527021946.dataset	Dataset Record	May 27 2022, 2:54 PM	user-ukb.robot
app12505_20220607013533.dataset	Dataset Record	Jun 07 2022, 2:19 PM	user-ukb.robot
app12505_20220623025127.dataset	Dataset Record	Jun 23 2022, 3:06 PM	user-ukb.robot
app12505_20220805071235.dataset	Dataset Record	Aug 05 2022, 7:38 PM	user-ukb.robot
app12505_20221006125448.dataset	Dataset Record	Oct 07 2022, 1:38 AM	user-ukb.robot
app12505_20230116125511.dataset	Dataset Record	Jan 17 2023, 1:24 AM	user-ukb.robot
app12505_20230416044510.dataset	Dataset Record	Apr 16 2023, 6:23 PM	user-ukb.robot
app12505_20230601051238.dataset	Dataset Record	Jun 01 2023, 6:10 PM	user-ukb.robot
Cognitive	CohortBrowser, Dat... Record	Jan 12 2023, 1:33 PM	user-uqjsidor

https://dnanexus.gitbook.io/uk-biobank-rap/getting-started/working-with-ukb-data
https://documentation.dnanexus.com/developer/datasets

Cohort Browser



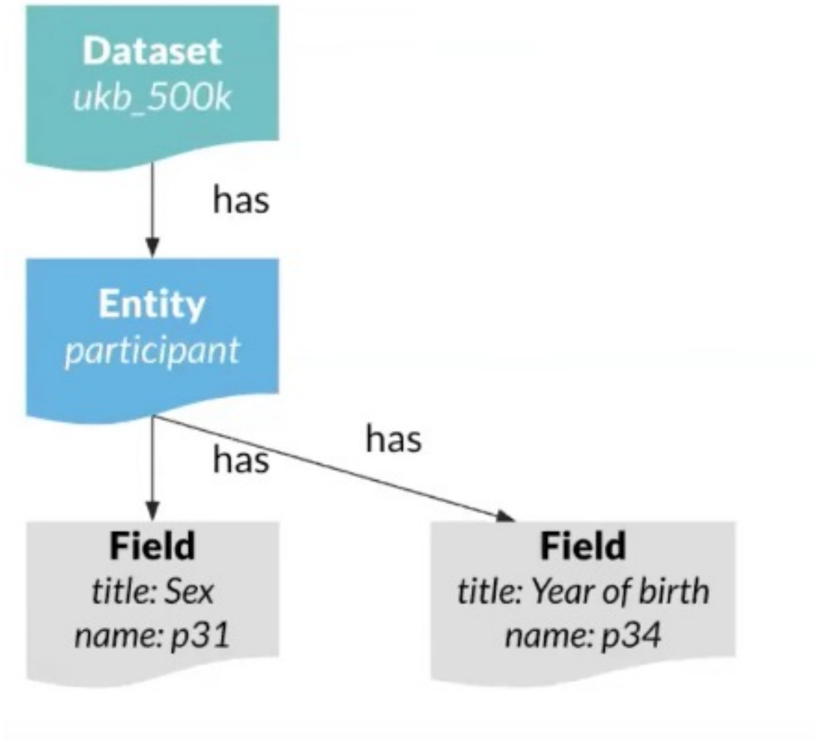
Structure of dataset



Top level structure of data.
Collection of entities. Contains *entities*.

Equivalent to a table; Contains *fields*.

Equivalent to a column in a table. Has associated metadata. Can have a *coding*, or mapping to categorical values



https://dnanexus.gitbook.io/uk-biobank-rap/getting-started/working-with-ukb-data
https://documentation.dnanexus.com/developer/datasets

Entities and fields

Entity list

entity
participant
death
death_cause
hesin_delivery
hesin_diag
hesin_maternity
hesin_oper
hesin_psych
hesin_critical
hesin
covid19_result_england
covid19_result_scotland
covid19_result_wales
gp_clinical
gp_registrations
gp_scripts
olink_instance_0
olink_instance_2
olink_instance_3

Entity and field list

entity	name	type
participant	eid	string
participant	p3_i0	integer
participant	p3_i1	integer
participant	p3_i2	integer
participant	p3_i3	integer
participant	p4_i0	integer
participant	p4_i1	integer
participant	p4_i2	integer

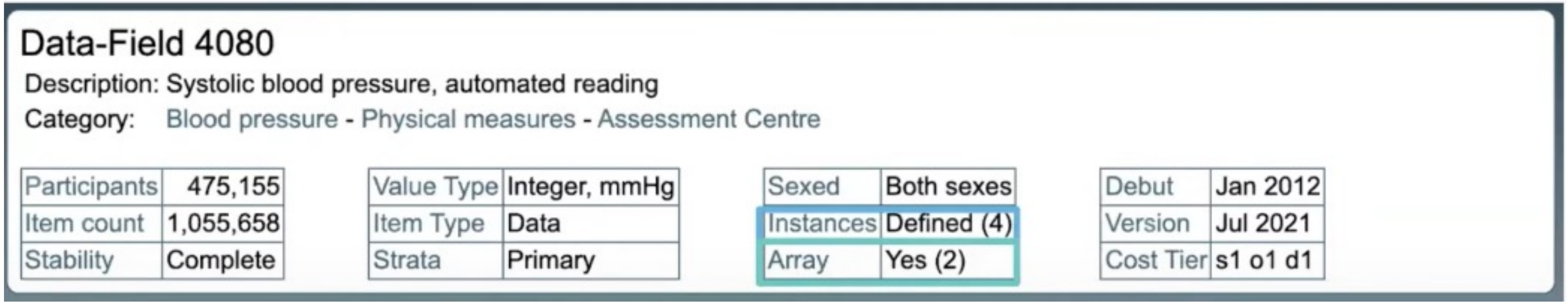
~27,000 fields in the entity of “participant”

https://github.com/dnanexus/OpenBio/blob/master/dx-toolkit/dx_extract_dataset_bash.ipynb

Extracted by “dx extract_dataset” command

Field name notation

- Example: Systolic blood pressure (field 4080)
  - 4 instances - assessment centre visits
    - Initial assessment visit (2006-2010)
    - First repeat assessment visit (2012-13)
    - Imaging visit (2014+)
    - First repeat imaging visit (2019+)
  - Array of length 2
    - Two measures of blood pressure were taken a few moments apart



Data-Field 4080

Description: Systolic blood pressure, automated reading

Category: Blood pressure - Physical measures - Assessment Centre

Participants	475,155
Item count	1,055,658
Stability	Complete

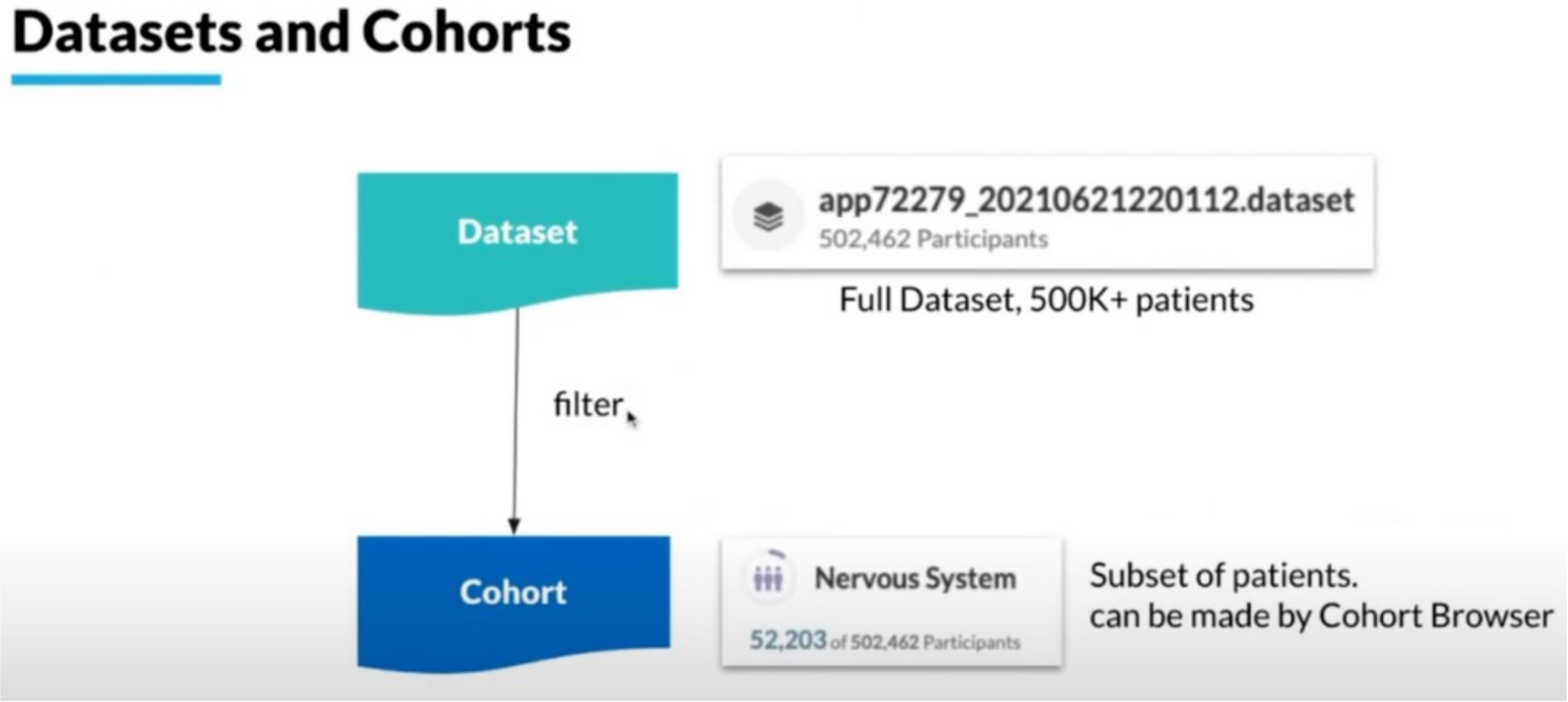
Value Type	Integer, mmHg
Item Type	Data
Strata	Primary

Sexed	Both sexes
Instances	Defined (4)
Array	Yes (2)

Debut	Jan 2012
Version	Jul 2021
Cost Tier	s1 o1 d1

- `p<FIELD-ID>_i<INSTANCE-ID>_a<ARRAY-ID>`
- Example
  - First blood pressure measurement during initial assessment visit
    - p4080_i0_a0
  - Second measurement during first repeat imaging visit
    - p4080_i3_a1

Datasets and cohorts



Datasets and Cohorts

Dataset

app72279_20210621220112.dataset
502,462 Participants

Full Dataset, 500K+ patients

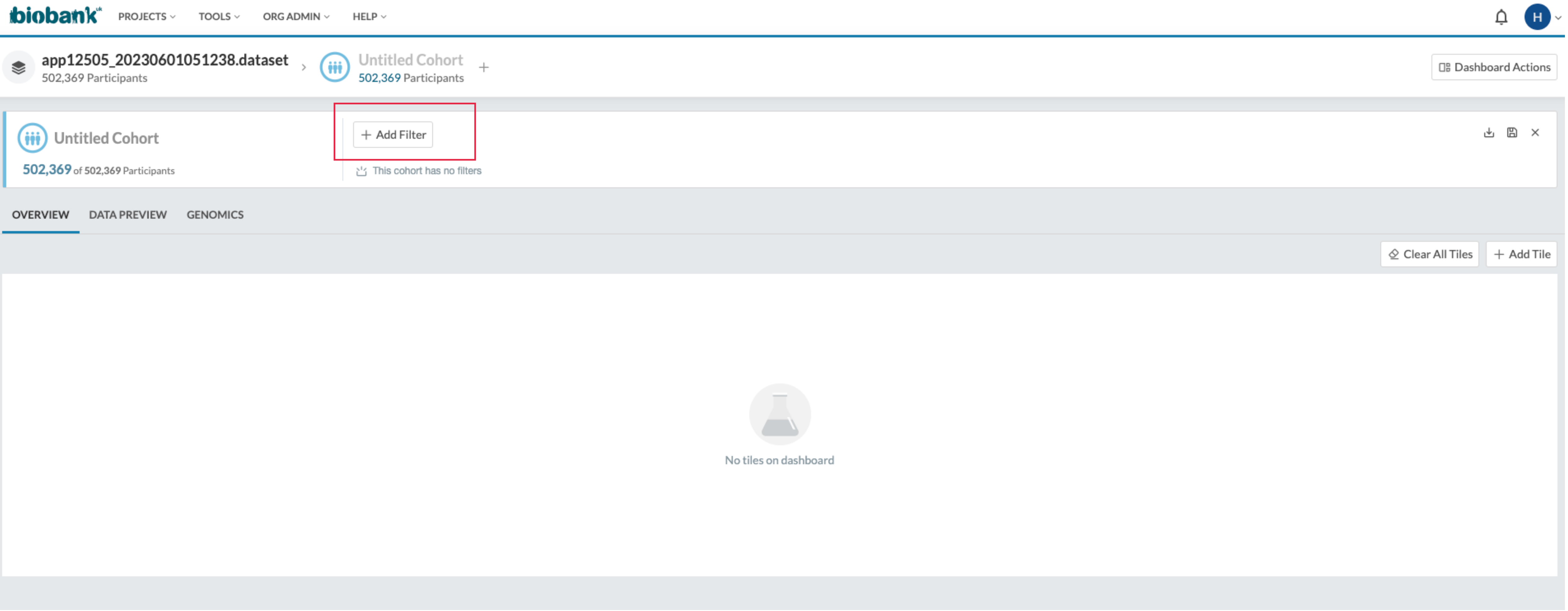
filter

Cohort

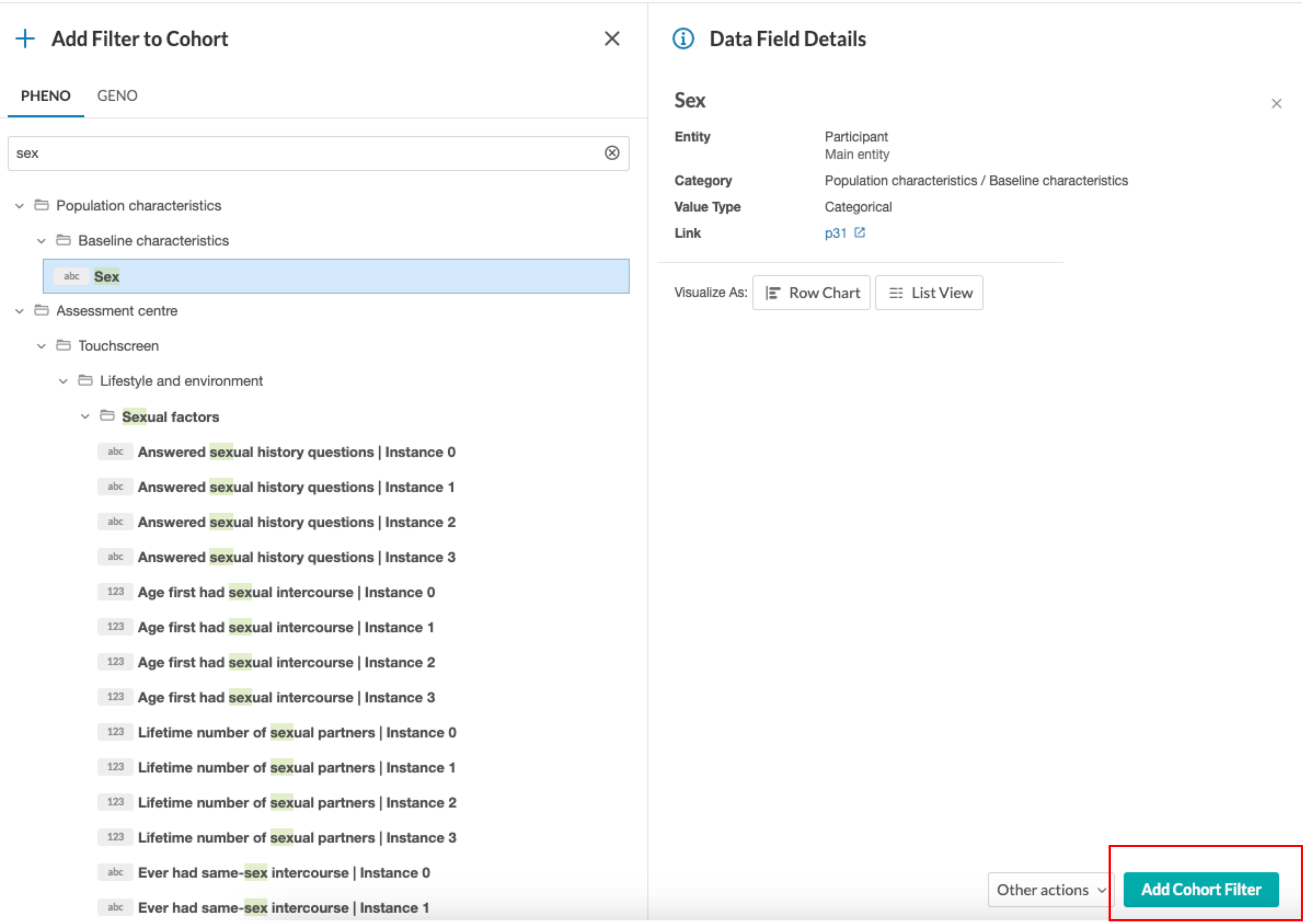
Nervous System
52,203 of 502,462 Participants

Subset of patients.
can be made by Cohort Browser

Add Filter

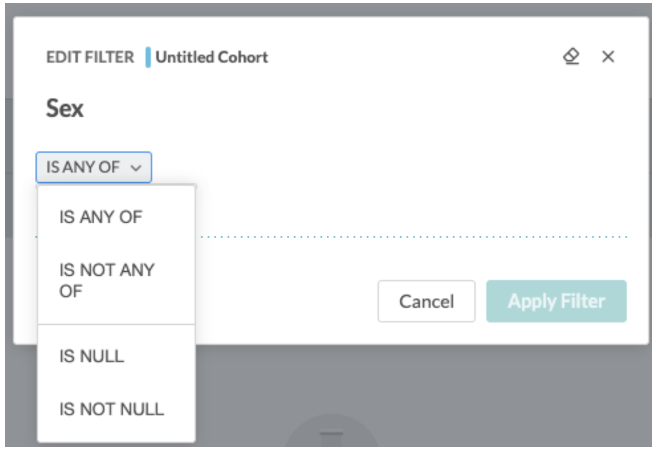

biobank
PROJECTS
TOOLS
ORG ADMIN
HELP
H
app12505_20230601051238.dataset
502,369 Participants
Untitled Cohort
502,369 Participants
Dashboard Actions
Untitled Cohort
502,369 of 502,369 Participants
+ Add Filter
This cohort has no filters
OVERVIEW
DATA PREVIEW
GENOMICS
Clear All Tiles
+ Add Tile
No tiles on dashboard

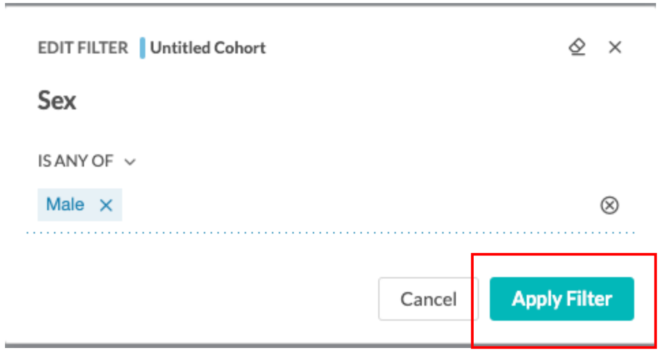
Pheno Filter



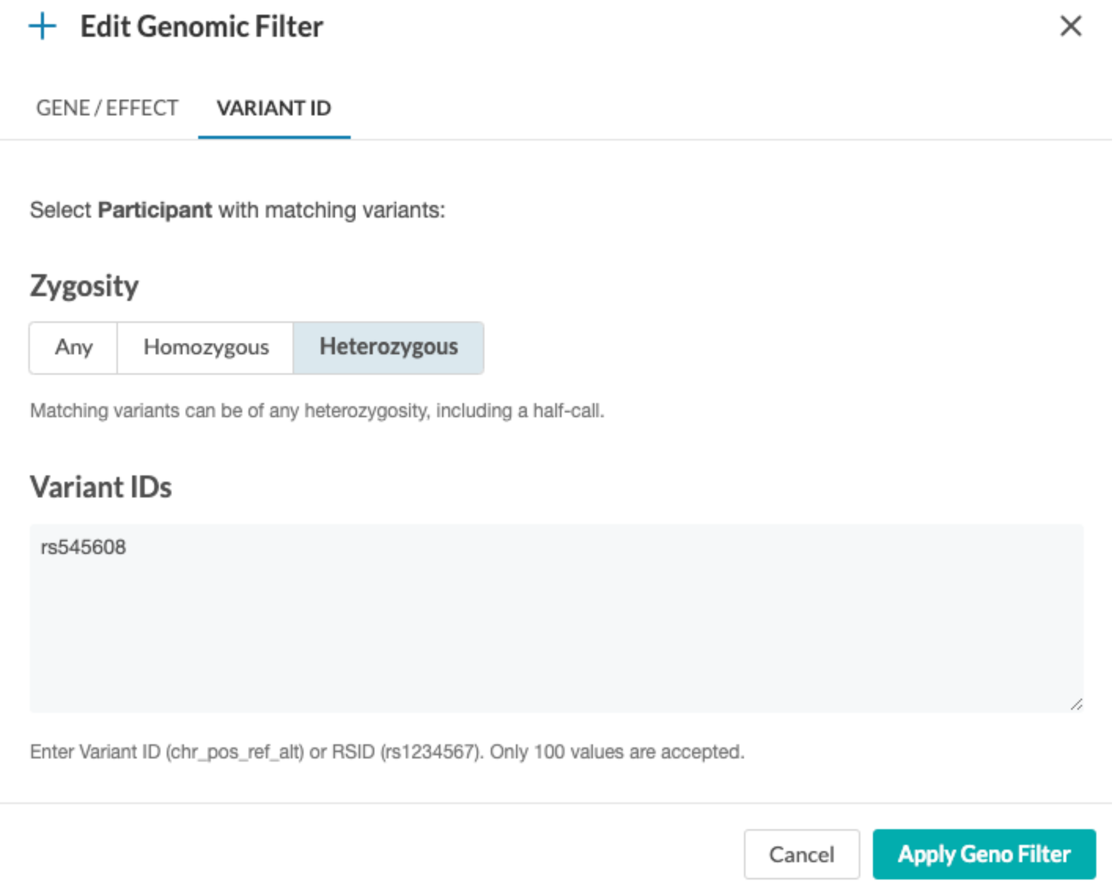
Sex

Entity	Participant Main entity
Category	Population characteristics / Baseline characteristics
Value Type	Categorical
Link	p31





Genomic Filter



Select **Participant** with matching variants:

Zygosity

Matching variants can be of any heterozygosity, including a half-call.

