WHITE PAPER

Working with UK Biobank Data

A Research Guide

DNAnexus®

Contents

Introduction

Do you have the right informatics strategy to get to insights quickly while working with the UK Biobank (UKB) dataset?

It's never easy to conduct a study that draws on both genomic and clinical data. The challenges are even greater when working with a dataset as large and complex as the UK Biobank. While its breadth and depth make it an incredibly rich resource for researchers, exploring it, and selecting the right data for a particular research study can be time-consuming. And when it comes to the analysis stage, many organizations find that their in-house technology just isn't powerful enough to handle the load.

What novel associations between genetic traits and health outcomes will you discover, by using UKB data in your research?

What is UK Biobank?

UK Biobank is a uniquely valuable resource for health researchers, incorporating genomic and clinical data collected from 500,000 volunteer participants. The project has drawn international praise for its ambitious scope, and its focus on helping medical professionals better understand, and thus better treat and prevent a range of serious illnesses.

UKB exome data is being released to approved researchers in batches. Researchers from pharmaceutical companies and research organizations around the world are already publishing studies drawing on the first batch, which consists of data on 50,000 exomes.[1] Another tranche, covering 100,000 participants, is due for release in mid-2020.

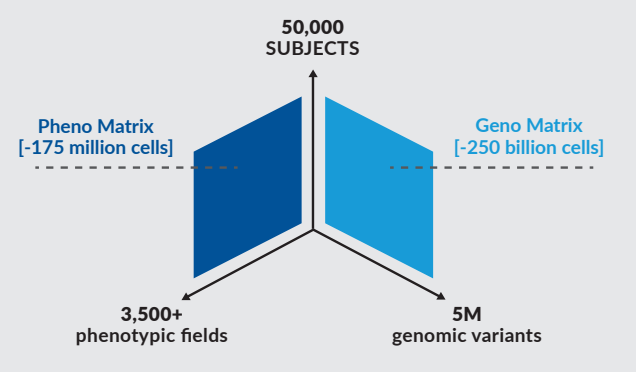
[1] https://www.ukbiobank.ac.uk/2019/03/new-genetic-data-on-50000-uk-biobank-participants-made-available-to-the-global-health-research-community/

1 How to Work at Population Scale

A Closer Look

Let's look closer at the UK Biobank dataset, to get a sense of its true scale and complexity. The first tranche of whole-exome sequence (WES) data, covering 50,000 individuals, offers researchers an incredibly rich resource: 250 billion genomic, and 175 million phenotypic data points. And consider that this release covers only 10% of the total participant population. The full dataset, covering 500,000 participants, will contain trillions of data points.

> ***The full dataset, covering 500,000 participants, will contain trillions of data points.***



As UK Biobank releases more whole exome and phenotypic data, and eventually whole genome data, researchers will find it increasingly difficult to download, access, and store it all securely, let alone conduct analyses. To do all this, they'll need a bioinformatics system that scales easily and effectively, as their datasets become larger and more complex.

Enter Modern Cloud Technology

> ***Many on-premise systems face challenges when dealing with very large, multidimensional datasets.***

Many on-premise systems face challenges when dealing with very large, multidimensional datasets. Storing, accessing, filtering, querying, and sharing data can all become very difficult for end users, as on-premise systems run up against the limits of their capabilities. Contrast this with the performance of cloud-based systems, which are designed and built to scale effortlessly. As demand increases, new computing resources can be automatically provisioned, preventing delays in processing. For compute-intensive jobs, on-demand clusters ensure performance. For researchers running complex analyses that draw on enormous datasets, this means the ability to work faster, and get to insights quicker. And with their data, tools, and projects in the cloud, they can quickly and securely share results with collaborators, wherever they might be, further accelerating discovery.

WORKSHEET

Is Your Organization Ready for UK Biobank-scale Datasets?

For each statement, select the number that reflects your organization's level of readiness.

	N/A	STRONGLY DISAGREE		NEUTRAL		STRONGLY AGREE
My organization has the strategy and infrastructure in place to execute bioinformatics analysis in the cloud.	0	1	2	3	4	5
My current informatics system is equipped to handle the storage and analysis of data on 500,000 whole exomes.	0	1	2	3	4	5
My organization has enough bioinformaticians to conduct the analysis needed for my research project.	0	1	2	3	4	5
I have expert support in troubleshooting the pipeline and analysis issues I'll encounter, when dealing with very large datasets.	0	1	2	3	4	5

Your Result:

>16 - Your informatics solution should be able to handle the UKB dataset.

10-16 - You're moving in that direction. You've got the fundamentals down, but your organization will still face challenges in dealing with UKB-scale datasets. You should upgrade your informatics system, and strongly consider a purpose-built cloud solution.