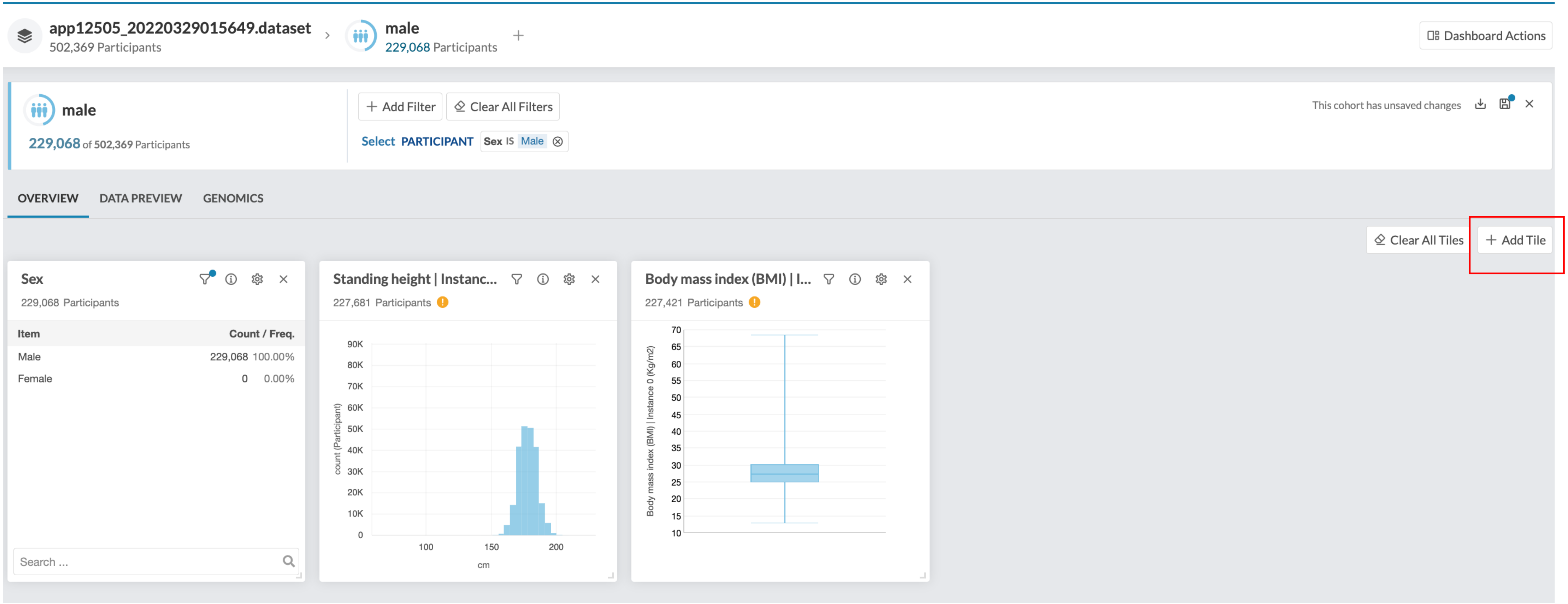
Variant IDs

rs545608

Enter Variant ID (chr_pos_ref_alt) or RSID (rs1234567). Only 100 values are accepted.

https://documentation.dnanexus.com/user/cohort-browser#add-genomic-filter

Overview tab



app12505_20220329015649.dataset
502,369 Participants

male
229,068 Participants

male
229,068 of 502,369 Participants

Select PARTICIPANT Sex IS Male

This cohort has unsaved changes

Sex
229,068 Participants

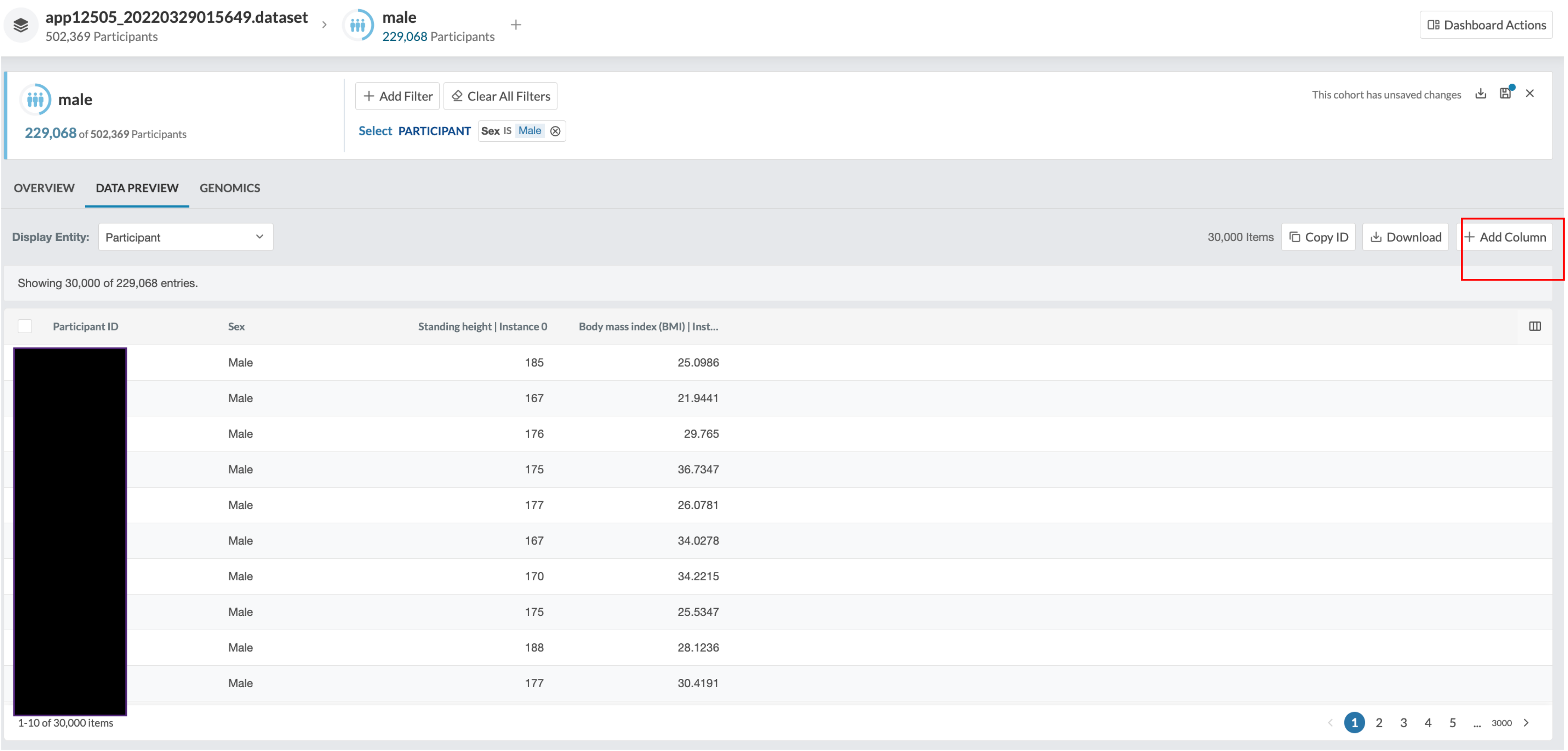
Item	Count / Freq.	
Male	229,068	100.00%
Female	0	0.00%

Standing height | Instanc...
227,681 Participants

Body mass index (BMI) | I...
227,421 Participants

A max of 15 tiles can be added to the dashboard.

Data Preview tab



Select PARTICIPANT **Sex** IS Male

Display Entity: Participant

30,000 Items

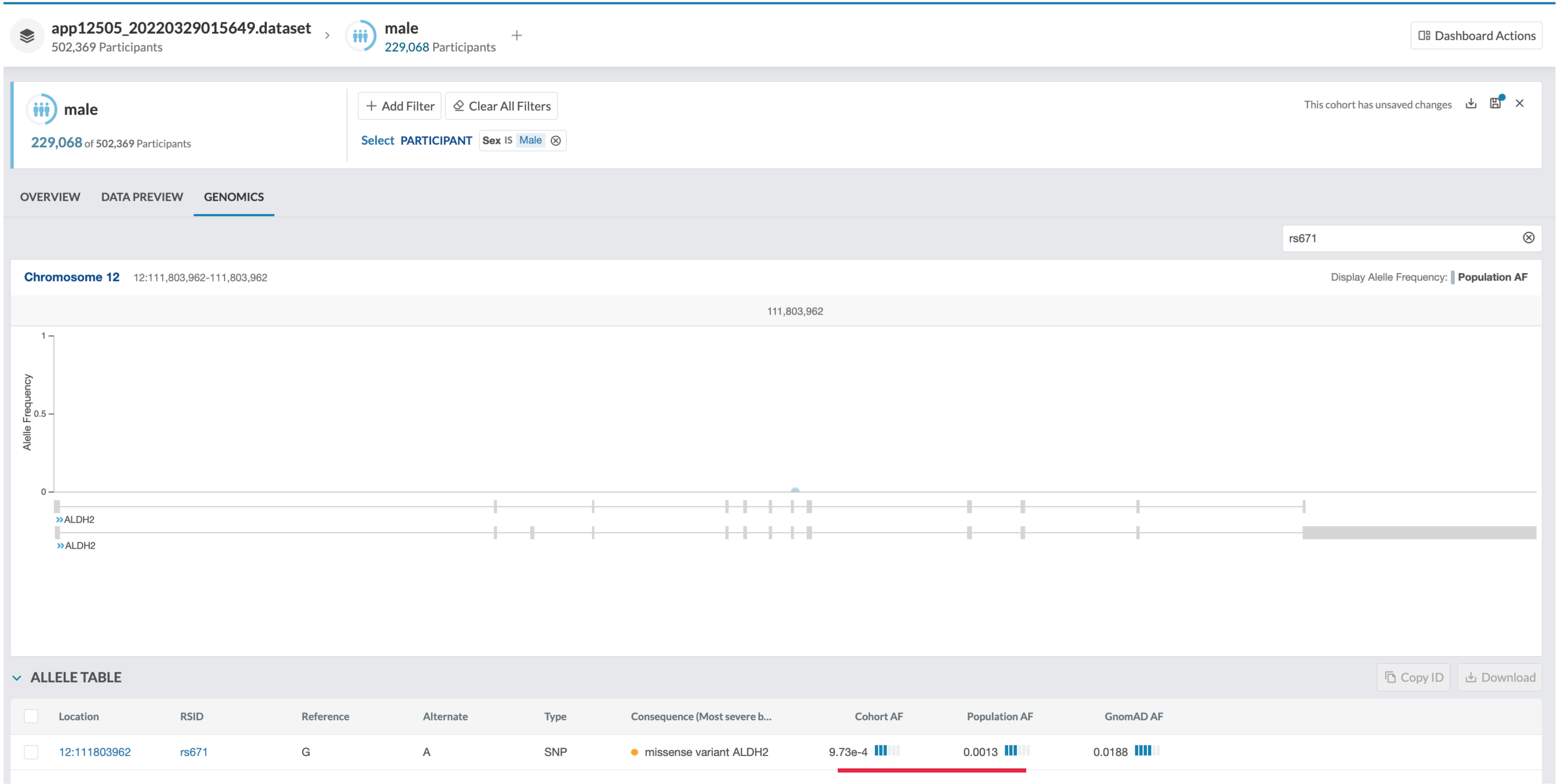
Showing 30,000 of 229,068 entries.

☐	Participant ID	Sex	Standing height | Instance 0	Body mass index (BMI) | Inst...
	[redacted]	Male	185	25.0986
	[redacted]	Male	167	21.9441
	[redacted]	Male	176	29.765
	[redacted]	Male	175	36.7347
	[redacted]	Male	177	26.0781
	[redacted]	Male	167	34.0278
	[redacted]	Male	170	34.2215
	[redacted]	Male	175	25.5347
	[redacted]	Male	188	28.1236
	[redacted]	Male	177	30.4191

1-10 of 30,000 items

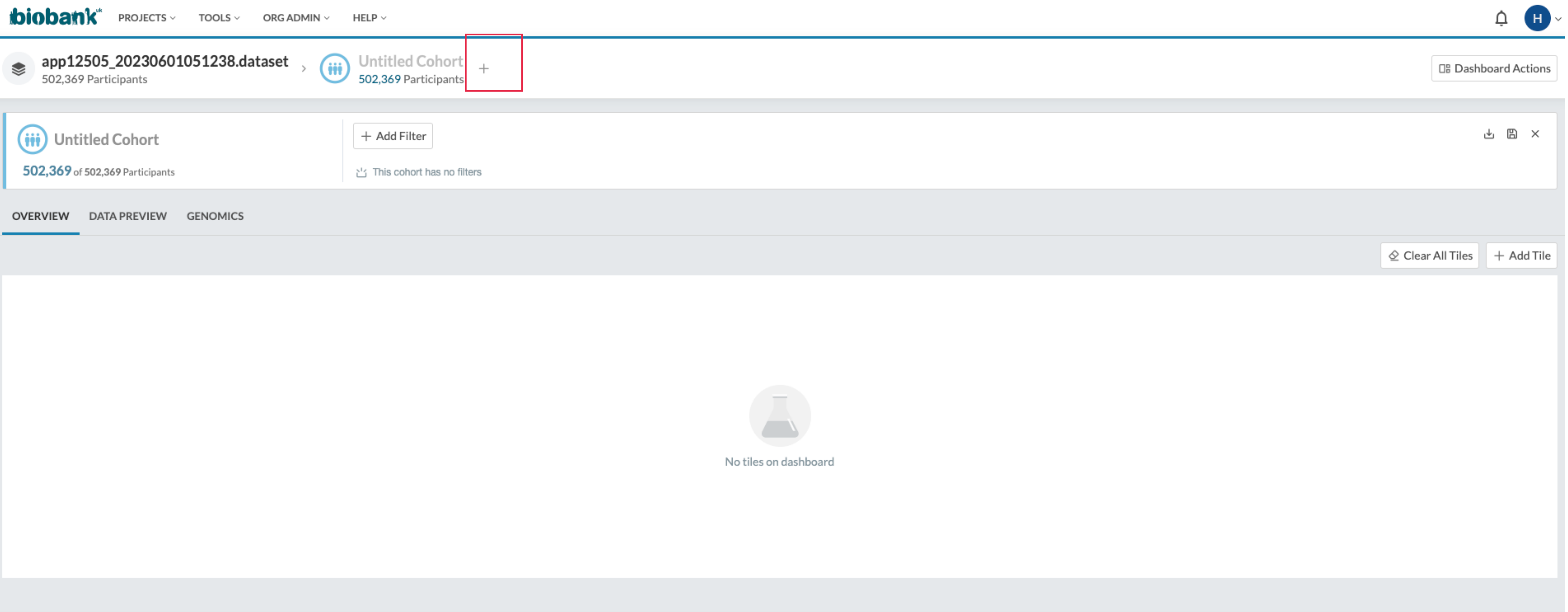
The Cohort Table can visualize up to 30 columns of data per tab.

Genomics

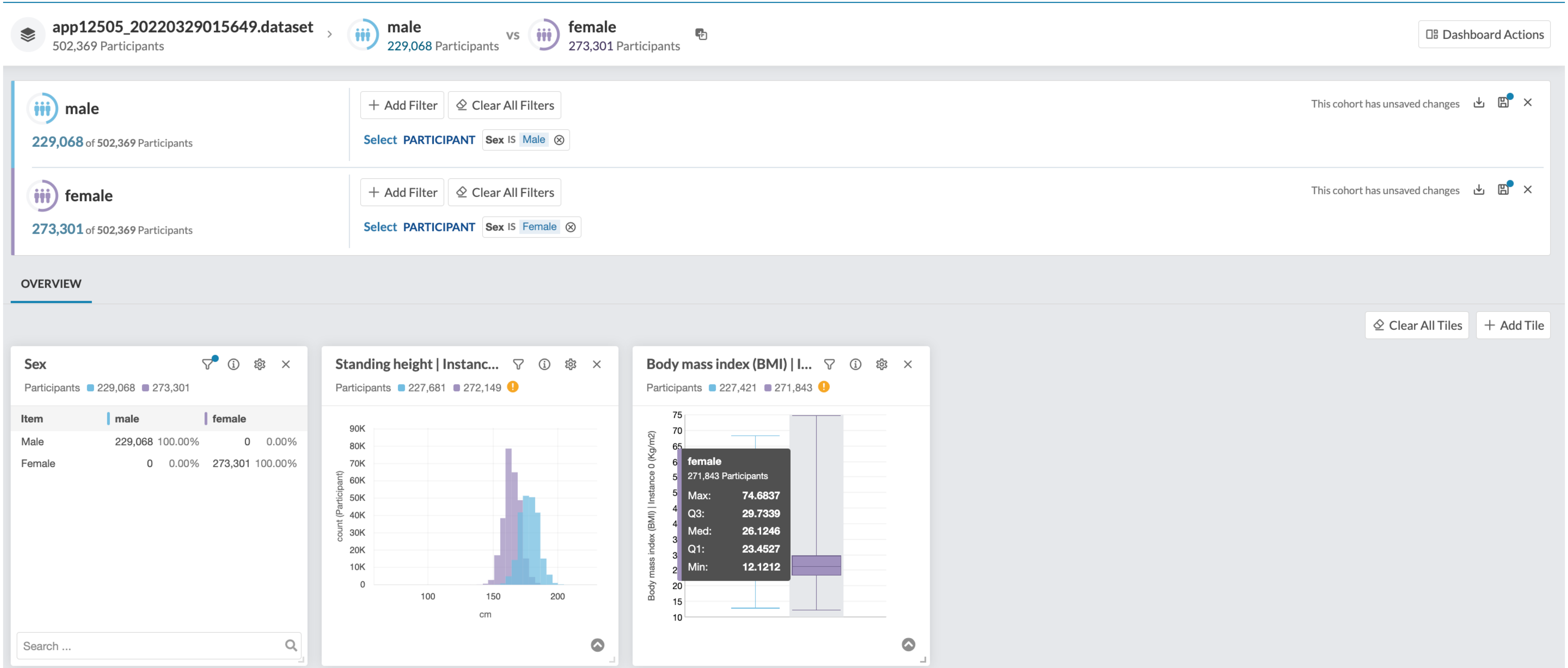


☐	Location	RSID	Reference	Alternate	Type	Consequence (Most severe b...	Cohort AF	Population AF	GnomAD AF
☐	12:111803962	rs671	G	A	SNP	● missense variant ALDH2	9.73e-4	0.0013	0.0188

Comparing Cohorts



Comparing Cohorts



app12505_20220329015649.dataset
502,369 Participants

male
229,068 Participants

vs

female
273,301 Participants

Dashboard Actions

male
229,068 of 502,369 Participants

Add Filter | Clear All Filters

Select PARTICIPANT **Sex** IS Male

This cohort has unsaved changes

female
273,301 of 502,369 Participants

Add Filter | Clear All Filters

Select PARTICIPANT **Sex** IS Female

This cohort has unsaved changes

OVERVIEW

Clear All Tiles | Add Tile

Sex

Participants 229,068 273,301

Item	male		female	
Male	229,068	100.00%	0	0.00%
Female	0	0.00%	273,301	100.00%

Search ...

Standing height | Instanc...

Participants 227,681 272,149

Body mass index (BMI) | I...

Participants 227,421 271,843

female
271,843 Participants

Max:	74.6837
Q3:	29.7339
Med:	26.1246
Q1:	23.4527
Min:	12.1212

https://documentation.dnanexus.com/user/cohort-browser#comparing-cohorts

Extract cohort data

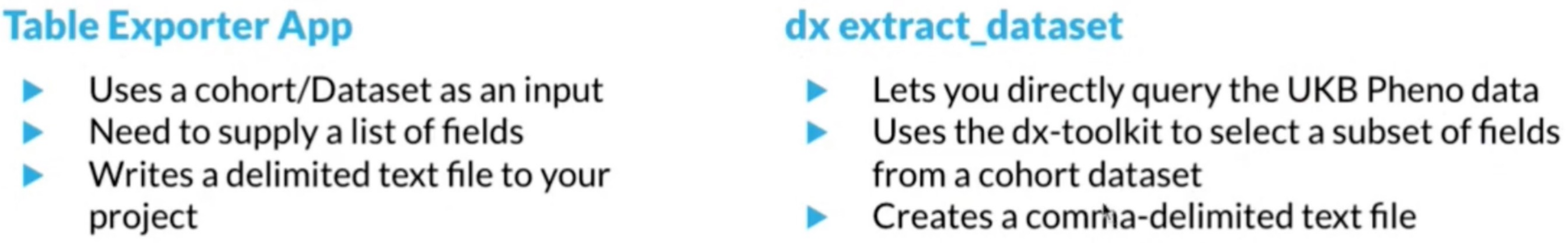


Table Exporter App

- Uses a cohort/Dataset as an input
- Need to supply a list of fields
- Writes a delimited text file to your project

dx extract_dataset

- Lets you directly query the UKB Pheno data
- Uses the dx-toolkit to select a subset of fields from a cohort dataset
- Creates a comma-delimited text file

Learning objects

- **How to set-up your RAP account**
- **How to explore the UKB data on RAP**

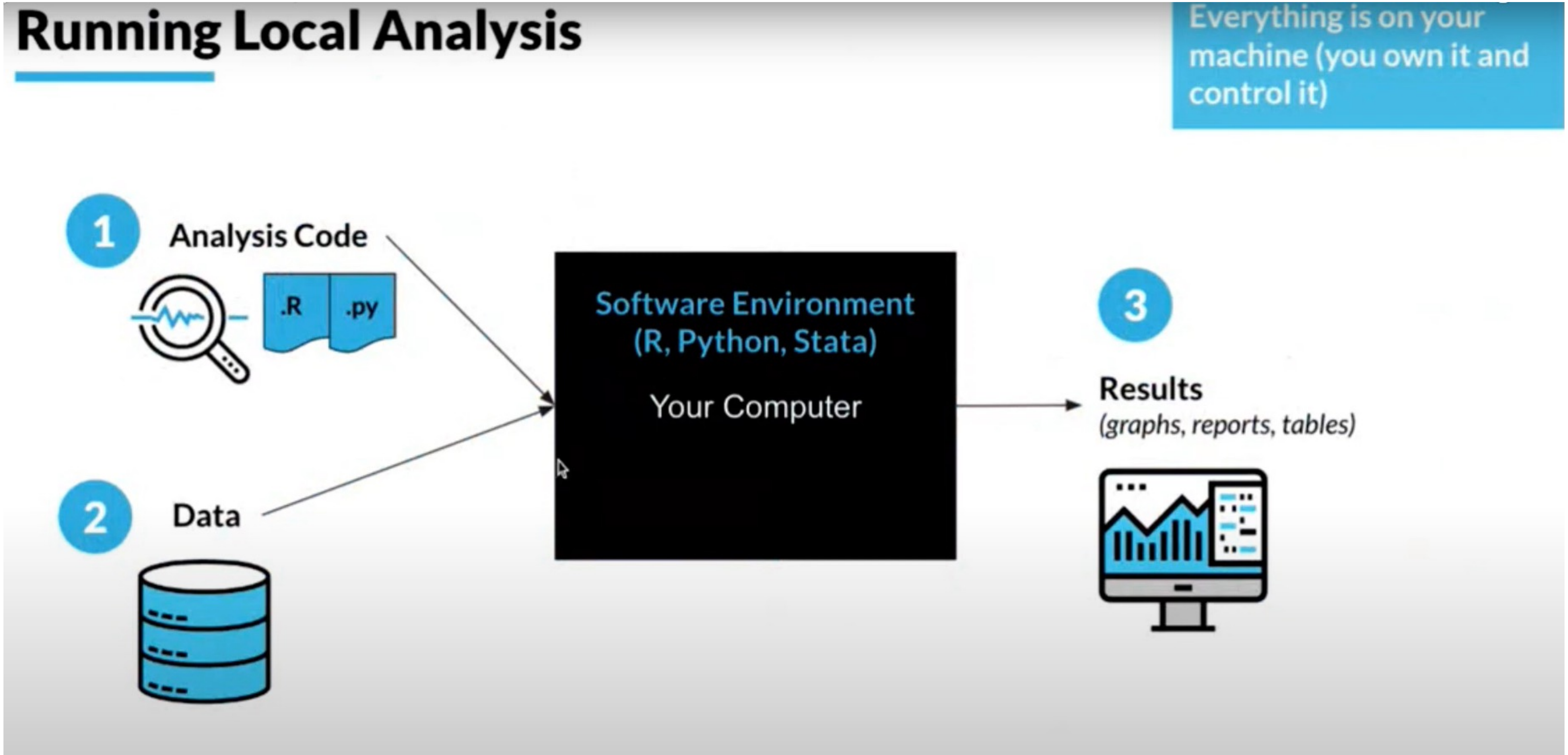
- What are the key differences between cloud- and cluster-based computing
- How to find existing apps on RAP
- How to submit a job on UI (user interface)
- How to install dx, the CLI (command line interface)
- How to submit a job using dx

Have a break

See you in 5 minutes

Cloud- vs cluster-computing

Cluster-computing



Running Local Analysis

Everything is on your machine (you own it and control it)

1 Analysis Code

2 Data

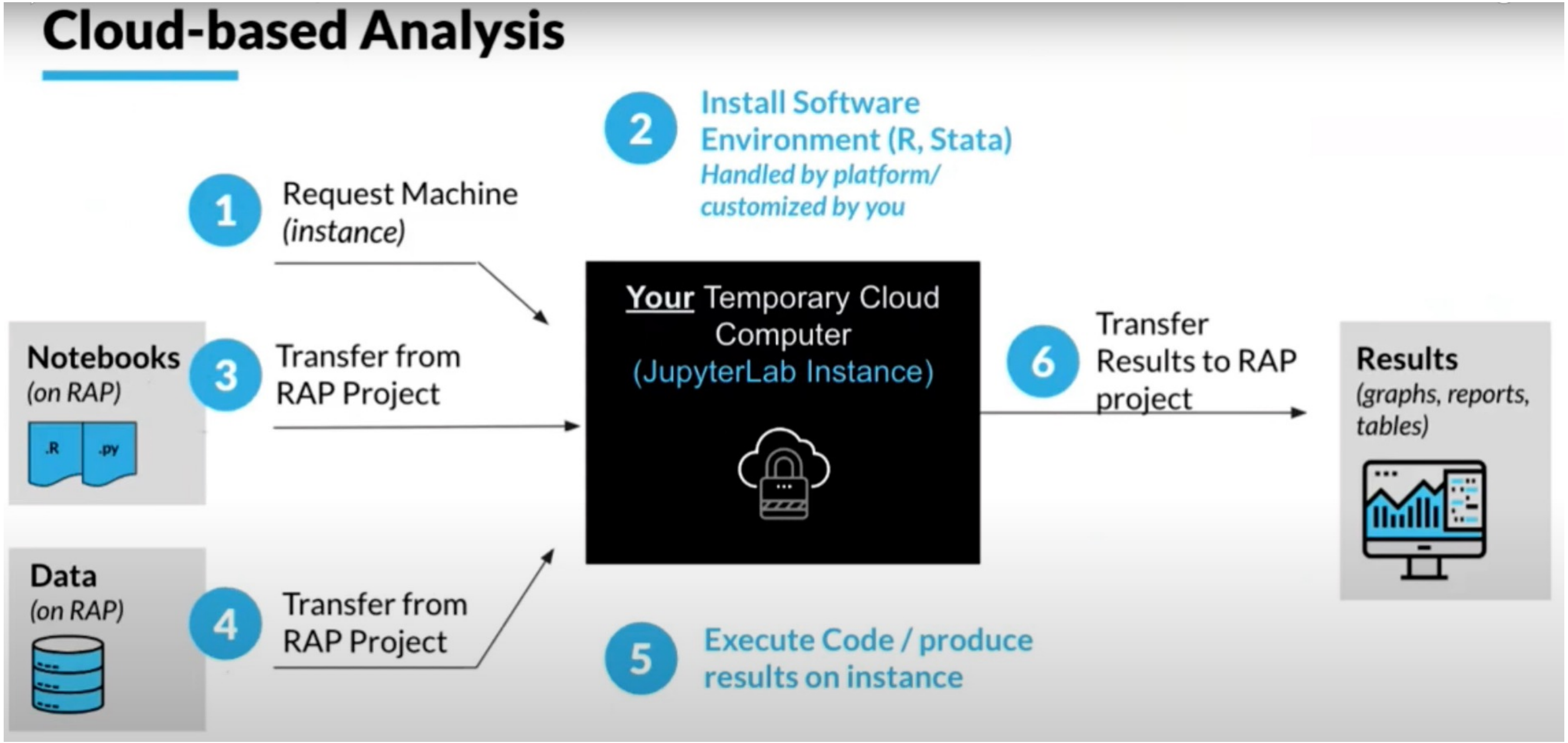
Software Environment (R, Python, Stata)

Your Computer

3 Results

(graphs, reports, tables)

Cloud-computing



Cloud-based Analysis

1. Request Machine *(instance)*

2. Install Software Environment (R, Stata)
Handled by platform/ customized by you

Notebooks
(on RAP)

3. Transfer from RAP Project

Data
(on RAP)

4. Transfer from RAP Project

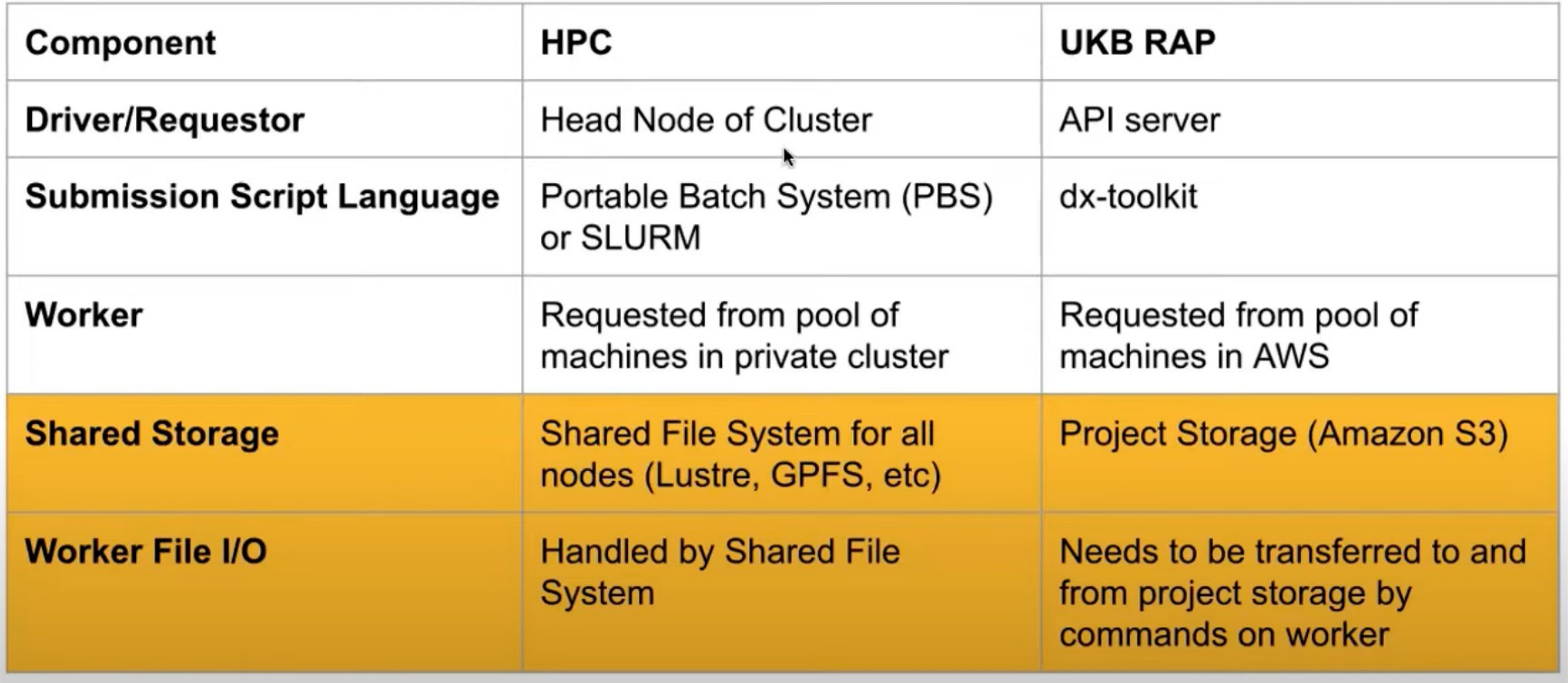
Your Temporary Cloud Computer
(JupyterLab Instance)

5. Execute Code / produce results on instance

6. Transfer Results to RAP project

Results
(graphs, reports, tables)

Similarities and differences



Component	HPC	UKB RAP
Driver/Requestor	Head Node of Cluster	API server
Submission Script Language	Portable Batch System (PBS) or SLURM	dx-toolkit
Worker	Requested from pool of machines in private cluster	Requested from pool of machines in AWS
Shared Storage	Shared File System for all nodes (Lustre, GPFS, etc)	Project Storage (Amazon S3)
Worker File I/O	Handled by Shared File System	Needs to be transferred to and from project storage by commands on worker

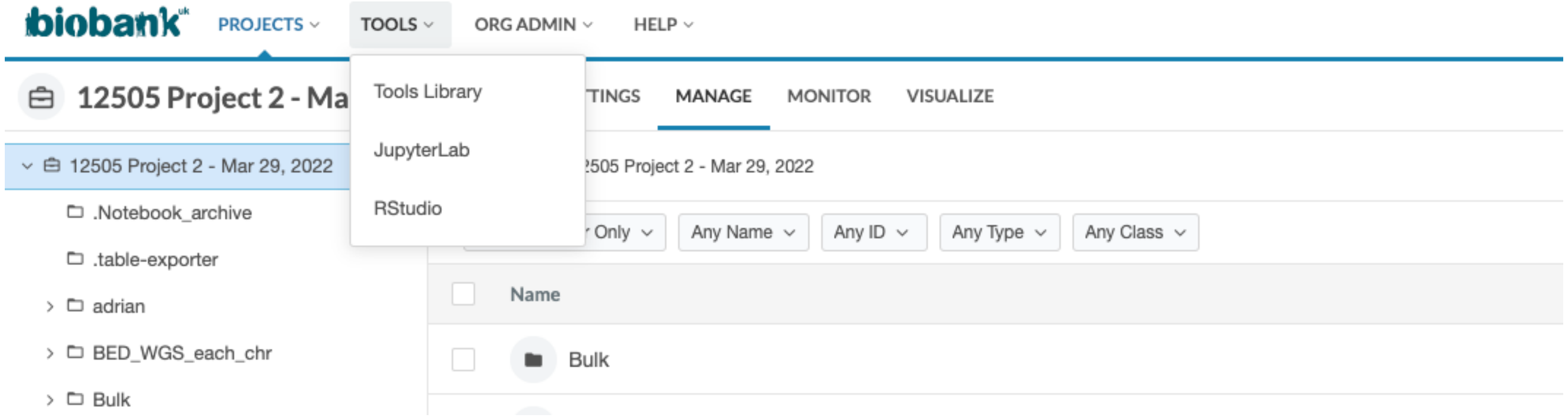
Tools

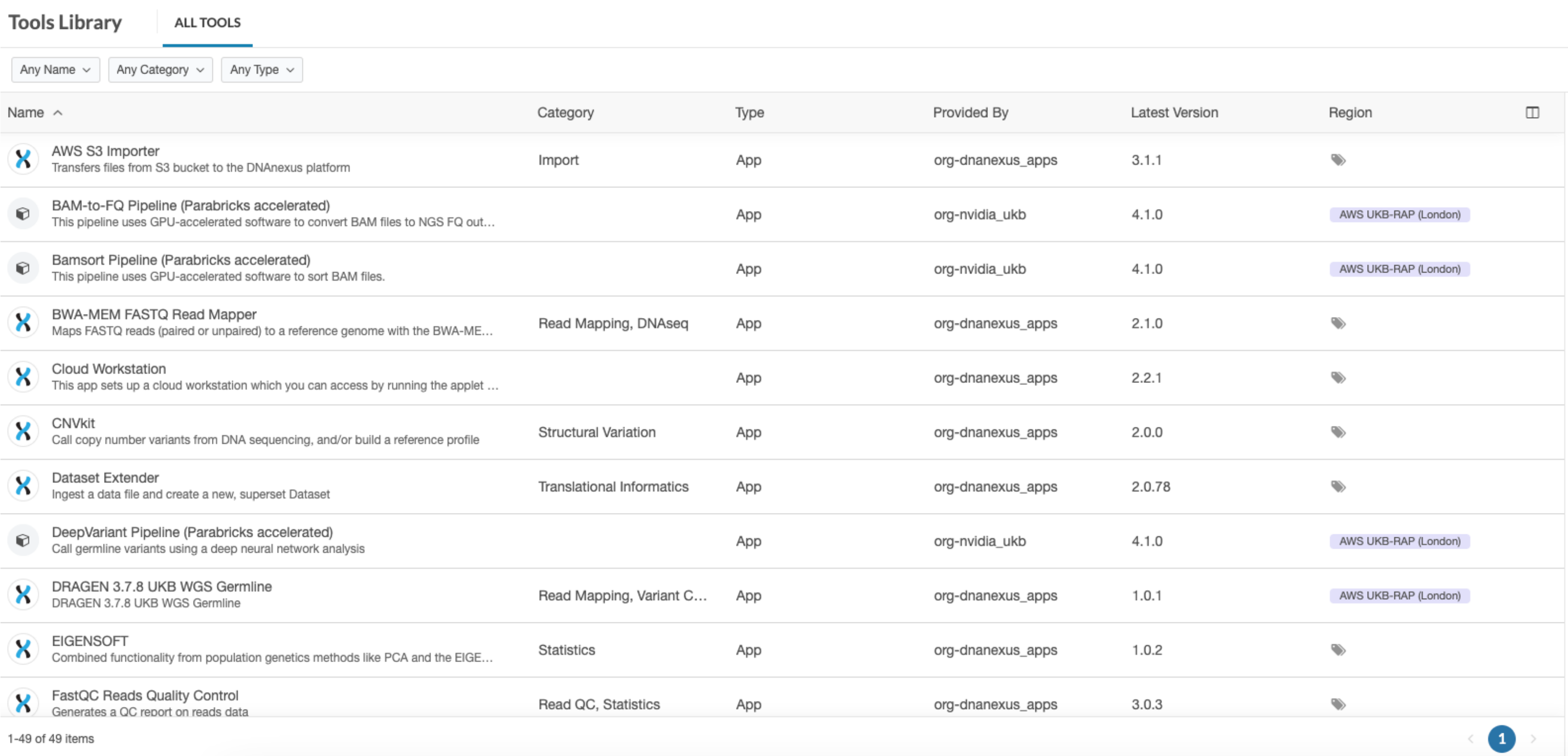
Tools

- Apps/Applets: executable data objects (software environment with scripts)
  - Applets: live inside the projects
  - App: do not live inside projects, and make it public to allow other users to run them in projects of their choosing.
- Workflows: apps/applets linked together
  - global workflow
- Tools: apps and workflows are generically referred to as "tools"

https://www.youtube.com/watch?v=lGxna-ZVYUg&t=167s
https://documentation.dnanexus.com/getting-started/key-concepts/apps-and-workflows

Tools library

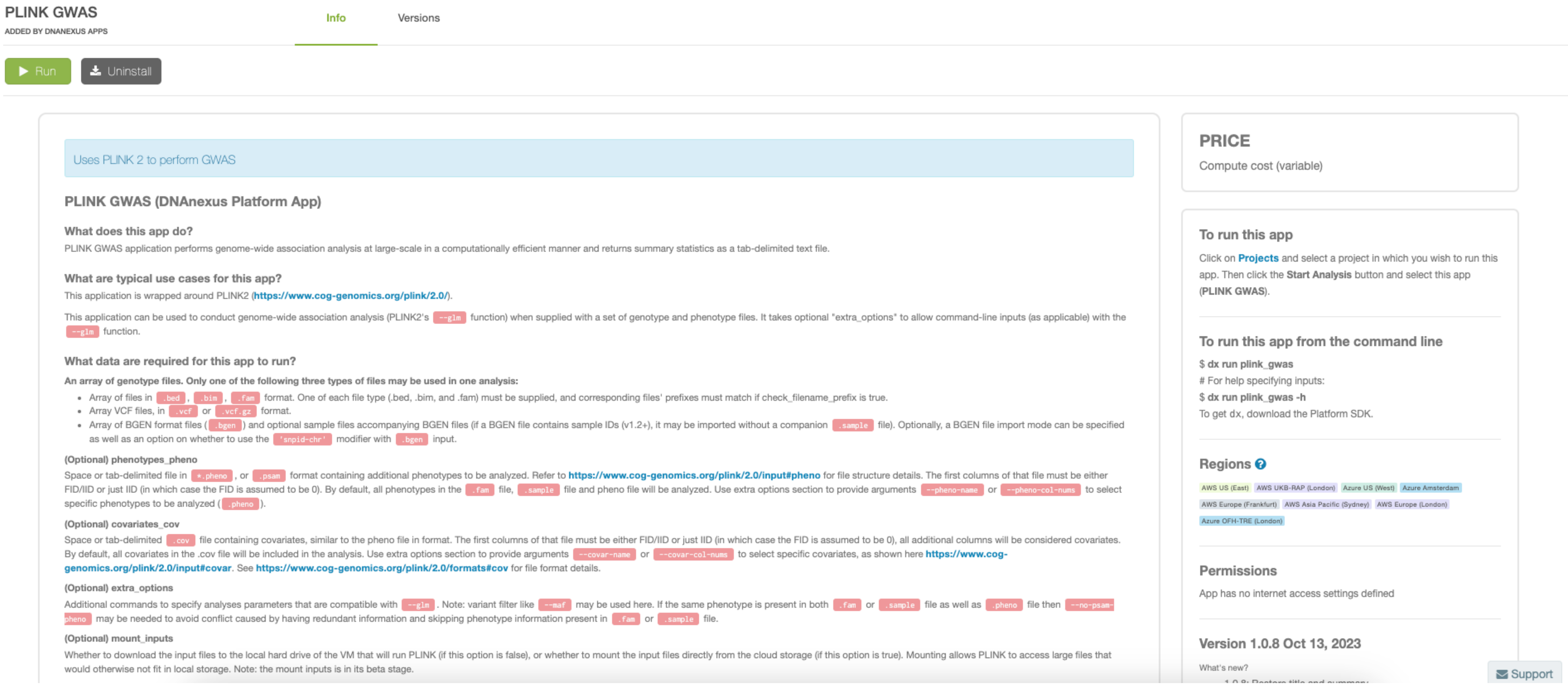




Name	Category	Type	Provided By	Latest Version	Region
AWS S3 Importer Transfers files from S3 bucket to the DNAnexus platform	Import	App	org-dnanexus_apps	3.1.1	
BAM-to-FQ Pipeline (Parabricks accelerated) This pipeline uses GPU-accelerated software to convert BAM files to NGS FQ out…		App	org-nvidia_ukb	4.1.0	AWS UKB-RAP (London)
Bamsort Pipeline (Parabricks accelerated) This pipeline uses GPU-accelerated software to sort BAM files.		App	org-nvidia_ukb	4.1.0	AWS UKB-RAP (London)
BWA-MEM FASTQ Read Mapper Maps FASTQ reads (paired or unpaired) to a reference genome with the BWA-ME…	Read Mapping, DNAseq	App	org-dnanexus_apps	2.1.0	
Cloud Workstation This app sets up a cloud workstation which you can access by running the applet …		App	org-dnanexus_apps	2.2.1	
CNVkit Call copy number variants from DNA sequencing, and/or build a reference profile	Structural Variation	App	org-dnanexus_apps	2.0.0	
Dataset Extender Ingest a data file and create a new, superset Dataset	Translational Informatics	App	org-dnanexus_apps	2.0.78	
DeepVariant Pipeline (Parabricks accelerated) Call germline variants using a deep neural network analysis		App	org-nvidia_ukb	4.1.0	AWS UKB-RAP (London)
DRAGEN 3.7.8 UKB WGS Germline DRAGEN 3.7.8 UKB WGS Germline	Read Mapping, Variant C…	App	org-dnanexus_apps	1.0.1	AWS UKB-RAP (London)
EIGENSOFT Combined functionality from population genetics methods like PCA and the EIGE…	Statistics	App	org-dnanexus_apps	1.0.2	
FastQC Reads Quality Control Generates a QC report on reads data	Read QC, Statistics	App	org-dnanexus_apps	3.0.3	

1-49 of 49 items

Info page



PLINK GWAS

ADDED BY DNANEXUS APPS

Uses PLINK 2 to perform GWAS

PLINK GWAS (DNAnexus Platform App)

What does this app do?

PLINK GWAS application performs genome-wide association analysis at large-scale in a computationally efficient manner and returns summary statistics as a tab-delimited text file.

What are typical use cases for this app?

This application is wrapped around PLINK2 (https://www.cog-genomics.org/plink/2.0/).

This application can be used to conduct genome-wide association analysis (PLINK2's `--glm` function) when supplied with a set of genotype and phenotype files. It takes optional "extra_options" to allow command-line inputs (as applicable) with the `--glm` function.

What data are required for this app to run?

An array of genotype files. Only one of the following three types of files may be used in one analysis:

- Array of files in `.bed`, `.bim`, `.fam` format. One of each file type (.bed, .bim, and .fam) must be supplied, and corresponding files' prefixes must match if check_filename_prefix is true.
- Array VCF files, in `.vcf` or `.vcf.gz` format.
- Array of BGEN format files (`.bgen`) and optional sample files accompanying BGEN files (if a BGEN file contains sample IDs (v1.2+), it may be imported without a companion `.sample` file). Optionally, a BGEN file import mode can be specified as well as an option on whether to use the `'snpid-chr'` modifier with `.bgen` input.

(Optional) phenotypes_pheno

Space or tab-delimited file in `*.pheno`, or `.psam` format containing additional phenotypes to be analyzed. Refer to https://www.cog-genomics.org/plink/2.0/input#pheno for file structure details. The first columns of that file must be either FID/IID or just IID (in which case the FID is assumed to be 0). By default, all phenotypes in the `.fam` file, `.sample` file and pheno file will be analyzed. Use extra options section to provide arguments `--pheno-name` or `--pheno-col-nums` to select specific phenotypes to be analyzed (`.pheno`).

(Optional) covariates_cov

Space or tab-delimited `.cov` file containing covariates, similar to the pheno file in format. The first columns of that file must be either FID/IID or just IID (in which case the FID is assumed to be 0), all additional columns will be considered covariates. By default, all covariates in the .cov file will be included in the analysis. Use extra options section to provide arguments `--covar-name` or `--covar-col-nums` to select specific covariates, as shown here https://www.cog-genomics.org/plink/2.0/input#covar. See https://www.cog-genomics.org/plink/2.0/formats#cov for file format details.

(Optional) extra_options

Additional commands to specify analyses parameters that are compatible with `--glm`. Note: variant filter like `--maf` may be used here. If the same phenotype is present in both `.fam` or `.sample` file as well as `.pheno` file then `--no-psam-pheno` may be needed to avoid conflict caused by having redundant information and skipping phenotype information present in `.fam` or `.sample` file.

(Optional) mount_inputs

Whether to download the input files to the local hard drive of the VM that will run PLINK (if this option is false), or whether to mount the input files directly from the cloud storage (if this option is true). Mounting allows PLINK to access large files that would otherwise not fit in local storage. Note: the mount inputs is in its beta stage.

PRICE

Compute cost (variable)

To run this app

Click on **Projects** and select a project in which you wish to run this app. Then click the **Start Analysis** button and select this app (**PLINK GWAS**).

To run this app from the command line

```
$ dx run plink_gwas
# For help specifying inputs:
$ dx run plink_gwas -h
```

To get dx, download the Platform SDK.

Regions ?