<10 - You should invest in a system that will support you not only today, but also well into the future.

2 How to Meet Privacy & Security Requirements

Accessing UK Biobank Data

To access the UKB dataset, you'll need to make a formal application, certifying among other things, that your bioinformatics system meets stringent privacy and security requirements.

You're best served by choosing a solution with a track record of helping customers meet complex, ever-changing compliance and security requirements. Your system should incorporate robust authentication and authorization functionality. It should support fine-grained, systematic and comprehensive tracking of users, objects and data, at all times. In short, it should empower you to monitor and control access to everything in the system, by anyone who uses it. And with these features in place, you'll be able not only to safeguard the data you're using today, but also be prepared to address new security and compliance requirements, as they emerge.

Your system should not only safeguard the data you're using today, but be able to address future security and compliance requirements.

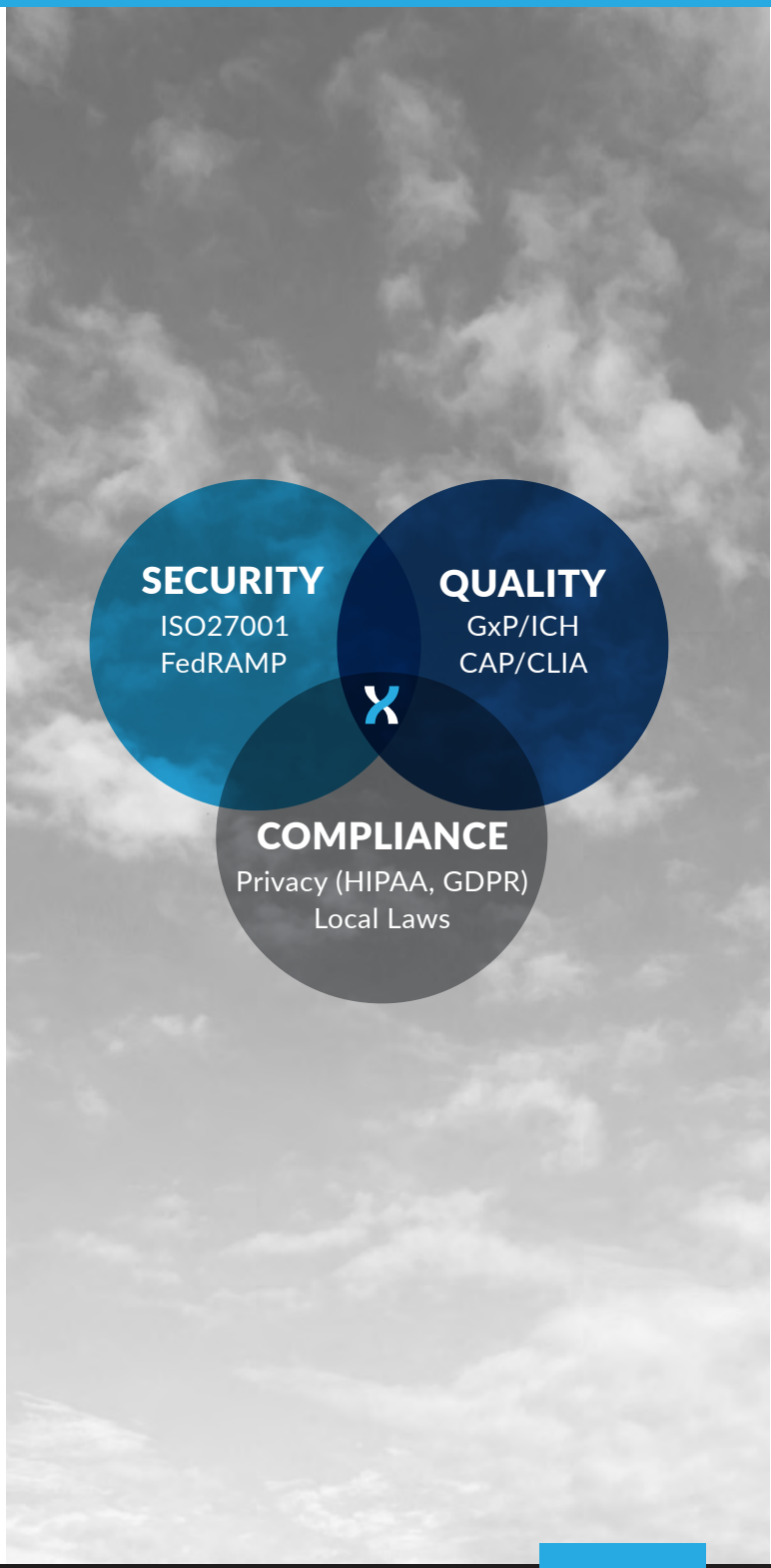
2 How to Meet Privacy & Security Requirements?