AWS US (East) | AWS UKB-RAP (London) | Azure US (West) | Azure Amsterdam | AWS Europe (Frankfurt) | AWS Asia Pacific (Sydney) | AWS Europe (London) | Azure OFH-TRE (London)

Permissions

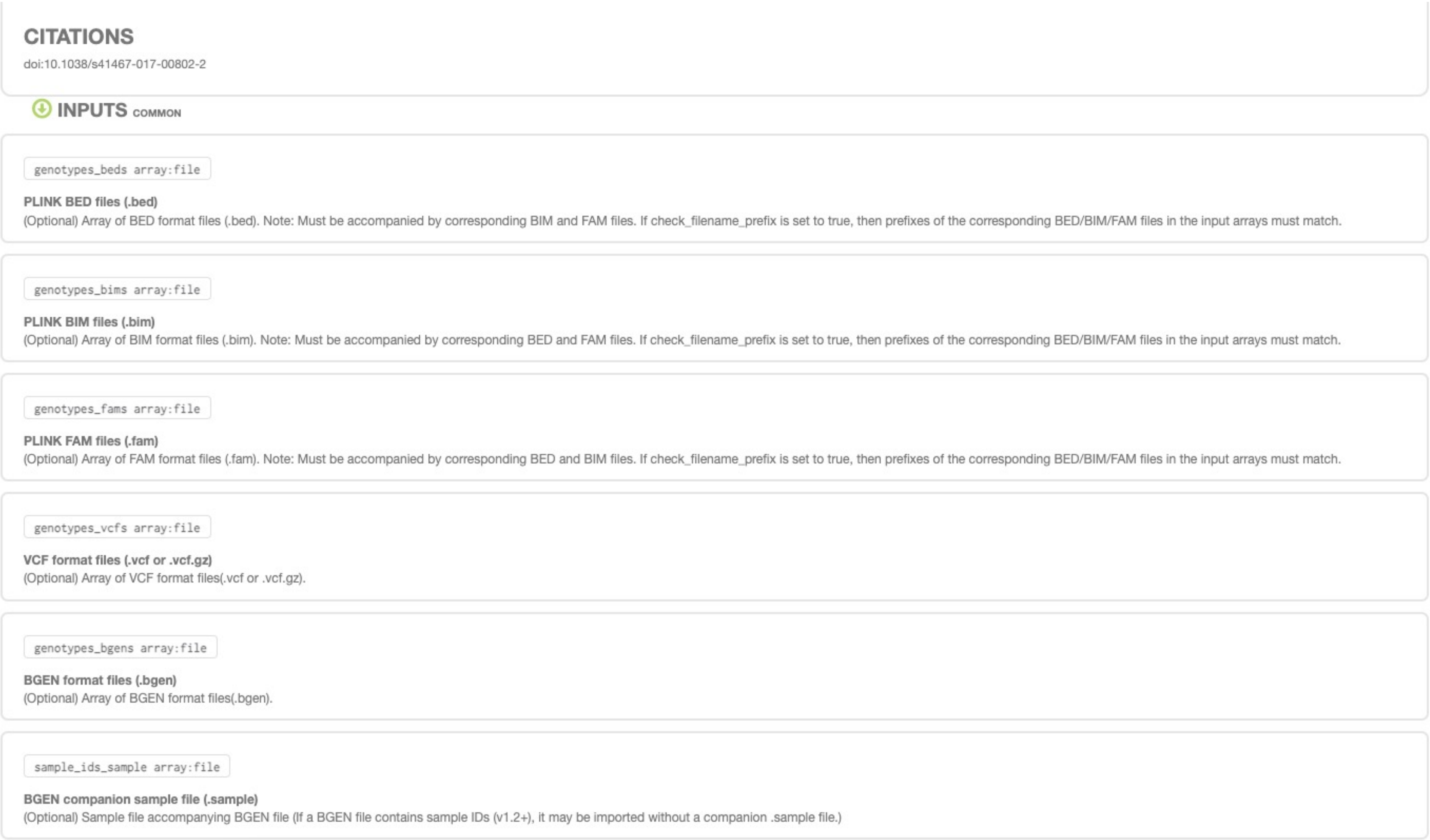
App has no internet access settings defined

Version 1.0.8 Oct 13, 2023

What's new?

https://ukbiobank.dnanexus.com/app/plink_gwas

Info page



CITATIONS

doi:10.1038/s41467-017-00802-2

INPUTS COMMON

`genotypes_beds array:file`

PLINK BED files (.bed)

(Optional) Array of BED format files (.bed). Note: Must be accompanied by corresponding BIM and FAM files. If check_filename_prefix is set to true, then prefixes of the corresponding BED/BIM/FAM files in the input arrays must match.

`genotypes_bims array:file`

PLINK BIM files (.bim)

(Optional) Array of BIM format files (.bim). Note: Must be accompanied by corresponding BED and FAM files. If check_filename_prefix is set to true, then prefixes of the corresponding BED/BIM/FAM files in the input arrays must match.

`genotypes_fams array:file`

PLINK FAM files (.fam)

(Optional) Array of FAM format files (.fam). Note: Must be accompanied by corresponding BED and BIM files. If check_filename_prefix is set to true, then prefixes of the corresponding BED/BIM/FAM files in the input arrays must match.

`genotypes_vcfs array:file`

VCF format files (.vcf or .vcf.gz)

(Optional) Array of VCF format files(.vcf or .vcf.gz).

`genotypes_bgens array:file`

BGEN format files (.bgen)

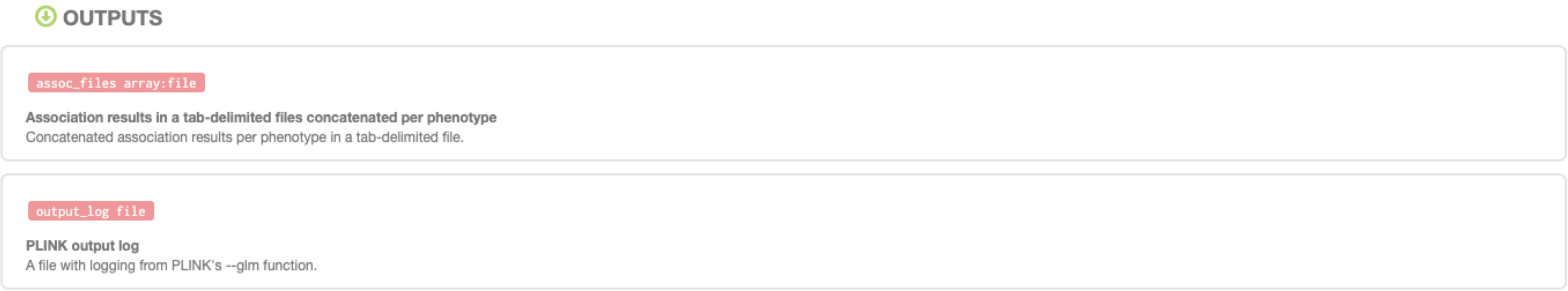
(Optional) Array of BGEN format files(.bgen).

`sample_ids_sample array:file`

BGEN companion sample file (.sample)

(Optional) Sample file accompanying BGEN file (If a BGEN file contains sample IDs (v1.2+), it may be imported without a companion .sample file.)

Info page



OUTPUTS

`assoc_files array:file`

Association results in a tab-delimited files concatenated per phenotype

Concatenated association results per phenotype in a tab-delimited file.

`output_log file`

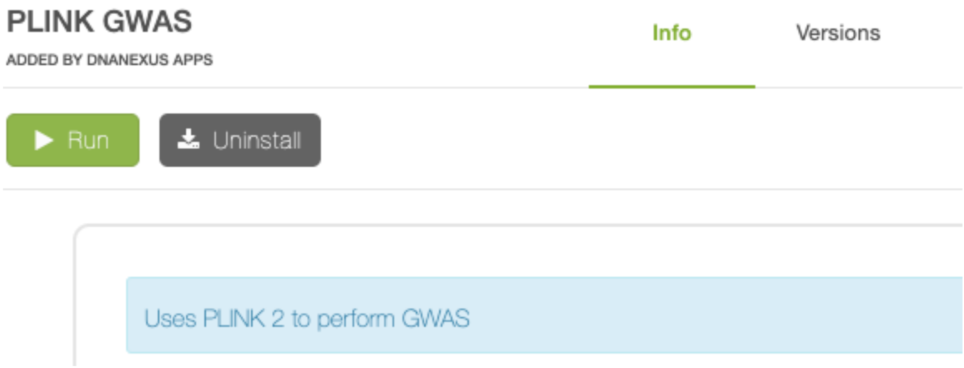
PLINK output log

A file with logging from PLINK's --glm function.

Submit a job on UI (user interface)

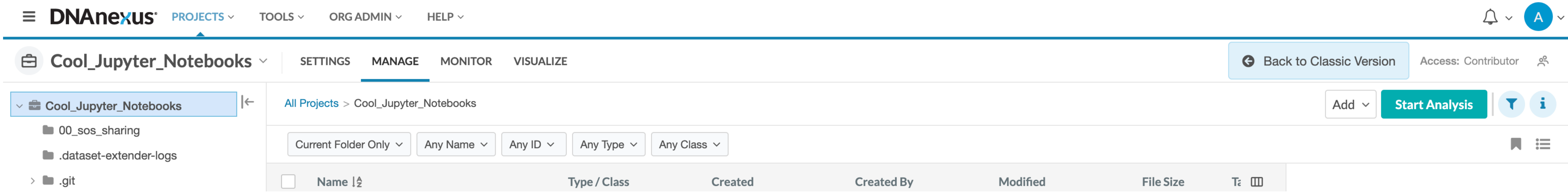
Run the tool

Launching from the Tools Library

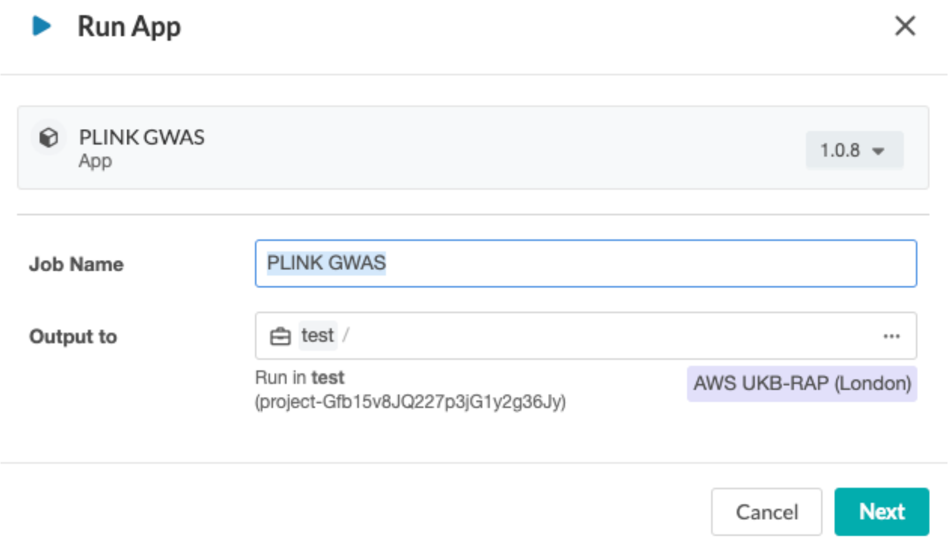


Uses PLINK 2 to perform GWAS

Launching from a Project



Configuration



PLINK GWAS
App
1.0.8

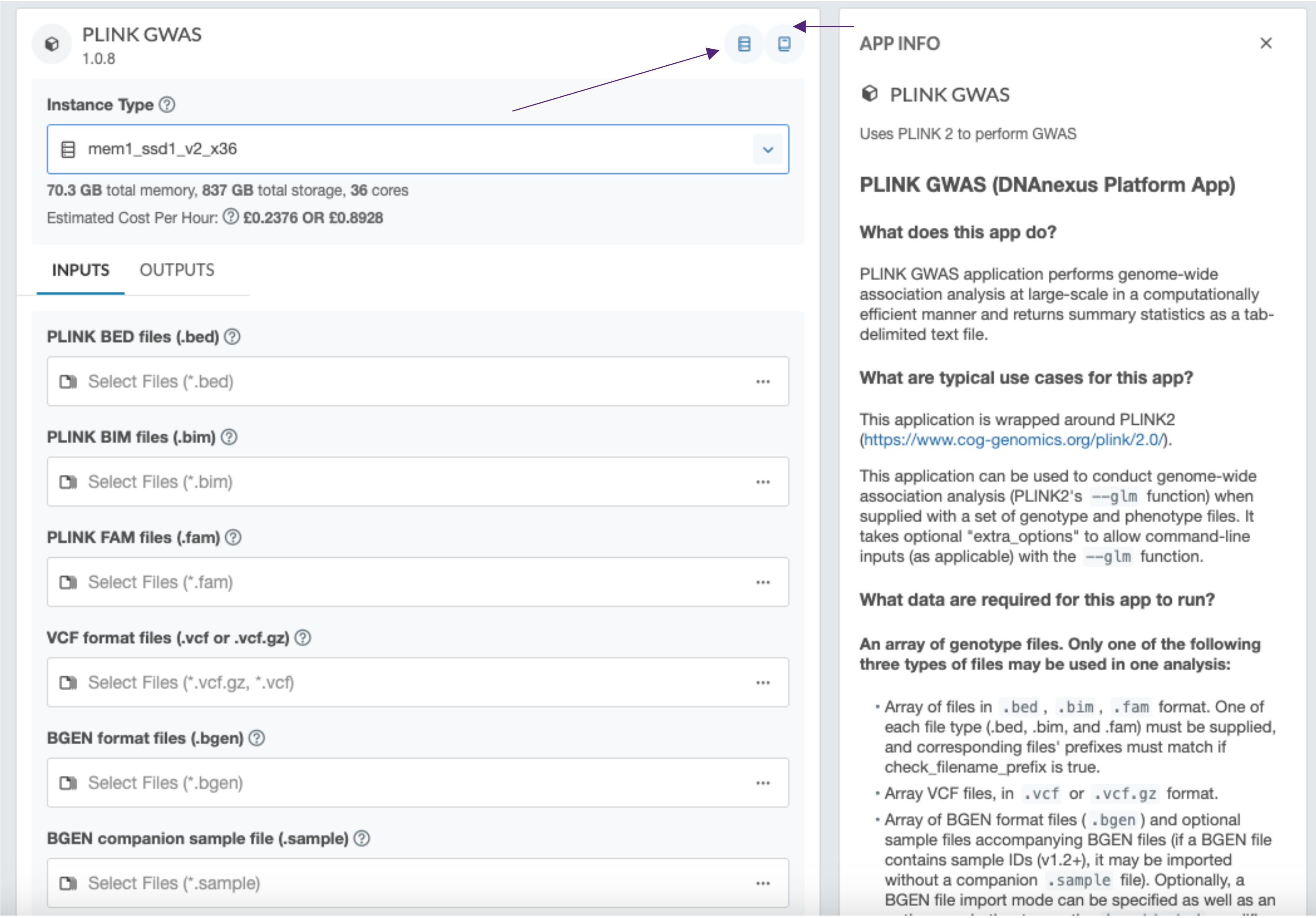
Job Name: PLINK GWAS

Output to: test /

Run in **test**
(project-Gfb15v8JQ227p3jG1y2g36Jy)

AWS UKB-RAP (London)

Inputs, APP INFO, Instance Type



PLINK GWAS

1.0.8

Instance Type

mem1_ssd1_v2_x36

70.3 GB total memory, **837 GB** total storage, **36** cores

Estimated Cost Per Hour: **£0.2376 OR £0.8928**

PLINK BED files (.bed)

Select Files (*.bed)

PLINK BIM files (.bim)

Select Files (*.bim)

PLINK FAM files (.fam)

Select Files (*.fam)

VCF format files (.vcf or .vcf.gz)

Select Files (*.vcf.gz, *.vcf)

BGEN format files (.bgen)

Select Files (*.bgen)

BGEN companion sample file (.sample)

Select Files (*.sample)

APP INFO

PLINK GWAS

Uses PLINK 2 to perform GWAS

PLINK GWAS (DNAnexus Platform App)

What does this app do?

PLINK GWAS application performs genome-wide association analysis at large-scale in a computationally efficient manner and returns summary statistics as a tab-delimited text file.

What are typical use cases for this app?

This application is wrapped around PLINK2 (https://www.cog-genomics.org/plink/2.0/).

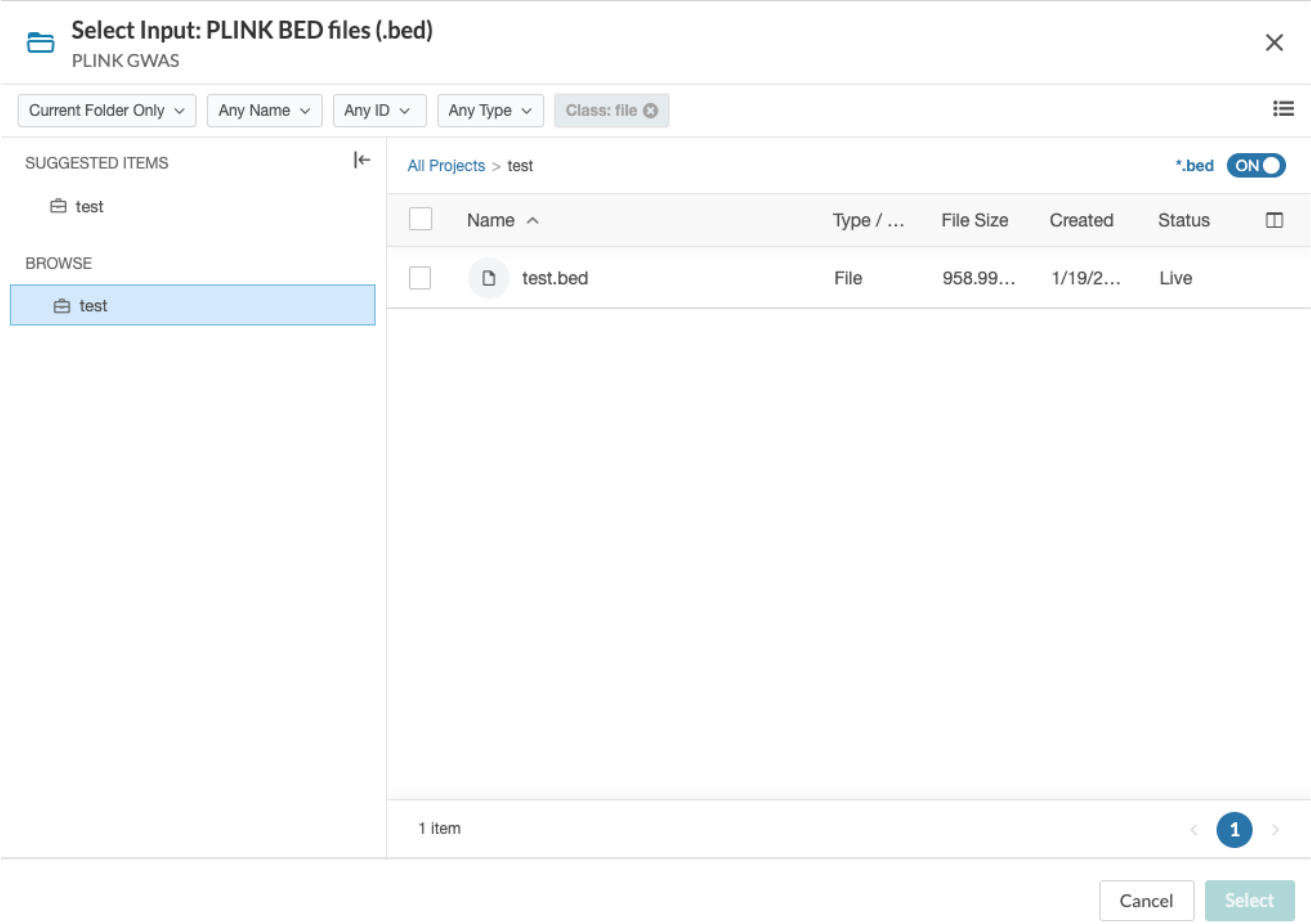
This application can be used to conduct genome-wide association analysis (PLINK2's `--glm` function) when supplied with a set of genotype and phenotype files. It takes optional "extra_options" to allow command-line inputs (as applicable) with the `--glm` function.

What data are required for this app to run?

An array of genotype files. Only one of the following three types of files may be used in one analysis:

- Array of files in `.bed`, `.bim`, `.fam` format. One of each file type (.bed, .bim, and .fam) must be supplied, and corresponding files' prefixes must match if check_filename_prefix is true.
- Array VCF files, in `.vcf` or `.vcf.gz` format.
- Array of BGEN format files (`.bgen`) and optional sample files accompanying BGEN files (if a BGEN file contains sample IDs (v1.2+), it may be imported without a companion `.sample` file). Optionally, a BGEN file import mode can be specified as well as an

Select Input File



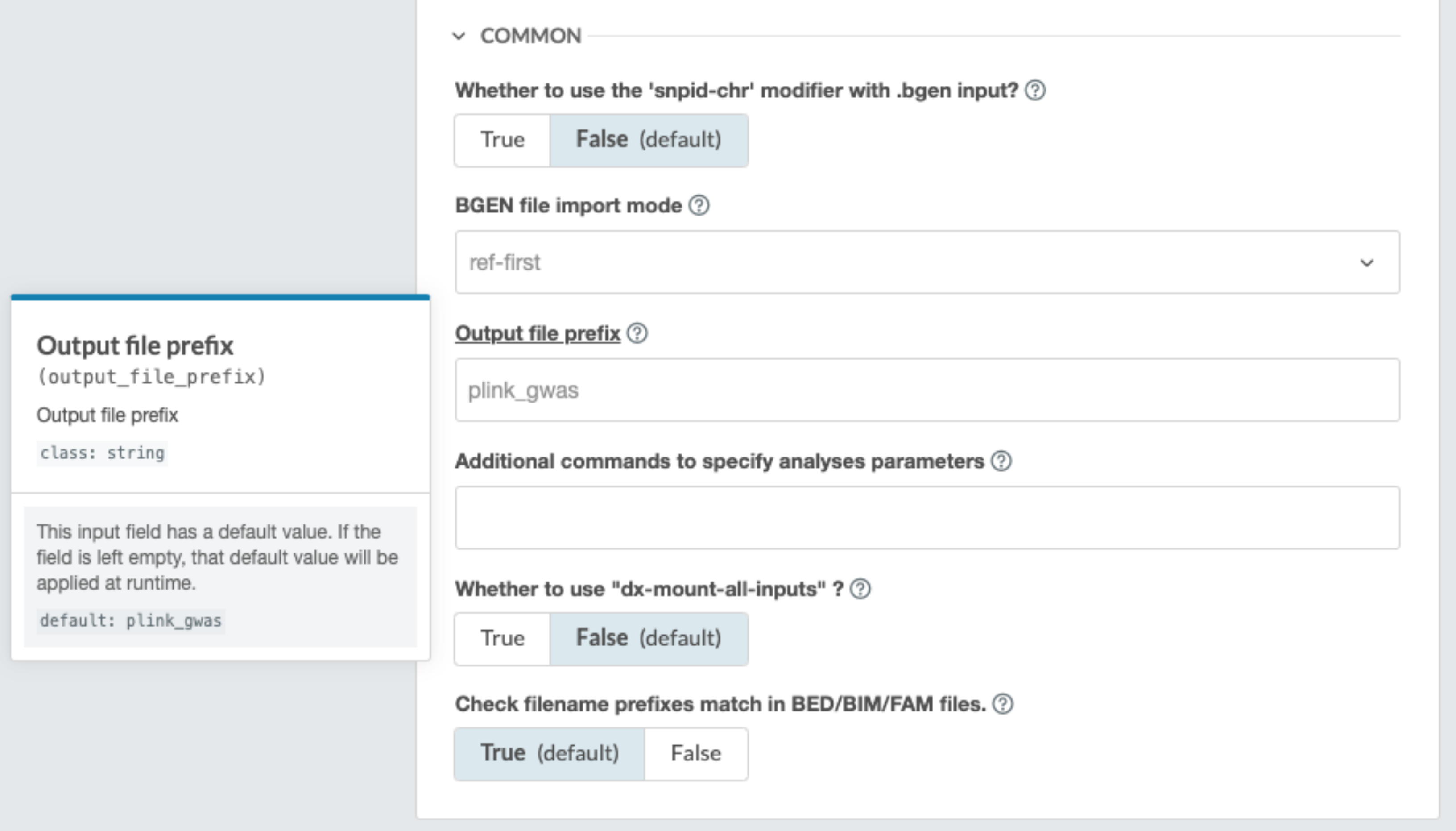
Select Input: PLINK BED files (.bed)

PLINK GWAS

☐	Name	Type / …	File Size	Created	Status
☐	test.bed	File	958.99…	1/19/2…	Live

1 item

Common parameters



COMMON

Whether to use the 'snpid-chr' modifier with .bgen input?

True | **False** (default)

BGEN file import mode

ref-first

Output file prefix

plink_gwas

Additional commands to specify analyses parameters

Whether to use "dx-mount-all-inputs" ?

True | **False** (default)

Check filename prefixes match in BED/BIM/FAM files.

True (default) | False

Output file prefix

`(output_file_prefix)`

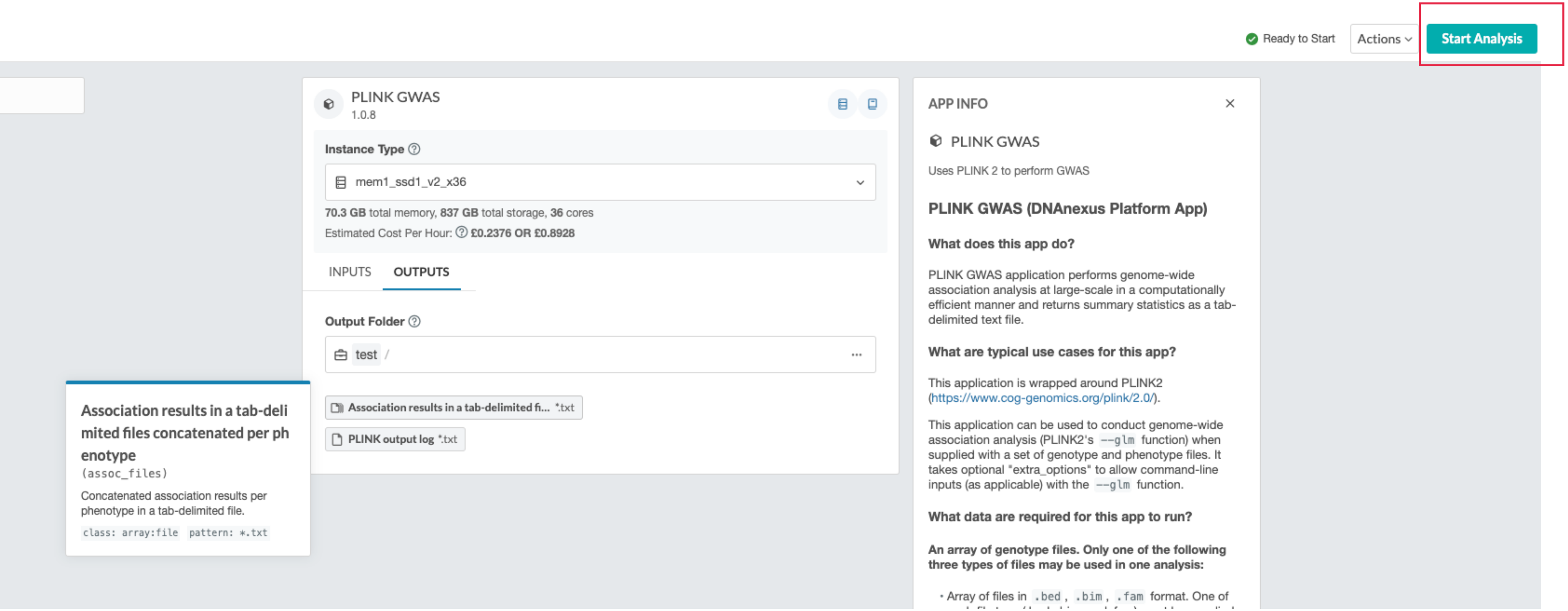
Output file prefix

`class: string`

This input field has a default value. If the field is left empty, that default value will be applied at runtime.

`default: plink_gwas`

Outputs



PLINK GWAS
1.0.8

Instance Type

mem1_ssd1_v2_x36

70.3 GB total memory, **837 GB** total storage, **36** cores

Estimated Cost Per Hour: **£0.2376 OR £0.8928**

Output Folder

test /

Association results in a tab-delimited fi... *.txt

PLINK output log *.txt

Association results in a tab-delimited files concatenated per phenotype
`(assoc_files)`

Concatenated association results per phenotype in a tab-delimited file.

`class: array:file` `pattern: *.txt`

APP INFO

PLINK GWAS

Uses PLINK 2 to perform GWAS

PLINK GWAS (DNAnexus Platform App)

What does this app do?

PLINK GWAS application performs genome-wide association analysis at large-scale in a computationally efficient manner and returns summary statistics as a tab-delimited text file.

What are typical use cases for this app?

This application is wrapped around PLINK2 (https://www.cog-genomics.org/plink/2.0/).

This application can be used to conduct genome-wide association analysis (PLINK2's `--glm` function) when supplied with a set of genotype and phenotype files. It takes optional "extra_options" to allow command-line inputs (as applicable) with the `--glm` function.

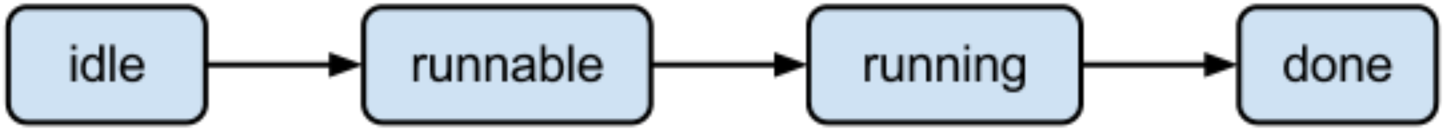
What data are required for this app to run?

An array of genotype files. Only one of the following three types of files may be used in one analysis:

- Array of files in `.bed`, `.bim`, `.fam` format. One of

Monitor your job

Job lifecycle



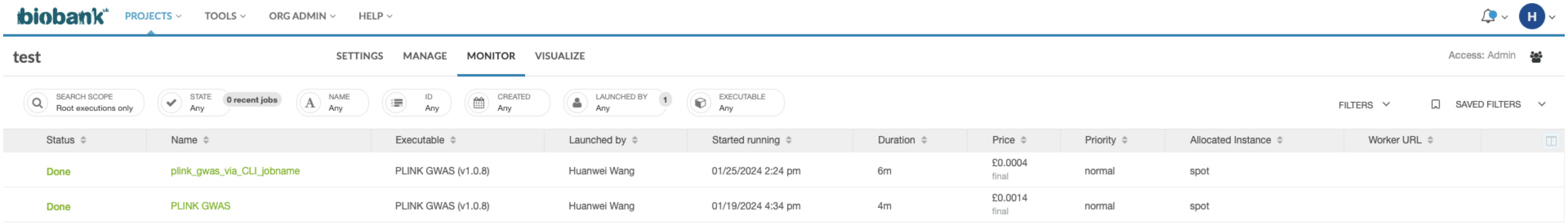
idle

runnable

running

done

Monitor page



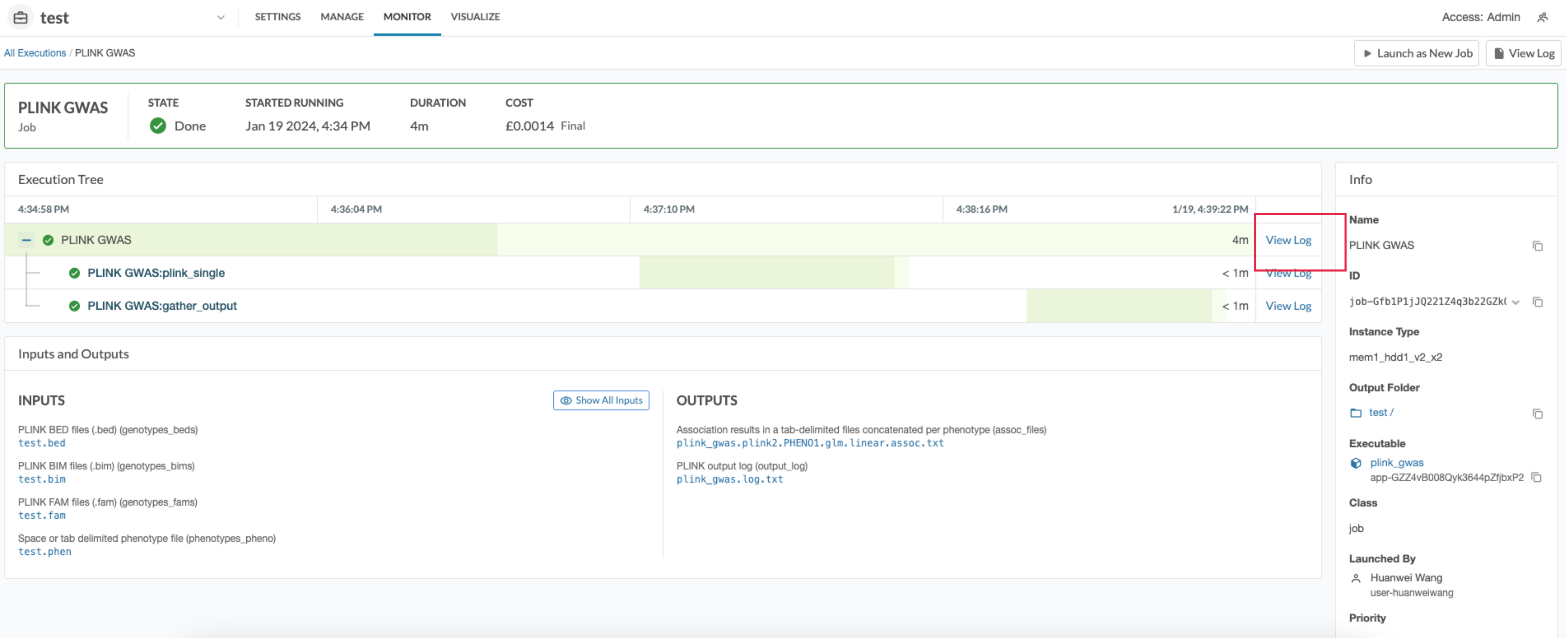
test

Access: Admin

SEARCH SCOPE: Root executions only
STATE: Any
0 recent jobs
NAME: Any
ID: Any
CREATED: Any
LAUNCHED BY: Any 1
EXECUTABLE: Any

Status	Name	Executable	Launched by	Started running	Duration	Price	Priority	Allocated Instance	Worker URL
Done	plink_gwas_via_CLI_jobname	PLINK GWAS (v1.0.8)	Huanwei Wang	01/25/2024 2:24 pm	6m	£0.0004 final	normal	spot	
Done	PLINK GWAS	PLINK GWAS (v1.0.8)	Huanwei Wang	01/19/2024 4:34 pm	4m	£0.0014 final	normal	spot	

Monitor a job



PLINK GWAS
Job

STATE	STARTED RUNNING	DURATION	COST
Done	Jan 19 2024, 4:34 PM	4m	£0.0014 Final

Execution Tree

4:34:58 PM	4:36:04 PM	4:37:10 PM	4:38:16 PM	1/19, 4:39:22 PM	
PLINK GWAS				4m	View Log
PLINK GWAS:plink_single				< 1m	View Log
PLINK GWAS:gather_output				< 1m	View Log

Inputs and Outputs

INPUTS

PLINK BED files (.bed) (genotypes_beds)
`test.bed`

PLINK BIM files (.bim) (genotypes_bims)
`test.bim`

PLINK FAM files (.fam) (genotypes_fams)
`test.fam`

Space or tab delimited phenotype file (phenotypes_pheno)
`test.phen`

OUTPUTS

Association results in a tab-delimited files concatenated per phenotype (assoc_files)
`plink_gwas.plink2.PHENO1.glm.linear.assoc.txt`

PLINK output log (output_log)
`plink_gwas.log.txt`

Info

Name
PLINK GWAS

ID
`job-Gfb1P1jJQ221Z4q3b22GZk(`

Instance Type
mem1_hdd1_v2_x2

Output Folder
test /

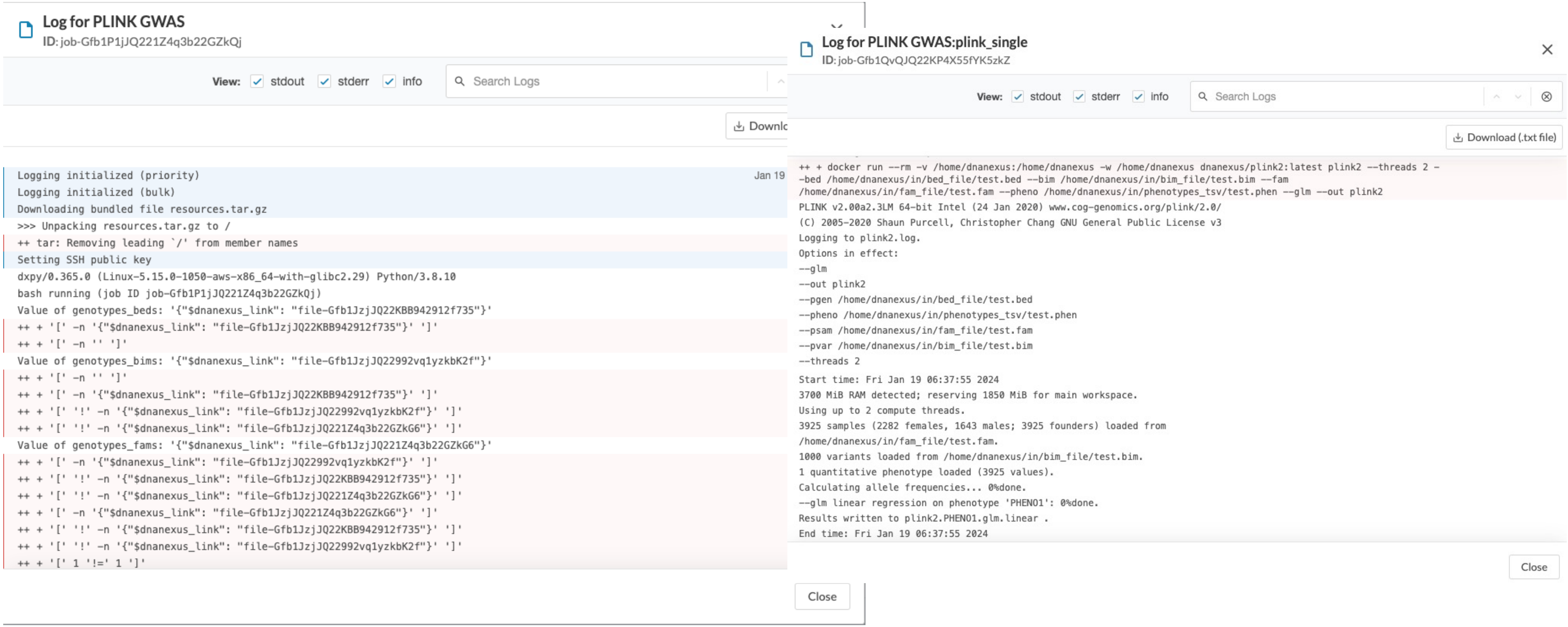
Executable
plink_gwas
app-GZZ4vB008Qyk3644pZfjbxP2

Class
job

Launched By
Huanwei Wang
user-huanweiwang

Priority

View log



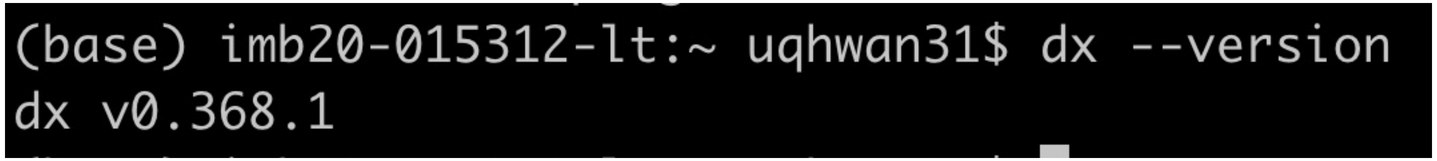
```
Logging initialized (priority)
Logging initialized (bulk)
Downloading bundled file resources.tar.gz
>>> Unpacking resources.tar.gz to /
++ tar: Removing leading `/' from member names
Setting SSH public key
dxpy/0.365.0 (Linux-5.15.0-1050-aws-x86_64-with-glibc2.29) Python/3.8.10
bash running (job ID job-Gfb1P1jJQ221Z4q3b22GZkQj)
Value of genotypes_beds: '{"$dnanexus_link": "file-Gfb1JzjJQ22KBB942912f735"}'
++ + '[' -n '{"$dnanexus_link": "file-Gfb1JzjJQ22KBB942912f735"}' ']'
++ + '[' -n '' ']'
Value of genotypes_bims: '{"$dnanexus_link": "file-Gfb1JzjJQ22992vq1yzkbK2f"}'
++ + '[' -n '' ']'
++ + '[' -n '{"$dnanexus_link": "file-Gfb1JzjJQ22KBB942912f735"}' ']'
++ + '[' '!' -n '{"$dnanexus_link": "file-Gfb1JzjJQ22992vq1yzkbK2f"}' ']'
++ + '[' '!' -n '{"$dnanexus_link": "file-Gfb1JzjJQ221Z4q3b22GZkG6"}' ']'
Value of genotypes_fams: '{"$dnanexus_link": "file-Gfb1JzjJQ221Z4q3b22GZkG6"}'
++ + '[' -n '{"$dnanexus_link": "file-Gfb1JzjJQ22992vq1yzkbK2f"}' ']'
++ + '[' '!' -n '{"$dnanexus_link": "file-Gfb1JzjJQ22KBB942912f735"}' ']'
++ + '[' '!' -n '{"$dnanexus_link": "file-Gfb1JzjJQ221Z4q3b22GZkG6"}' ']'
++ + '[' -n '{"$dnanexus_link": "file-Gfb1JzjJQ221Z4q3b22GZkG6"}' ']'
++ + '[' '!' -n '{"$dnanexus_link": "file-Gfb1JzjJQ22KBB942912f735"}' ']'
++ + '[' '!' -n '{"$dnanexus_link": "file-Gfb1JzjJQ22992vq1yzkbK2f"}' ']'
++ + '[' 1 '!=' 1 ']'
```

```
++ + docker run --rm -v /home/dnanexus:/home/dnanexus -w /home/dnanexus dnanexus/plink2:latest plink2 --threads 2 -
-bed /home/dnanexus/in/bed_file/test.bed --bim /home/dnanexus/in/bim_file/test.bim --fam
/home/dnanexus/in/fam_file/test.fam --pheno /home/dnanexus/in/phenotypes_tsv/test.phen --glm --out plink2
PLINK v2.00a2.3LM 64-bit Intel (24 Jan 2020) www.cog-genomics.org/plink/2.0/
(C) 2005-2020 Shaun Purcell, Christopher Chang GNU General Public License v3
Logging to plink2.log.
Options in effect:
--glm
--out plink2
--pgen /home/dnanexus/in/bed_file/test.bed
--pheno /home/dnanexus/in/phenotypes_tsv/test.phen
--psam /home/dnanexus/in/fam_file/test.fam
--pvar /home/dnanexus/in/bim_file/test.bim
--threads 2
Start time: Fri Jan 19 06:37:55 2024
3700 MiB RAM detected; reserving 1850 MiB for main workspace.
Using up to 2 compute threads.
3925 samples (2282 females, 1643 males; 3925 founders) loaded from
/home/dnanexus/in/fam_file/test.fam.
1000 variants loaded from /home/dnanexus/in/bim_file/test.bim.
1 quantitative phenotype loaded (3925 values).
Calculating allele frequencies... 0%done.
--glm linear regression on phenotype 'PHENO1': 0%done.
Results written to plink2.PHENO1.glm.linear .
End time: Fri Jan 19 06:37:55 2024
```

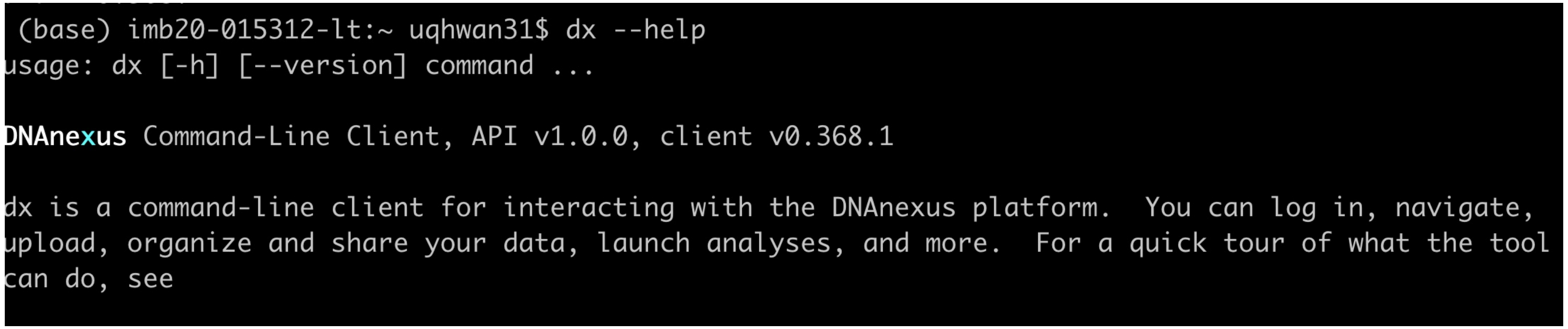
Command line interface (CLI)

Installation

- dx: command-line client
- via PyPI/PIP3
  - install: “pip3 install dxpy”
  - update: “pip3 install --upgrade dxpy”