The UK Biobank Material Transfer Agreement

To ensure you can meet the security requirements laid down in Section 4.2 of UK Biobank's Material Transfer Agreement[2], your bioinformatics solution should provide all of the following:

- Compliance with applicable privacy standards, such as HIPAA, PIPEDA, and GDPR
- Guardrails to protect against data transfers that violate data transfer agreements
- End-to-end encryption using AES-256 or better
- Strong authentication (multi-factor authentication at a minimum)
- Implementation of different data sensitivity levels, implying different users can have different authorizations
- Audit trail with sufficient granularity to track "who saw what, when and why?"
- The ability to help you pass your own security and privacy audits

Know that when you choose a bioinformatics solution, you're also choosing a partner. Look for a partner that supports you with a team of security, quality, and compliance experts – professionals whose specialty is staying on top of the constantly evolving regulatory space.



[2] The section reads: "The Applicant will retain the Materials in a secure location as regards Samples or a secure network system as regards Data at such standard as would be reasonably expected for the storage of valuable and proprietary samples and/or sensitive/-confidential data." http://www.ukbiobank.ac.uk/wp-content/uploads/2012/09/Material-Transfer-Agreement.pdf

3 How to Navigate the UK Biobank Dataset

Exploring the Data

If you are new to the UKB dataset, open the UK Biobank Showcase to start exploring.[3] The Showcase provides a detailed overview of the dataset's structure and content, including a full list of fields, organized by category. You can examine aggregate data for the more than 3,500 individual fields, and quickly learn a good deal about the phenotypic and genotypic characteristics of the participant population. There is also information about collection methods and timing, any issues with the data in a particular category or field, number of participants covered in each, and other resources essential to helping you understand what UKB data will be most useful in your research.

Building and Analyzing Cohorts

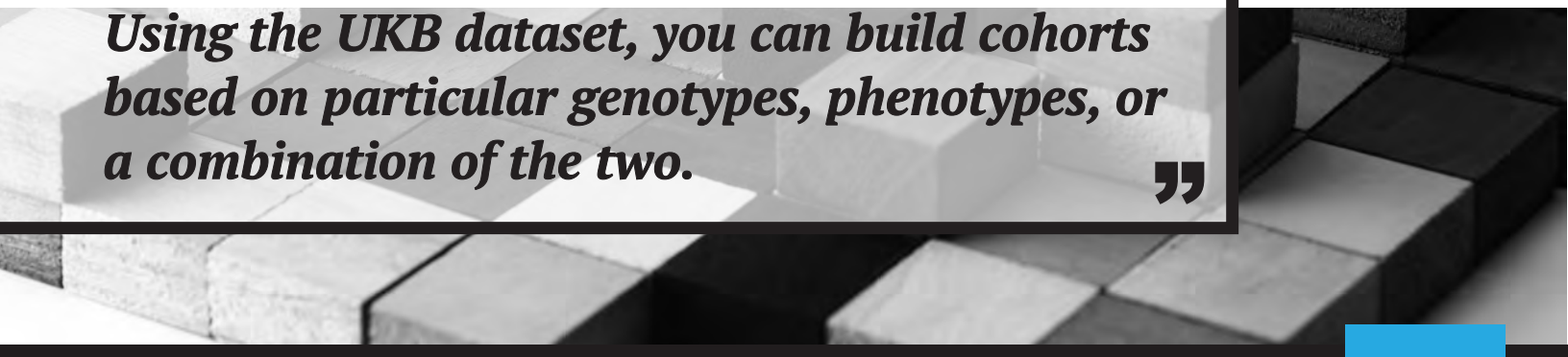
You'll want a way to quickly define cohorts, and understand the features of each. Using the UKB dataset, you can build cohorts based on particular genotypes, phenotypes, or a combination of genotypes and phenotypes. For example, we've met researchers who are interested in a particular variant, and want to use UKB data to explore the phenotypes of carriers, looking for associations. The UKB dataset, because of its scale and scope, offers exciting opportunities for doing just this sort of investigation.

You will face certain challenges here. One such challenge is easily querying genetic and phenotypic data together, in order to build a dataset that covers a specific set of participants. When doing this, rely on a bioinformatics solution that links these datasets seamlessly, by drawing on UKB-produced bridge and linking files.

Another challenge is tied to the dataset's size and richness. Because it's so large and diverse, legacy tools may struggle to support analyses that go beyond basic exploration. And as more exome sequencing data is released, this problem will only become more acute.