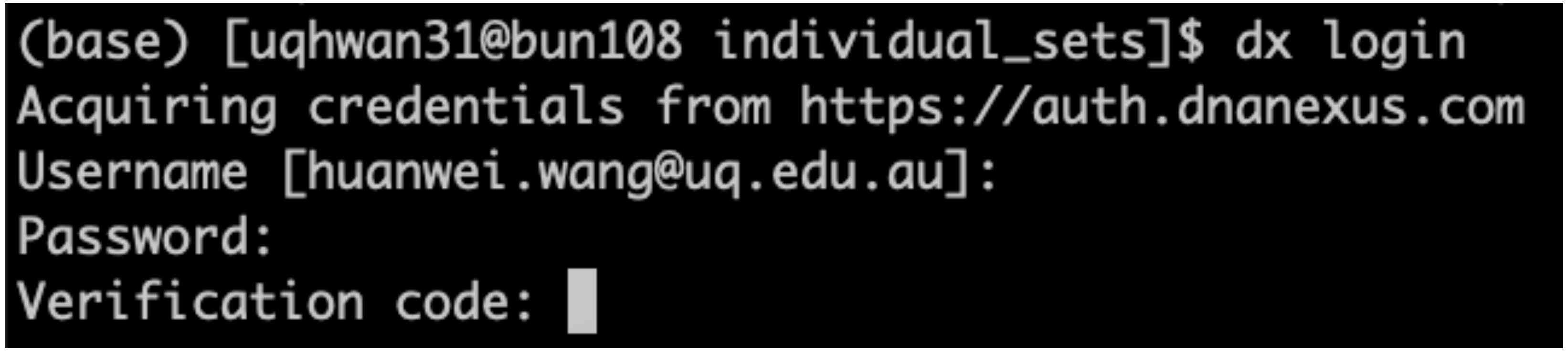
```
(base) imb20-015312-lt:~ uqhwan31$ dx --version
dx v0.368.1
```



```
 (base) imb20-015312-lt:~ uqhwan31$ dx --help
usage: dx [-h] [--version] command ...

DNAnexus Command-Line Client, API v1.0.0, client v0.368.1

dx is a command-line client for interacting with the DNAnexus platform.  You can log in, navigate,
upload, organize and share your data, launch analyses, and more.  For a quick tour of what the tool
can do, see
```

https://documentation.dnanexus.com/downloads#DNAnexus-Platform-SDK

dx login

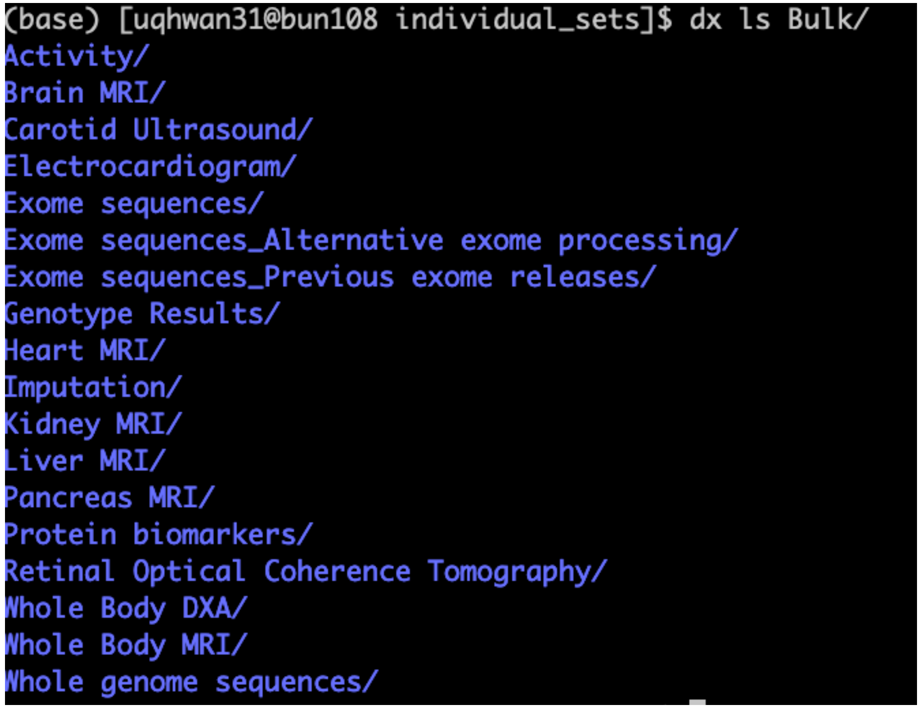


```
(base) [uqhwan31@bun108 individual_sets]$ dx login
Acquiring credentials from https://auth.dnanexus.com
Username [huanwei.wang@uq.edu.au]:
Password:
Verification code: 
```

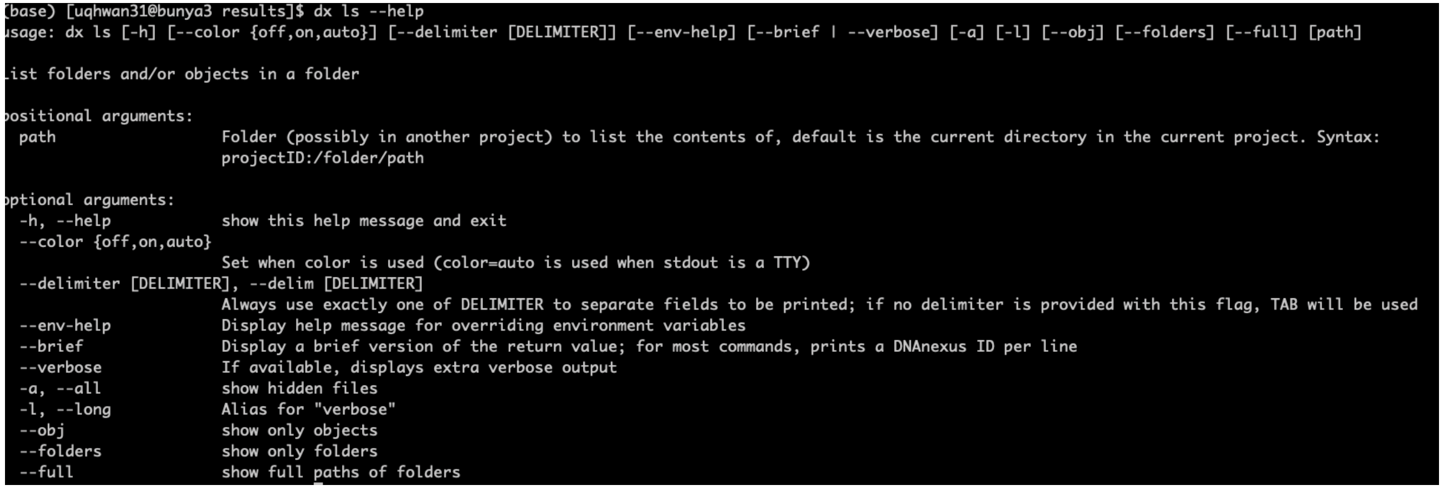
https://documentation.dnanexus.com/user/helpstrings-of-sdk-command-line-utilities

dx commands

- dx ls
- dx cd
- dx download
- dx mv
- dx mkdir
- dx ls --help
- …



```
(base) [uqhwan31@bun108 individual_sets]$ dx ls Bulk/
Activity/
Brain MRI/
Carotid Ultrasound/
Electrocardiogram/
Exome sequences/
Exome sequences_Alternative exome processing/
Exome sequences_Previous exome releases/
Genotype Results/
Heart MRI/
Imputation/
Kidney MRI/
Liver MRI/
Pancreas MRI/
Protein biomarkers/
Retinal Optical Coherence Tomography/
Whole Body DXA/
Whole Body MRI/
Whole genome sequences/
```

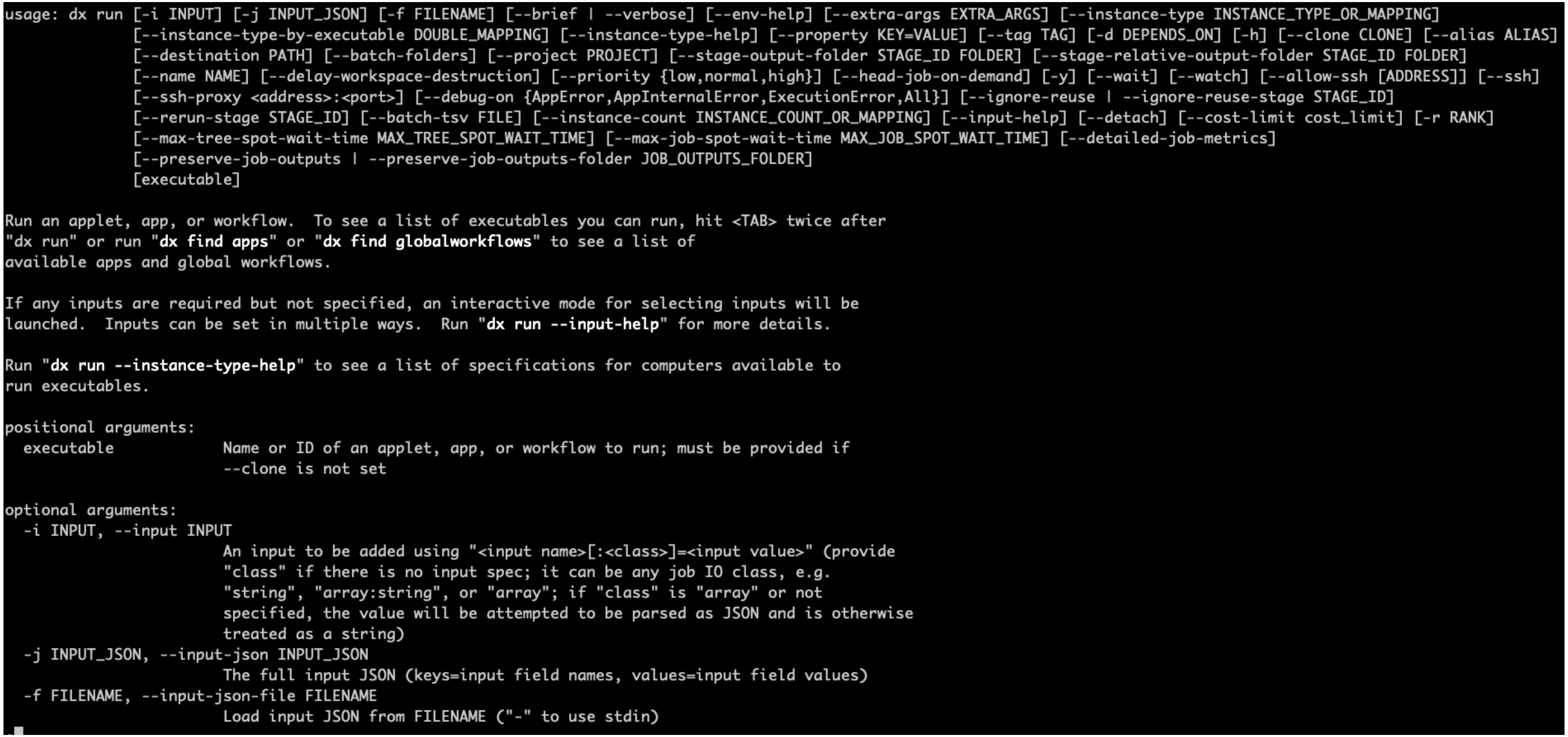


```
(base) [uqhwan31@bunya3 results]$ dx ls --help
usage: dx ls [-h] [--color {off,on,auto}] [--delimiter [DELIMITER]] [--env-help] [--brief | --verbose] [-a] [-l] [--obj] [--folders] [--full] [path]

List folders and/or objects in a folder

positional arguments:
  path                  Folder (possibly in another project) to list the contents of, default is the current directory in the current project. Syntax:
                        projectID:/folder/path

optional arguments:
  -h, --help            show this help message and exit
  --color {off,on,auto}
                        Set when color is used (color=auto is used when stdout is a TTY)
  --delimiter [DELIMITER], --delim [DELIMITER]
                        Always use exactly one of DELIMITER to separate fields to be printed; if no delimiter is provided with this flag, TAB will be used
  --env-help            Display help message for overriding environment variables
  --brief               Display a brief version of the return value; for most commands, prints a DNAnexus ID per line
  --verbose             If available, displays extra verbose output
  -a, --all             show hidden files
  -l, --long            Alias for "verbose"
  --obj                 show only objects
  --folders             show only folders
  --full                show full paths of folders
```

https://documentation.dnanexus.com/user/helpstrings-of-sdk-command-line-utilities

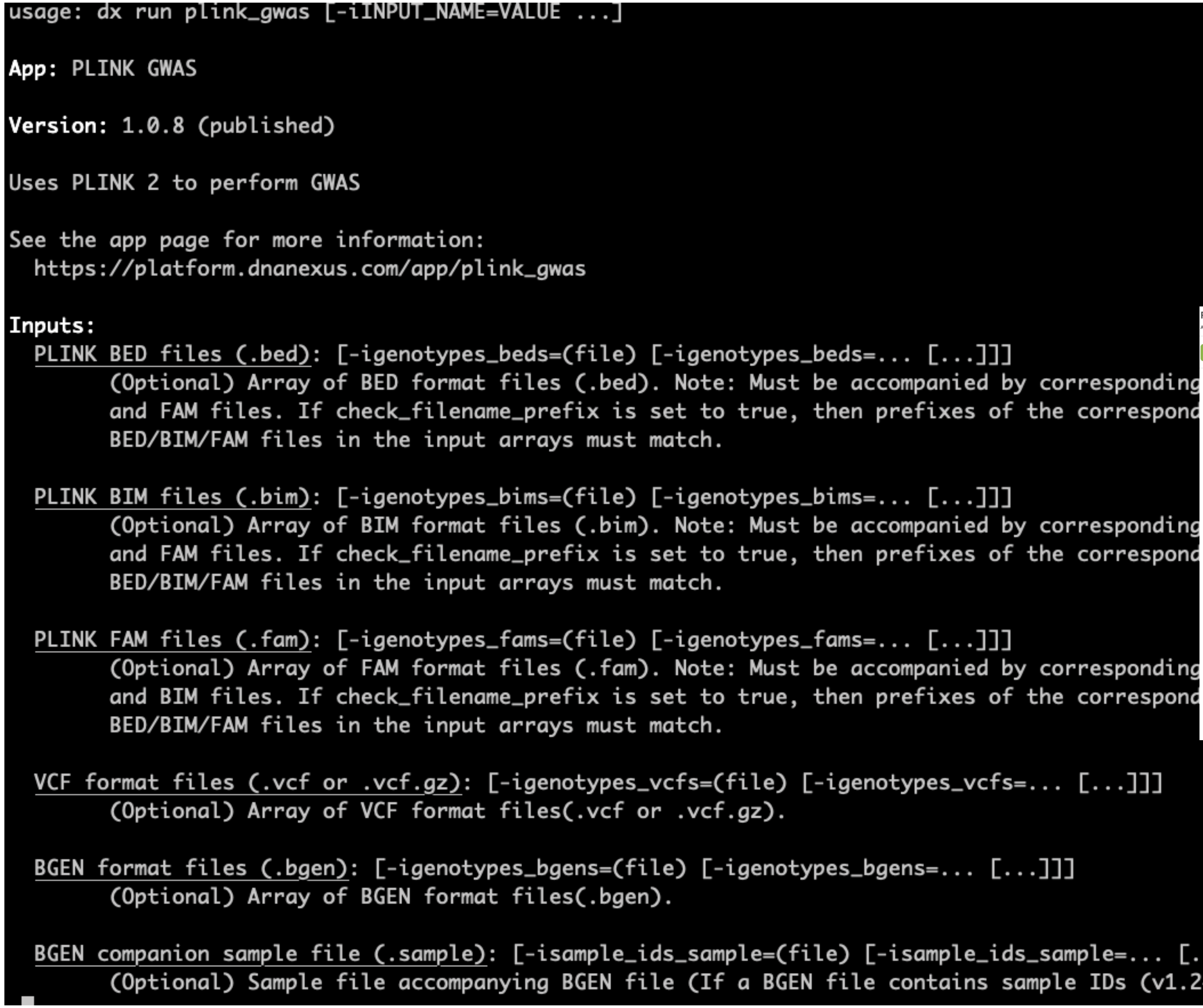
Submit a job on CLI

dx run --help

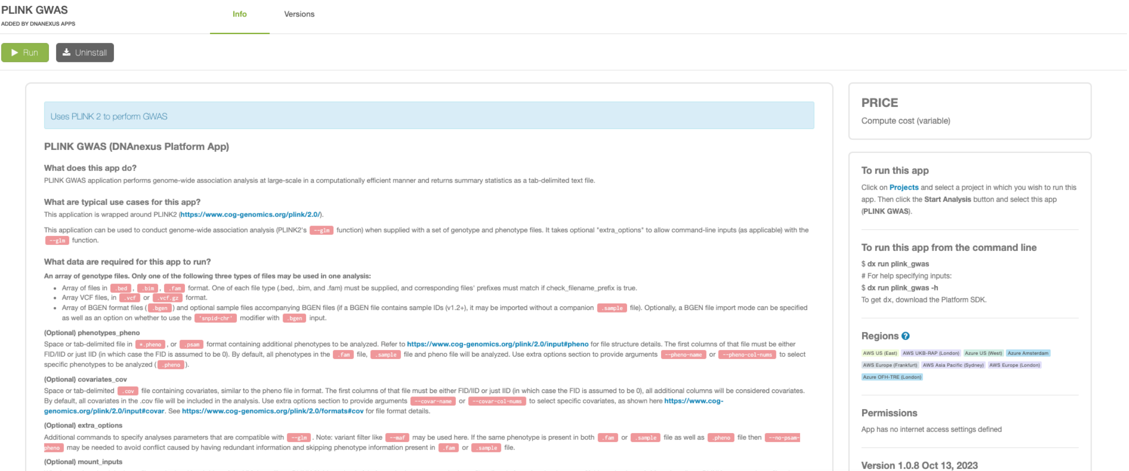


```
usage: dx run [-i INPUT] [-j INPUT_JSON] [-f FILENAME] [--brief | --verbose] [--env-help] [--extra-args EXTRA_ARGS] [--instance-type INSTANCE_TYPE_OR_MAPPING]
              [--instance-type-by-executable DOUBLE_MAPPING] [--instance-type-help] [--property KEY=VALUE] [--tag TAG] [-d DEPENDS_ON] [-h] [--clone CLONE] [--alias ALIAS]
              [--destination PATH] [--batch-folders] [--project PROJECT] [--stage-output-folder STAGE_ID FOLDER] [--stage-relative-output-folder STAGE_ID FOLDER]
              [--name NAME] [--delay-workspace-destruction] [--priority {low,normal,high}] [--head-job-on-demand] [-y] [--wait] [--watch] [--allow-ssh [ADDRESS]] [--ssh]
              [--ssh-proxy <address>:<port>] [--debug-on {AppError,AppInternalError,ExecutionError,All}] [--ignore-reuse | --ignore-reuse-stage STAGE_ID]
              [--rerun-stage STAGE_ID] [--batch-tsv FILE] [--instance-count INSTANCE_COUNT_OR_MAPPING] [--input-help] [--detach] [--cost-limit cost_limit] [-r RANK]
              [--max-tree-spot-wait-time MAX_TREE_SPOT_WAIT_TIME] [--max-job-spot-wait-time MAX_JOB_SPOT_WAIT_TIME] [--detailed-job-metrics]
              [--preserve-job-outputs | --preserve-job-outputs-folder JOB_OUTPUTS_FOLDER]
              [executable]

Run an applet, app, or workflow.  To see a list of executables you can run, hit <TAB> twice after
"dx run" or run "dx find apps" or "dx find globalworkflows" to see a list of
available apps and global workflows.

If any inputs are required but not specified, an interactive mode for selecting inputs will be
launched.  Inputs can be set in multiple ways.  Run "dx run --input-help" for more details.

Run "dx run --instance-type-help" to see a list of specifications for computers available to
run executables.

positional arguments:
  executable            Name or ID of an applet, app, or workflow to run; must be provided if
                        --clone is not set

optional arguments:
  -i INPUT, --input INPUT
                        An input to be added using "<input name>[:<class>]=<input value>" (provide
                        "class" if there is no input spec; it can be any job IO class, e.g.
                        "string", "array:string", or "array"; if "class" is "array" or not
                        specified, the value will be attempted to be parsed as JSON and is otherwise
                        treated as a string)
  -j INPUT_JSON, --input-json INPUT_JSON
                        The full input JSON (keys=input field names, values=input field values)
  -f FILENAME, --input-json-file FILENAME
                        Load input JSON from FILENAME ("-" to use stdin)
```

dx run plink_gwas --help



```
usage: dx run plink_gwas [-iINPUT_NAME=VALUE ...]

App: PLINK GWAS

Version: 1.0.8 (published)

Uses PLINK 2 to perform GWAS

See the app page for more information:
  https://platform.dnanexus.com/app/plink_gwas

Inputs:
  PLINK BED files (.bed): [-igenotypes_beds=(file) [-igenotypes_beds=... [...]]]
        (Optional) Array of BED format files (.bed). Note: Must be accompanied by corresponding
        and FAM files. If check_filename_prefix is set to true, then prefixes of the correspond
        BED/BIM/FAM files in the input arrays must match.

  PLINK BIM files (.bim): [-igenotypes_bims=(file) [-igenotypes_bims=... [...]]]
        (Optional) Array of BIM format files (.bim). Note: Must be accompanied by corresponding
        and FAM files. If check_filename_prefix is set to true, then prefixes of the correspond
        BED/BIM/FAM files in the input arrays must match.

  PLINK FAM files (.fam): [-igenotypes_fams=(file) [-igenotypes_fams=... [...]]]
        (Optional) Array of FAM format files (.fam). Note: Must be accompanied by corresponding
        and BIM files. If check_filename_prefix is set to true, then prefixes of the correspond
        BED/BIM/FAM files in the input arrays must match.

  VCF format files (.vcf or .vcf.gz): [-igenotypes_vcfs=(file) [-igenotypes_vcfs=... [...]]]
        (Optional) Array of VCF format files(.vcf or .vcf.gz).

  BGEN format files (.bgen): [-igenotypes_bgens=(file) [-igenotypes_bgens=... [...]]]
        (Optional) Array of BGEN format files(.bgen).

  BGEN companion sample file (.sample): [-isample_ids_sample=(file) [-isample_ids_sample=... [...]]]
        (Optional) Sample file accompanying BGEN file (If a BGEN file contains sample IDs (v1.2+),
```



Uses PLINK 2 to perform GWAS

PLINK GWAS (DNAnexus Platform App)

What does this app do?

PLINK GWAS application performs genome-wide association analysis at large-scale in a computationally efficient manner and returns summary statistics as a tab-delimited text file.

What are typical use cases for this app?

This application is wrapped around PLINK2 (https://www.cog-genomics.org/plink/2.0/).

This application can be used to conduct genome-wide association analysis (PLINK2's `--glm` function) when supplied with a set of genotype and phenotype files. It takes optional "extra_options" to allow command-line inputs (as applicable) with the `--glm` function.

What data are required for this app to run?

An array of genotype files. Only one of the following three types of files may be used in one analysis:

- Array of files in `.bed`, `.bim`, `.fam` format. One of each file type (.bed, .bim, and .fam) must be supplied, and corresponding files' prefixes must match if check_filename_prefix is true.
- Array VCF files, in `.vcf` or `.vcf.gz` format.
- Array of BGEN format files (`.bgen`) and optional sample files accompanying BGEN files (if a BGEN file contains sample IDs (v1.2+), it may be imported without a companion `.sample` file). Optionally, a BGEN file import mode can be specified as well as an option on whether to use the `'snpid-chr'` modifier with `.bgen` input.

(Optional) phenotypes_pheno

Space or tab-delimited file in `*.pheno`, or `.psam` format containing additional phenotypes to be analyzed. Refer to https://www.cog-genomics.org/plink/2.0/input#pheno for file structure details. The first columns of that file must be either FID/IID or just IID (in which case the FID is assumed to be 0). By default, all phenotypes in the `.fam` file, `.sample` file and pheno file will be analyzed. Use extra options section to provide arguments `--pheno-name` or `--pheno-col-nums` to select specific phenotypes to be analyzed (`.pheno`).

(Optional) covariates_cov

Space or tab-delimited `.cov` file containing covariates, similar to the pheno file in format. The first columns of that file must be either FID/IID or just IID (in which case the FID is assumed to be 0), all additional columns will be considered covariates. By default, all covariates in the .cov file will be included in the analysis. Use extra options section to provide arguments `--covar-name` or `--covar-col-nums` to select specific covariates, as shown here https://www.cog-genomics.org/plink/2.0/input#covar. See https://www.cog-genomics.org/plink/2.0/formats#cov for file format details.

(Optional) extra_options

Additional commands to specify analyses parameters that are compatible with `--glm`. Note: variant filter like `--maf` may be used here. If the same phenotype is present in both `.fam` or `.sample` file as well as `.pheno` file then `--no-psam-pheno` may be needed to avoid conflict caused by having redundant information and skipping phenotype information present in `.fam` or `.sample` file.

(Optional) mount_inputs

Whether to download the input files to the local hard drive of the VM that will run PLINK (if this option is false), or whether to mount the input files directly from the cloud storage (if this option is true). Mounting allows PLINK to access large files that would otherwise not fit in local storage. Note: the mount inputs is in its beta stage.

PRICE

Compute cost (variable)

To run this app

Click on **Projects** and select a project in which you wish to run this app. Then click the **Start Analysis** button and select this app (**PLINK GWAS**).

To run this app from the command line

$ **dx run plink_gwas**
\# For help specifying inputs:
$ **dx run plink_gwas -h**
To get dx, download the Platform SDK.

Regions

AWS US (East) AWS UKB-RAP (London) Azure US (West) Azure Amsterdam AWS Europe (Frankfurt) AWS Asia Pacific (Sydney) AWS Europe (London) Azure OFH-TRE (London)

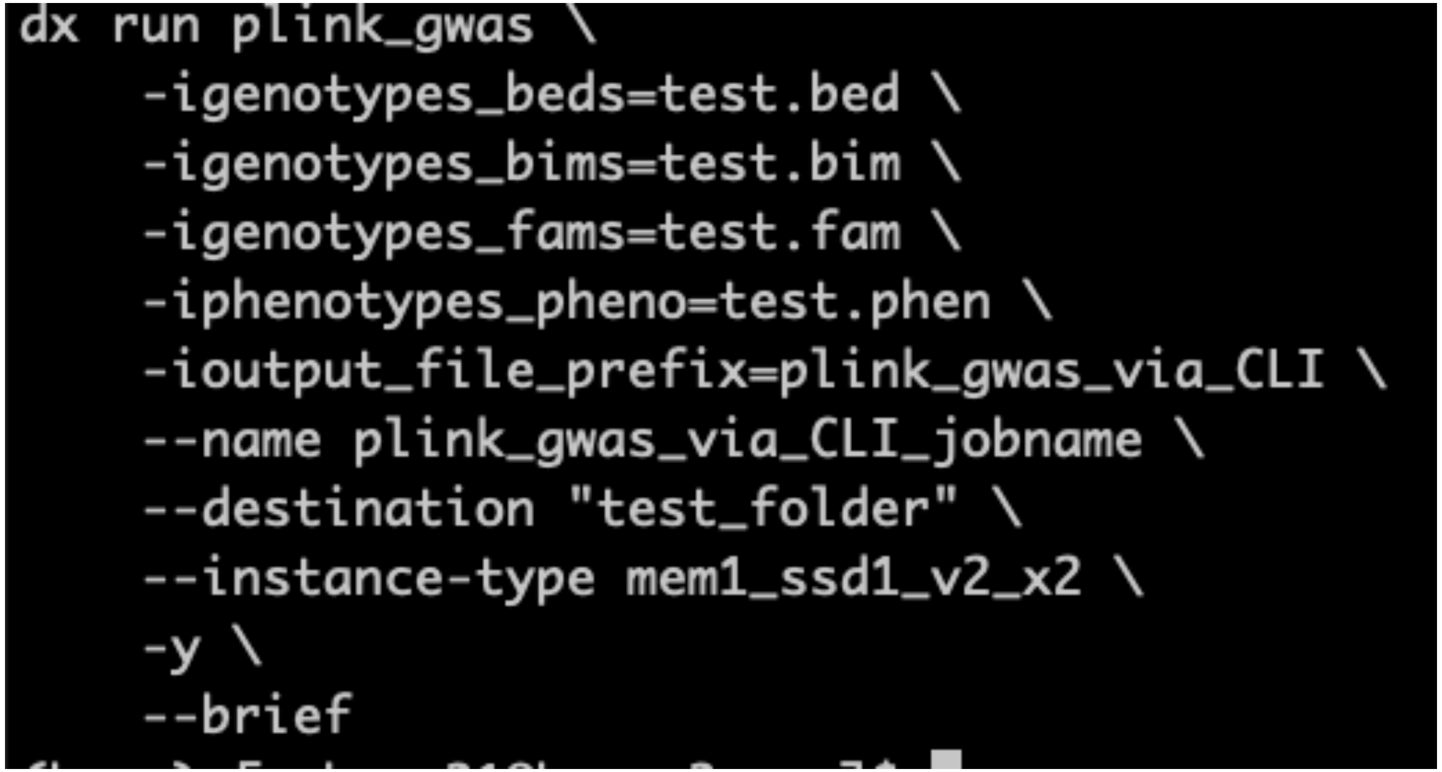
Permissions

App has no internet access settings defined

Version 1.0.8 Oct 13, 2023

What's new?

dx run plink_gwas

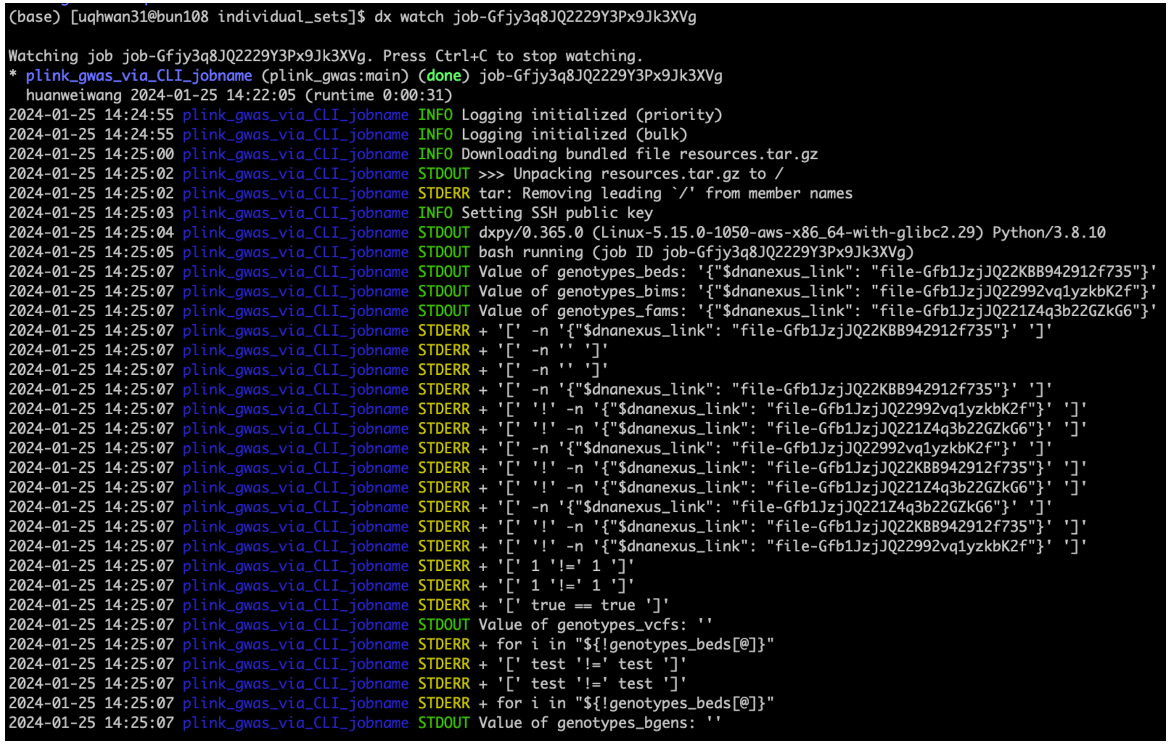


```
dx run plink_gwas \
    -igenotypes_beds=test.bed \
    -igenotypes_bims=test.bim \
    -igenotypes_fams=test.fam \
    -iphenotypes_pheno=test.phen \
    -ioutput_file_prefix=plink_gwas_via_CLI \
    --name plink_gwas_via_CLI_jobname \
    --destination "test_folder" \
    --instance-type mem1_ssd1_v2_x2 \
    -y \
    --brief
```

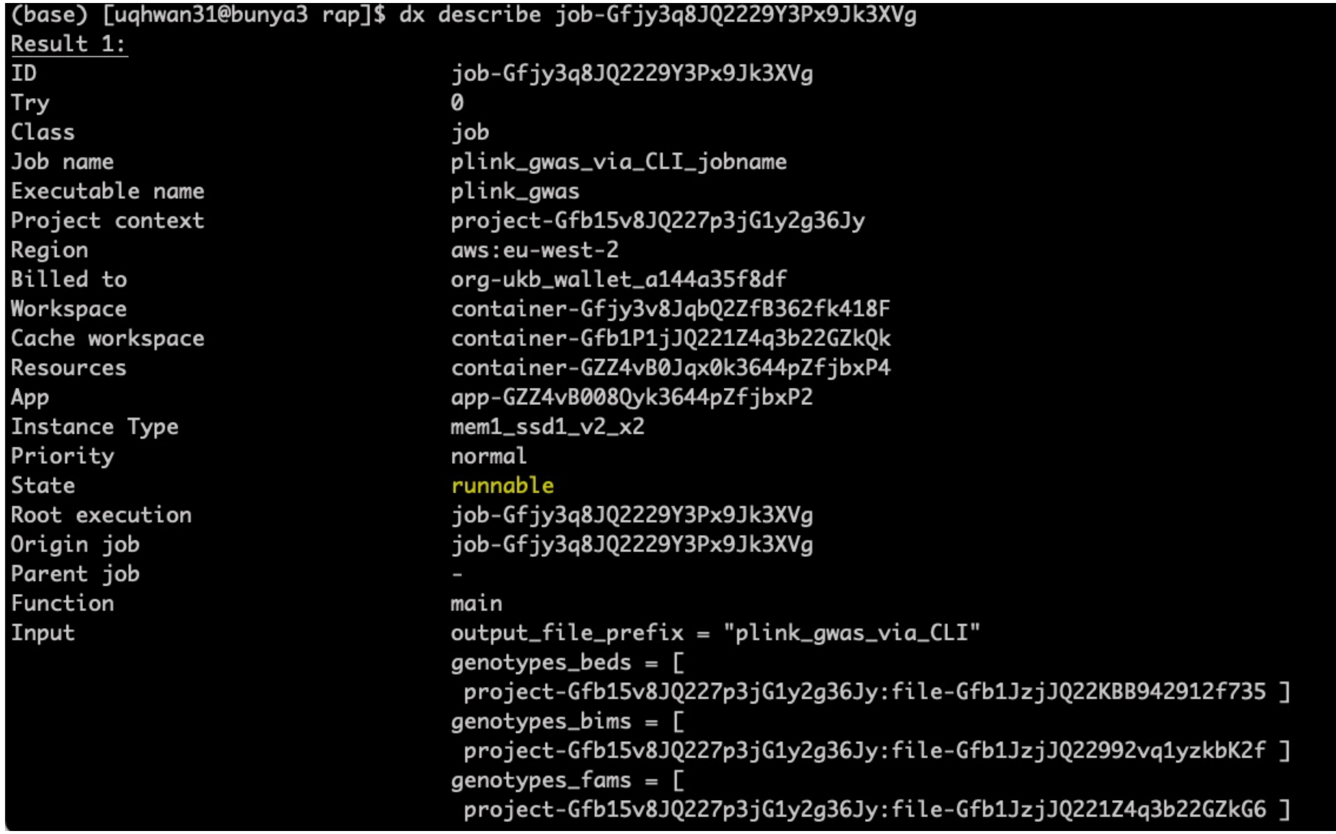
--name: job name
--destination: output folder
--instance-type:
-y: Do not ask for confirmation
--brief: Display a brief version of the return value

job-Gfjy3q8JQ2229Y3Px9Jk3XVg

Monitor a job in CLI

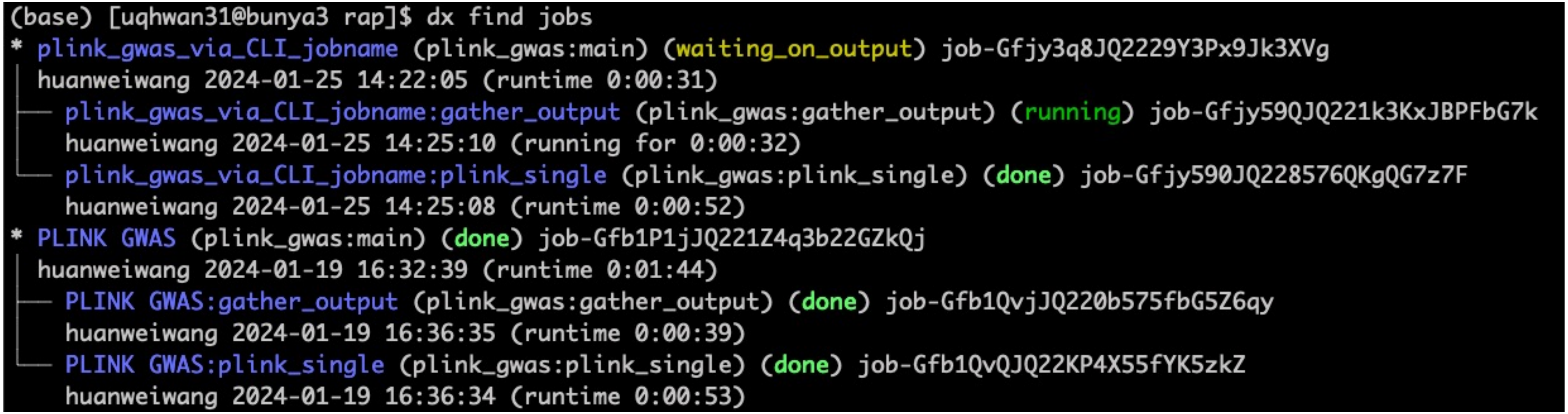


```
(base) [uqhwan31@bun108 individual_sets]$ dx watch job-Gfjy3q8JQ2229Y3Px9Jk3XVg

Watching job job-Gfjy3q8JQ2229Y3Px9Jk3XVg. Press Ctrl+C to stop watching.
* plink_gwas_via_CLI_jobname (plink_gwas:main) (done) job-Gfjy3q8JQ2229Y3Px9Jk3XVg
  huanweiwang 2024-01-25 14:22:05 (runtime 0:00:31)
2024-01-25 14:24:55 plink_gwas_via_CLI_jobname INFO Logging initialized (priority)
2024-01-25 14:24:55 plink_gwas_via_CLI_jobname INFO Logging initialized (bulk)
2024-01-25 14:25:00 plink_gwas_via_CLI_jobname INFO Downloading bundled file resources.tar.gz
2024-01-25 14:25:02 plink_gwas_via_CLI_jobname STDOUT >>> Unpacking resources.tar.gz to /
2024-01-25 14:25:02 plink_gwas_via_CLI_jobname STDERR tar: Removing leading `/' from member names
2024-01-25 14:25:03 plink_gwas_via_CLI_jobname INFO Setting SSH public key
2024-01-25 14:25:04 plink_gwas_via_CLI_jobname STDOUT dxpy/0.365.0 (Linux-5.15.0-1050-aws-x86_64-with-glibc2.29) Python/3.8.10
2024-01-25 14:25:05 plink_gwas_via_CLI_jobname STDOUT bash running (job ID job-Gfjy3q8JQ2229Y3Px9Jk3XVg)
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDOUT Value of genotypes_beds: '{"$dnanexus_link": "file-Gfb1JzjJQ22KBB942912f735"}'
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDOUT Value of genotypes_bims: '{"$dnanexus_link": "file-Gfb1JzjJQ22992vq1yzkbK2f"}'
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDOUT Value of genotypes_fams: '{"$dnanexus_link": "file-Gfb1JzjJQ221Z4q3b22GZkG6"}'
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDERR + '[' -n '{"$dnanexus_link": "file-Gfb1JzjJQ22KBB942912f735"}' ']'
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDERR + '[' -n '' ']'
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDERR + '[' -n '' ']'
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDERR + '[' -n '{"$dnanexus_link": "file-Gfb1JzjJQ22KBB942912f735"}' ']'
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDERR + '[' '!' -n '{"$dnanexus_link": "file-Gfb1JzjJQ22992vq1yzkbK2f"}' ']'
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDERR + '[' '!' -n '{"$dnanexus_link": "file-Gfb1JzjJQ221Z4q3b22GZkG6"}' ']'
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDERR + '[' -n '{"$dnanexus_link": "file-Gfb1JzjJQ22992vq1yzkbK2f"}' ']'
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDERR + '[' '!' -n '{"$dnanexus_link": "file-Gfb1JzjJQ22KBB942912f735"}' ']'
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDERR + '[' '!' -n '{"$dnanexus_link": "file-Gfb1JzjJQ221Z4q3b22GZkG6"}' ']'
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDERR + '[' -n '{"$dnanexus_link": "file-Gfb1JzjJQ221Z4q3b22GZkG6"}' ']'
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDERR + '[' '!' -n '{"$dnanexus_link": "file-Gfb1JzjJQ22KBB942912f735"}' ']'
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDERR + '[' '!' -n '{"$dnanexus_link": "file-Gfb1JzjJQ22992vq1yzkbK2f"}' ']'
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDERR + '[' 1 '!=' 1 ']'
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDERR + '[' 1 '!=' 1 ']'
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDERR + '[' true == true ']'
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDOUT Value of genotypes_vcfs: ''
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDERR + for i in "${!genotypes_beds[@]}"
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDERR + '[' test '!=' test ']'
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDERR + '[' test '!=' test ']'
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDERR + for i in "${!genotypes_beds[@]}"
2024-01-25 14:25:07 plink_gwas_via_CLI_jobname STDOUT Value of genotypes_bgens: ''
```



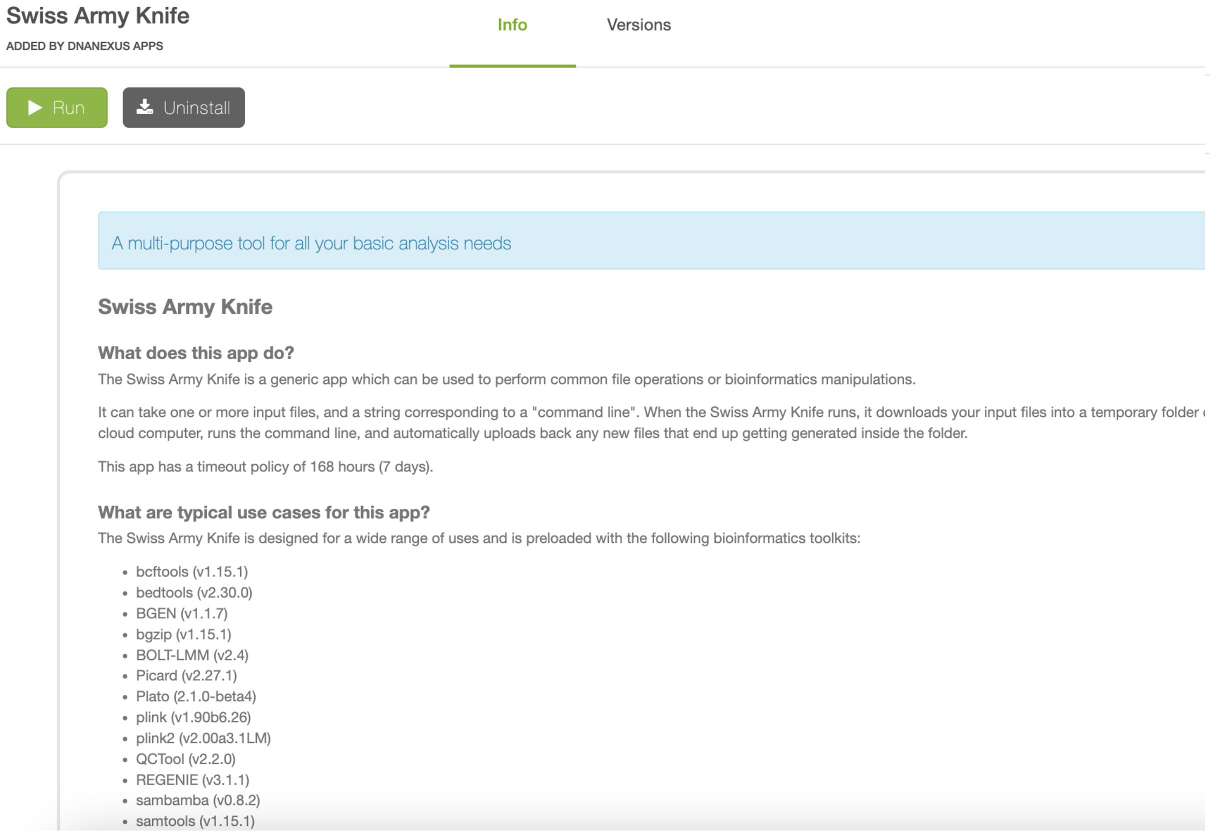
```
(base) [uqhwan31@bunya3 rap]$ dx describe job-Gfjy3q8JQ2229Y3Px9Jk3XVg
Result 1:
ID                                  job-Gfjy3q8JQ2229Y3Px9Jk3XVg
Try                                 0
Class                               job
Job name                            plink_gwas_via_CLI_jobname
Executable name                     plink_gwas
Project context                     project-Gfb15v8JQ227p3jG1y2g36Jy
Region                              aws:eu-west-2
Billed to                           org-ukb_wallet_a144a35f8df
Workspace                           container-Gfjy3v8JqbQ2ZfB362fk418F
Cache workspace                     container-Gfb1P1jJQ221Z4q3b22GZkQk
Resources                           container-GZZ4vB0Jqx0k3644pZfjbxP4
App                                 app-GZZ4vB008Qyk3644pZfjbxP2
Instance Type                       mem1_ssd1_v2_x2
Priority                            normal
State                               runnable
Root execution                      job-Gfjy3q8JQ2229Y3Px9Jk3XVg
Origin job                          job-Gfjy3q8JQ2229Y3Px9Jk3XVg
Parent job                          -
Function                            main
Input                               output_file_prefix = "plink_gwas_via_CLI"
                                    genotypes_beds = [
                                     project-Gfb15v8JQ227p3jG1y2g36Jy:file-Gfb1JzjJQ22KBB942912f735 ]
                                    genotypes_bims = [
                                     project-Gfb15v8JQ227p3jG1y2g36Jy:file-Gfb1JzjJQ22992vq1yzkbK2f ]
                                    genotypes_fams = [
                                     project-Gfb15v8JQ227p3jG1y2g36Jy:file-Gfb1JzjJQ221Z4q3b22GZkG6 ]
```

List the jobs in CLI



```
(base) [uqhwan31@bunya3 rap]$ dx find jobs
* plink_gwas_via_CLI_jobname (plink_gwas:main) (waiting_on_output) job-Gfjy3q8JQ2229Y3Px9Jk3XVg
│ huanweiwang 2024-01-25 14:22:05 (runtime 0:00:31)
├── plink_gwas_via_CLI_jobname:gather_output (plink_gwas:gather_output) (running) job-Gfjy59QJQ221k3KxJBPFbG7k
│   huanweiwang 2024-01-25 14:25:10 (running for 0:00:32)
└── plink_gwas_via_CLI_jobname:plink_single (plink_gwas:plink_single) (done) job-Gfjy590JQ228576QKgQG7z7F
    huanweiwang 2024-01-25 14:25:08 (runtime 0:00:52)
* PLINK GWAS (plink_gwas:main) (done) job-Gfb1P1jJQ221Z4q3b22GZkQj
│ huanweiwang 2024-01-19 16:32:39 (runtime 0:01:44)
├── PLINK GWAS:gather_output (plink_gwas:gather_output) (done) job-Gfb1QvjJQ220b575fbG5Z6qy
│   huanweiwang 2024-01-19 16:36:35 (runtime 0:00:39)
└── PLINK GWAS:plink_single (plink_gwas:plink_single) (done) job-Gfb1QvQJQ22KP4X55fYK5zkZ
    huanweiwang 2024-01-19 16:36:34 (runtime 0:00:53)
```

Another two APPs



Swiss Army Knife

ADDED BY DNANEXUS APPS

A multi-purpose tool for all your basic analysis needs

Swiss Army Knife

What does this app do?

The Swiss Army Knife is a generic app which can be used to perform common file operations or bioinformatics manipulations.

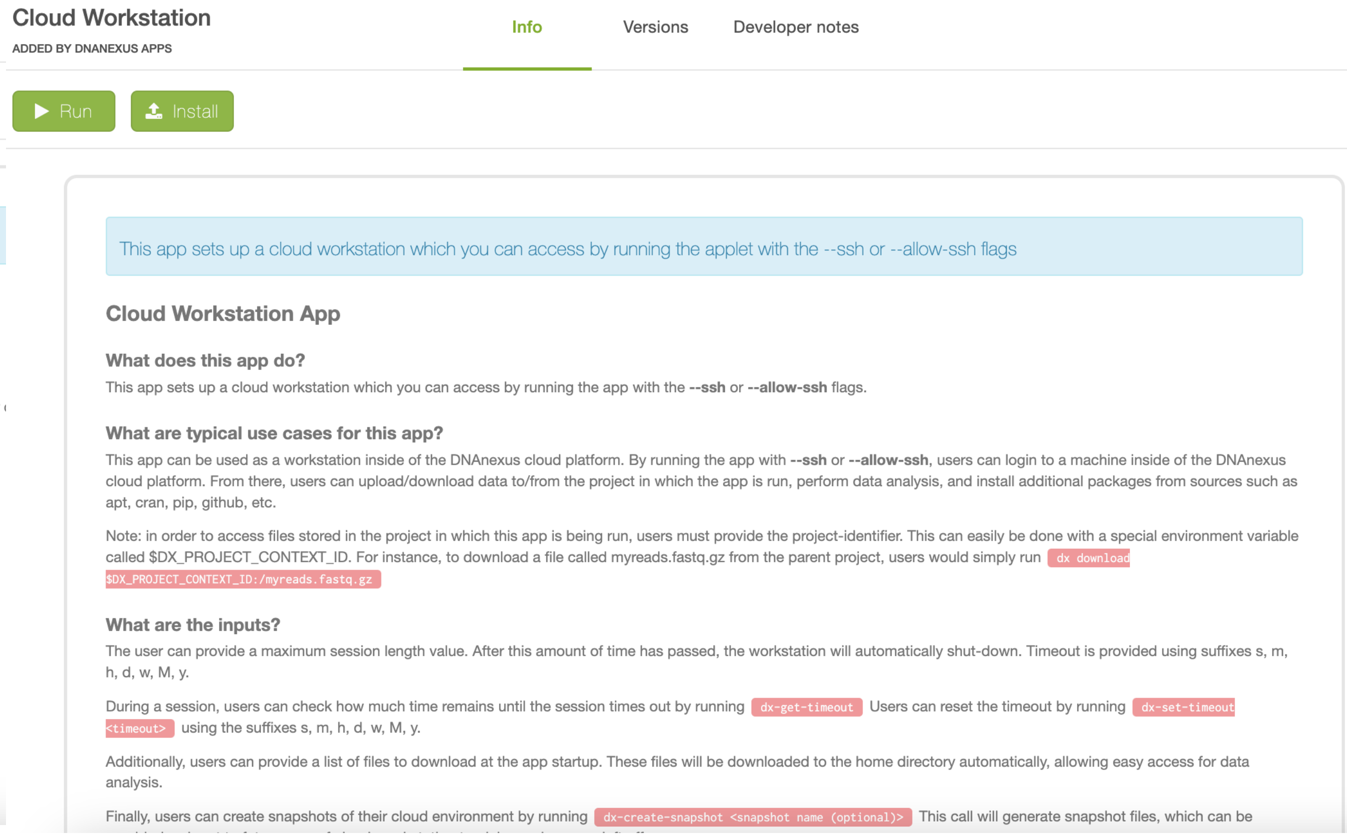
It can take one or more input files, and a string corresponding to a "command line". When the Swiss Army Knife runs, it downloads your input files into a temporary folder cloud computer, runs the command line, and automatically uploads back any new files that end up getting generated inside the folder.

This app has a timeout policy of 168 hours (7 days).

What are typical use cases for this app?

The Swiss Army Knife is designed for a wide range of uses and is preloaded with the following bioinformatics toolkits:

- bcftools (v1.15.1)
- bedtools (v2.30.0)
- BGEN (v1.1.7)
- bgzip (v1.15.1)
- BOLT-LMM (v2.4)
- Picard (v2.27.1)
- Plato (2.1.0-beta4)
- plink (v1.90b6.26)
- plink2 (v2.00a3.1LM)
- QCTool (v2.2.0)
- REGENIE (v3.1.1)
- sambamba (v0.8.2)
- samtools (v1.15.1)



Cloud Workstation

ADDED BY DNANEXUS APPS

This app sets up a cloud workstation which you can access by running the applet with the --ssh or --allow-ssh flags

Cloud Workstation App

What does this app do?

This app sets up a cloud workstation which you can access by running the app with the **--ssh** or **--allow-ssh** flags.

What are typical use cases for this app?

This app can be used as a workstation inside of the DNAnexus cloud platform. By running the app with **--ssh** or **--allow-ssh**, users can login to a machine inside of the DNAnexus cloud platform. From there, users can upload/download data to/from the project in which the app is run, perform data analysis, and install additional packages from sources such as apt, cran, pip, github, etc.

Note: in order to access files stored in the project in which this app is being run, users must provide the project-identifier. This can easily be done with a special environment variable called $DX_PROJECT_CONTEXT_ID. For instance, to download a file called myreads.fastq.gz from the parent project, users would simply run `dx download $DX_PROJECT_CONTEXT_ID:/myreads.fastq.gz`

What are the inputs?

The user can provide a maximum session length value. After this amount of time has passed, the workstation will automatically shut-down. Timeout is provided using suffixes s, m, h, d, w, M, y.

During a session, users can check how much time remains until the session times out by running `dx-get-timeout` Users can reset the timeout by running `dx-set-timeout <timeout>` using the suffixes s, m, h, d, w, M, y.

Additionally, users can provide a list of files to download at the app startup. These files will be downloaded to the home directory automatically, allowing easy access for data analysis.

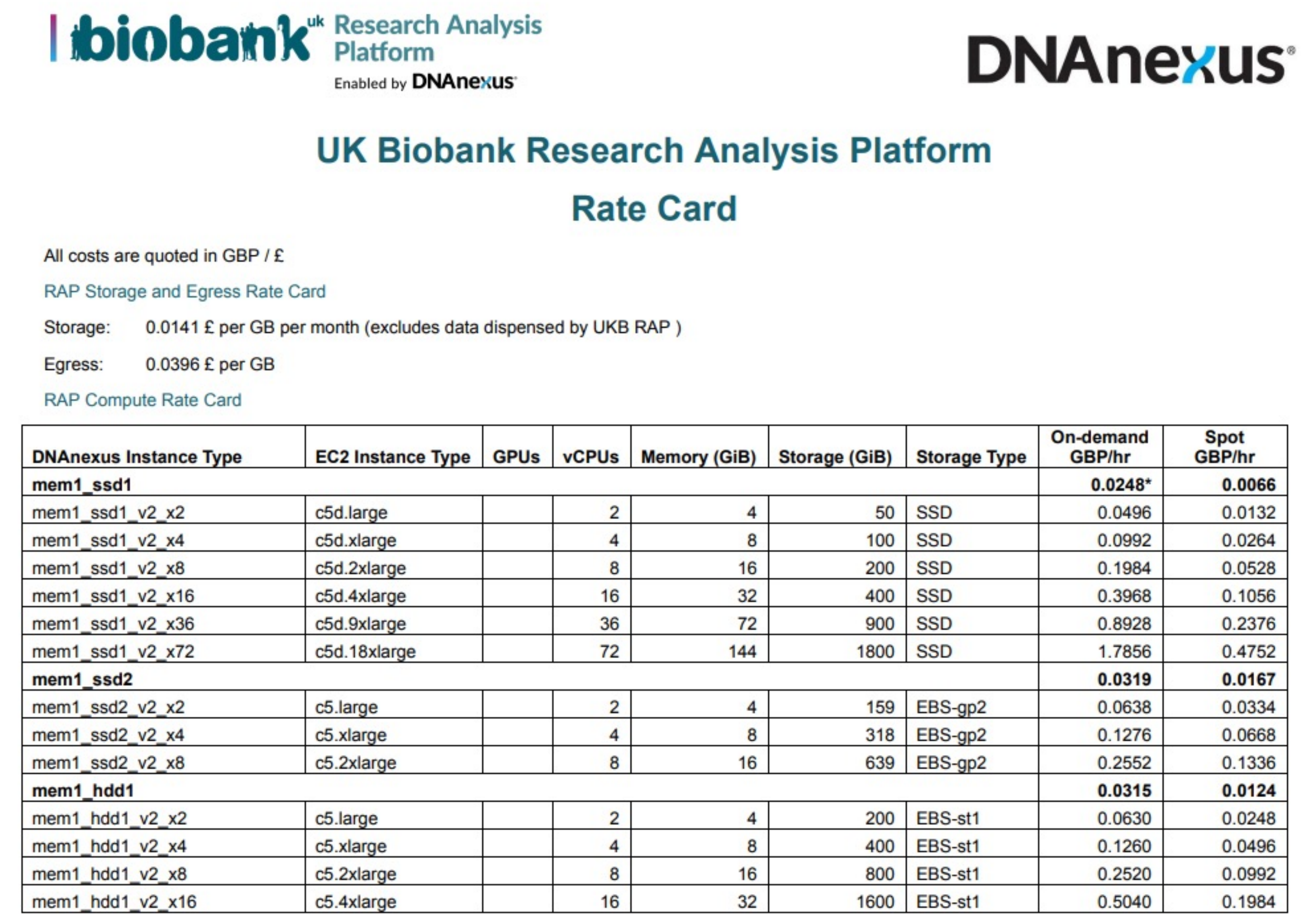
Finally, users can create snapshots of their cloud environment by running `dx-create-snapshot <snapshot name (optional)>` This call will generate snapshot files, which can be

https://ukbiobank.dnanexus.com/app/swiss-army-knife
https://ukbiobank.dnanexus.com/app/cloud_workstation

Costs

Rate card

Each new user receives a £40 credit toward covering usage costs



Research Analysis Platform

Enabled by DNAnexus

UK Biobank Research Analysis Platform

Rate Card

All costs are quoted in GBP / £

RAP Storage and Egress Rate Card

Storage: 0.0141 £ per GB per month (excludes data dispensed by UKB RAP)

Egress: 0.0396 £ per GB

RAP Compute Rate Card

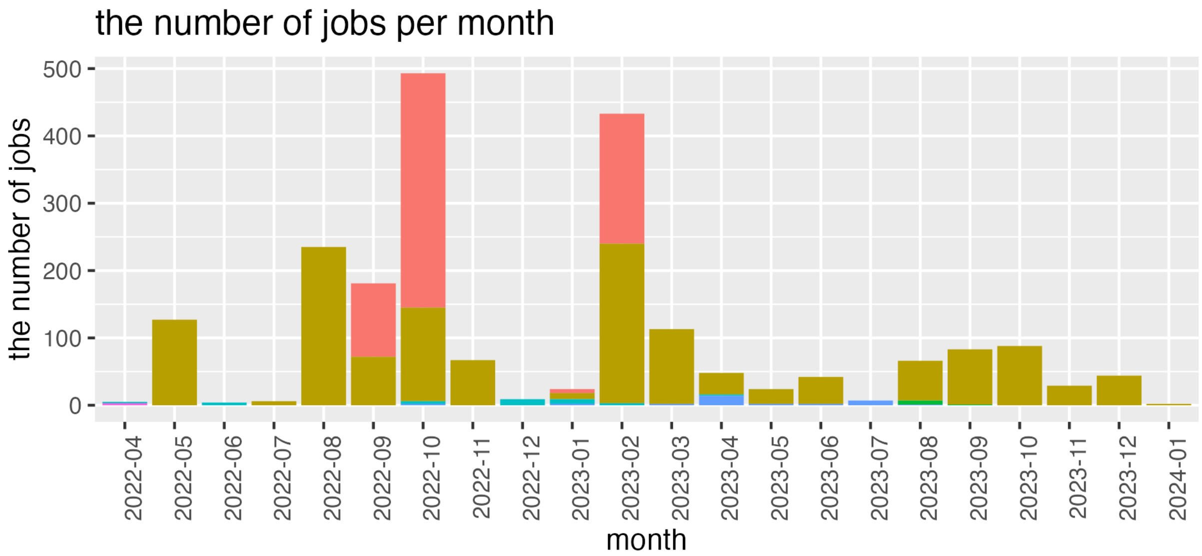
DNAnexus Instance Type	EC2 Instance Type	GPUs	vCPUs	Memory (GiB)	Storage (GiB)	Storage Type	On-demand GBP/hr	Spot GBP/hr
mem1_ssd1							**0.0248***	**0.0066**
mem1_ssd1_v2_x2	c5d.large		2	4	50	SSD	0.0496	0.0132
mem1_ssd1_v2_x4	c5d.xlarge		4	8	100	SSD	0.0992	0.0264
mem1_ssd1_v2_x8	c5d.2xlarge		8	16	200	SSD	0.1984	0.0528
mem1_ssd1_v2_x16	c5d.4xlarge		16	32	400	SSD	0.3968	0.1056
mem1_ssd1_v2_x36	c5d.9xlarge		36	72	900	SSD	0.8928	0.2376
mem1_ssd1_v2_x72	c5d.18xlarge		72	144	1800	SSD	1.7856	0.4752
mem1_ssd2							**0.0319**	**0.0167**
mem1_ssd2_v2_x2	c5.large		2	4	159	EBS-gp2	0.0638	0.0334
mem1_ssd2_v2_x4	c5.xlarge		4	8	318	EBS-gp2	0.1276	0.0668
mem1_ssd2_v2_x8	c5.2xlarge		8	16	639	EBS-gp2	0.2552	0.1336
mem1_hdd1							**0.0315**	**0.0124**
mem1_hdd1_v2_x2	c5.large		2	4	200	EBS-st1	0.0630	0.0248
mem1_hdd1_v2_x4	c5.xlarge		4	8	400	EBS-st1	0.1260	0.0496
mem1_hdd1_v2_x8	c5.2xlarge		8	16	800	EBS-st1	0.2520	0.0992
mem1_hdd1_v2_x16	c5.4xlarge		16	32	1600	EBS-st1	0.5040	0.1984

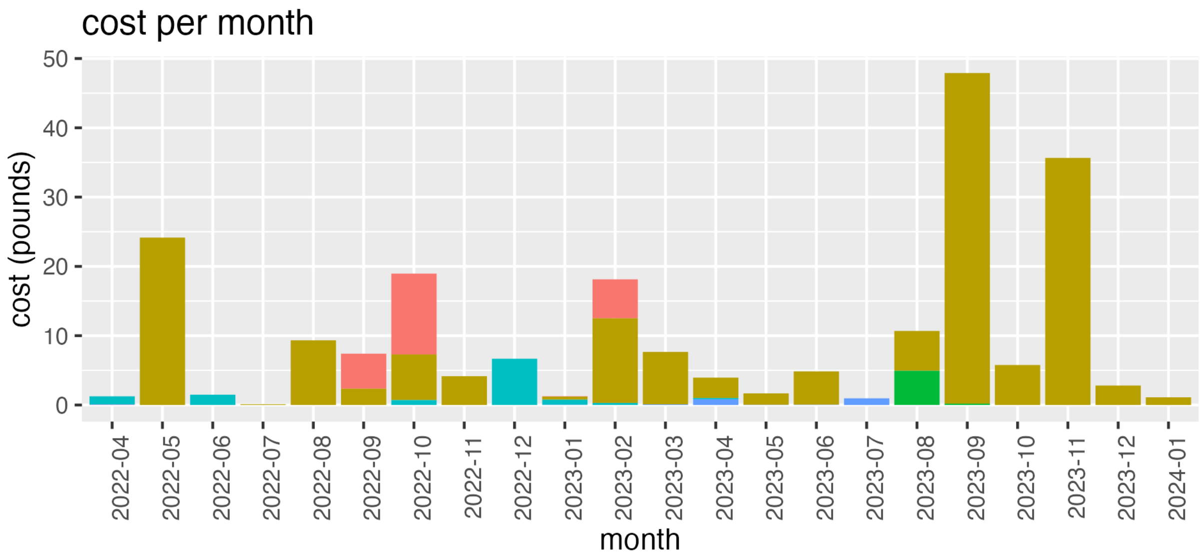
https://20779781.fs1.hubspotusercontent-na1.net/hubfs/20779781/Product%20Team%20Folder/Rate%20Cards/BiobankResearchAnalysisPlatform_Rate%20Card_Current.pdf

Computing cost for our group

- 2130 jobs
- 215.95 pounds (410.31 AUD based on 1 pound=1.9 AUD)
- an average of 0.1 pound per job

By 29/Jan/2024





AWS credits

Eligible researchers can apply for credits to offset the costs of compute and storage in the UKB-RAP

Getting Started Credits

- £1,000 fixed amount valid for 1 year
- Researchers can explore and understand the RAP
- Researchers can only receive one Getting Started Credit

Enhanced Credits

- In multiples of £1,000 valid for one year
- Researchers can request more than one Enhanced Credit

https://dnanexus.gitbook.io/uk-biobank-rap/working-on-the-research-analysis-platform/billing-and-costs

Further resource

Further resource

UKB-RAP Documentation

About the Research Analysis Platfor

What is UK Biobank?

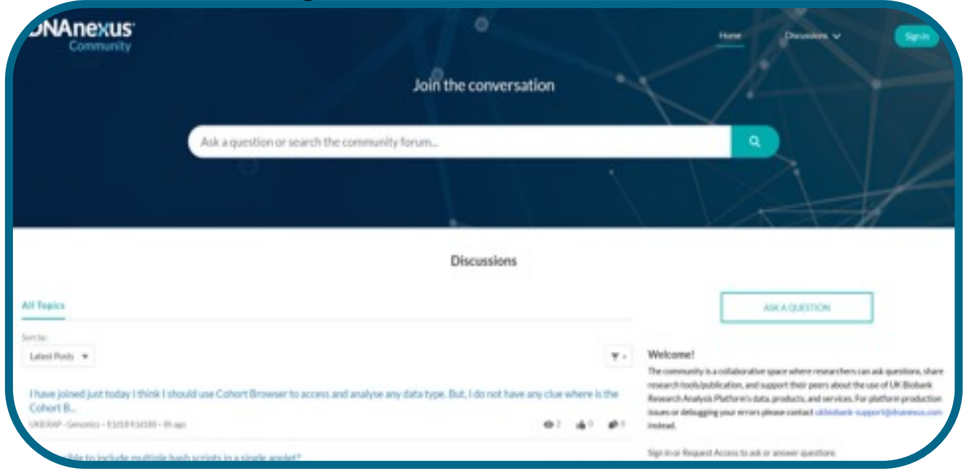
UK Biobank is a national and international health resource with unparalleled research oppor
to all bona fide health researchers. UK Biobank aims to improve the prevention, diagnosis an
wide range of serious and life-threatening illnesses – including cancer, heart diseases, stroke
arthritis, osteoporosis, eye disorders, depression and forms of dementia. It is following the he
being of 500,000 volunteer participants and provides health information, which does not ide
approved researchers in the UK and overseas, from academia and industry.

The UK Biobank Research Analysis Platform

The UK Biobank Research Analysis Platform is an informatics platform that provides acces
of, UK Biobank data by its registered researcher community. Read the announcement.

dnanexus.gitbook.io/uk-biobank-rap

Community Forum



Join the conversation

Discussions

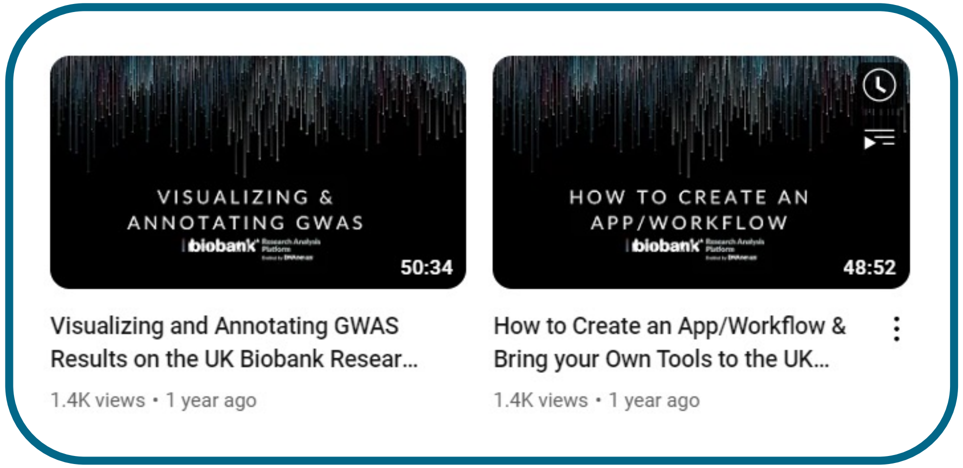
I have joined just today I think I should use Cohort Browser to access and analyse any data type. But, I do not have any clue where is the Cohort B...

Welcome!

The community is a collaborative space where researchers can ask questions, share research tools/publication, and support their peers about the use of UK Biobank Research Analysis Platform's data, products, and services. For platform production issues or debugging your errors please contact ukbiobank-support@dnanexus.com instead.

Sign in or Request Access to ask or answer questions.

DNAnexus Training Webinars

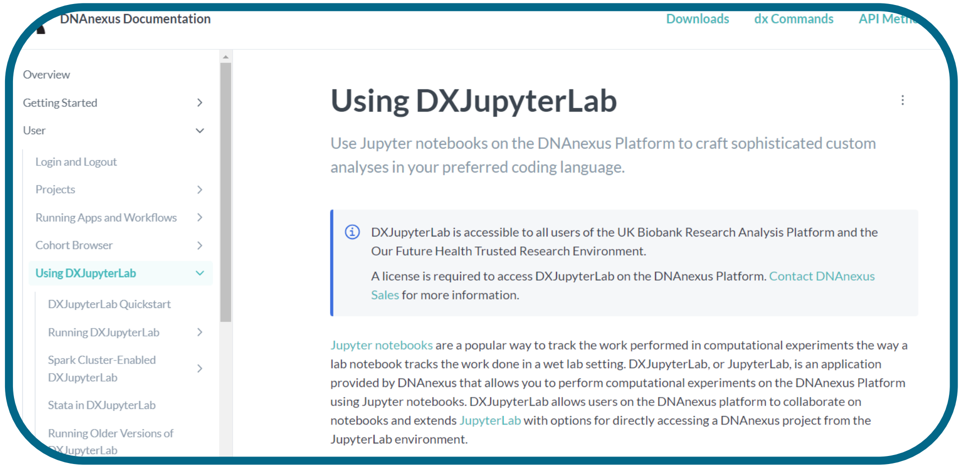


Visualizing and Annotating GWAS Results on the UK Biobank Resear...
1.4K views • 1 year ago

How to Create an App/Workflow & Bring your Own Tools to the UK...
1.4K views • 1 year ago

youtube.com/Dnanexus

DNAnexus Platform Documentation



Using DXJupyterLab

Use Jupyter notebooks on the DNAnexus Platform to craft sophisticated custom analyses in your preferred coding language.

DXJupyterLab is accessible to all users of the UK Biobank Research Analysis Platform and the Our Future Health Trusted Research Environment.

A license is required to access DXJupyterLab on the DNAnexus Platform. Contact DNAnexus Sales for more information.

Jupyter notebooks are a popular way to track the work performed in computational experiments the way a lab notebook tracks the work done in a wet lab setting. DXJupyterLab, or JupyterLab, is an application provided by DNAnexus that allows you to perform computational experiments on the DNAnexus Platform using Jupyter notebooks. DXJupyterLab allows users on the DNAnexus platform to collaborate on notebooks and extends JupyterLab with options for directly accessing a DNAnexus project from the JupyterLab environment.

documentation.dnanexus.com

Any questions?

Thank you for your attention