[3] http://biobank.ctsu.ox.ac.uk/crystal/

Using the UKB dataset, you can build cohorts based on particular genotypes, phenotypes, or a combination of the two.



4 Why Visualization Tools are Key to Success

The Right Tool Goes a Long Way

When working with a dataset that's so large and rich, researchers need an easy way to quickly build and explore cohorts, and form and test hypotheses about them, without the need for expert-level programming skills.

They need a tool that, among other things, enables them easily to visualize phenotypic data, including trait distribution for each phenotype. In advance of handing off cohort data to a bioinformatics expert for deeper analysis, this ensures that there will be enough data to power their-study, and also helps them correct for covariates as necessary.

Their visualization tool should also be able to handle all the various data types they'll use, and data formats they'll encounter when conducting both clinical and omics analyses. The UK Biobank dataset is a prime example of why this is so necessary, containing as it does an extremely wide array of phenotypic and genetic data, in a multitude of formats, including strings, floats, integers, and categoricals.

Researchers need an easy way to quickly build and explore cohorts, and form and test hypotheses about them...

4 Why Visualization Tools are Key to Success

A good visualization tool should also be able to handle encoded data. Again, the UKB dataset shows why. Many UKB fields are populated with encoded data. To make this data useful, a visualization tool should be able to auto-convert encoded values to more accessible terms.

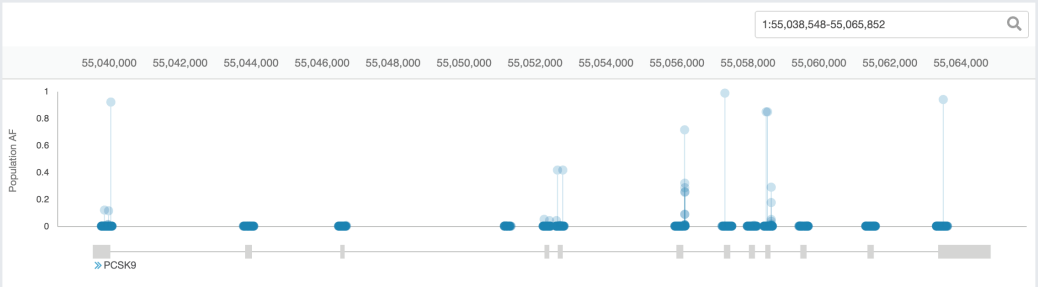
> ***A visualization tool should be able to auto-convert encoded values to more accessible terms.***

Or, at minimum, it should link to a glossary.

When it comes to creating visualizations, the tool should offer users a full range of types to choose from, from histograms and scatter plots for numeric fields, to pivot tables for categorical fields. Ideally it will also allow researchers to quickly build visualizations from data in multiple fields, so they can hone their approach during initial exploration – an example would be a visualization showing blood pressure readings for individuals in different risk categories.

Key, as well, is the ability to manage the particularities of visualizing genomic data, allowing researchers to quickly see variant frequencies, and assess their impact. A good tool will make this easier, by providing annotations drawing on such resources as SnpEff and GnomAD. It will also enable users to do quick filtering searches, to hone in on particular genes or variants. As with filtering by category when analyzing clinical data, filtering by gene or variant is a useful way to refine cohorts, and to get quickly to a deeper understanding of a dataset.

Finally, a top-notch bioinformatics tool needs to support running advanced analyses on combined pheno-geno cohorts, then displaying the results using standard visualization types. These should include Manhattan plots, volcano plots, and clustering charts.



DNAnexus Apollo Cohort Browser Lollipop Chart

5 Democratizing Data Access & Speeding Time-to-Insights

Case Study: A Top 20 Pharmaceutical Company

Empowering Biologists

An R&D team at a top 20 pharmaceutical company is using UK Biobank data to study associations between genetics, lifestyle, and disease. They needed a way to draw on UKB clinical data, parsing the complex encoding systems, as they iteratively defined the phenotypic inclusion criteria for their studies. They turned to **DNAneuxus Apollo for UK Biobank** to use a suite of tools purpose-built for working with the UKB dataset.

With the Apollo Cohort Browser, the team has an easy and visual way to explore data and build cohorts based on both genomic and phenotypic traits. Team members can rapidly experiment with different cohorts, to ensure they've selected the right subpopulation for their study, without having to go to their data science colleagues for help. The team creates custom dashboards, populated with visualizations drawing on the phenotypic fields, and can easily compare two cohorts for case-control comparisons, and comparing subpopulations to the whole. Using the Cohort Browser, biologists are able to do the necessary work on their own, quickly creating, exploring, comparing, and saving a variety of cohorts.

5 Democratizing Data Access & Speeding Time-to-Insights

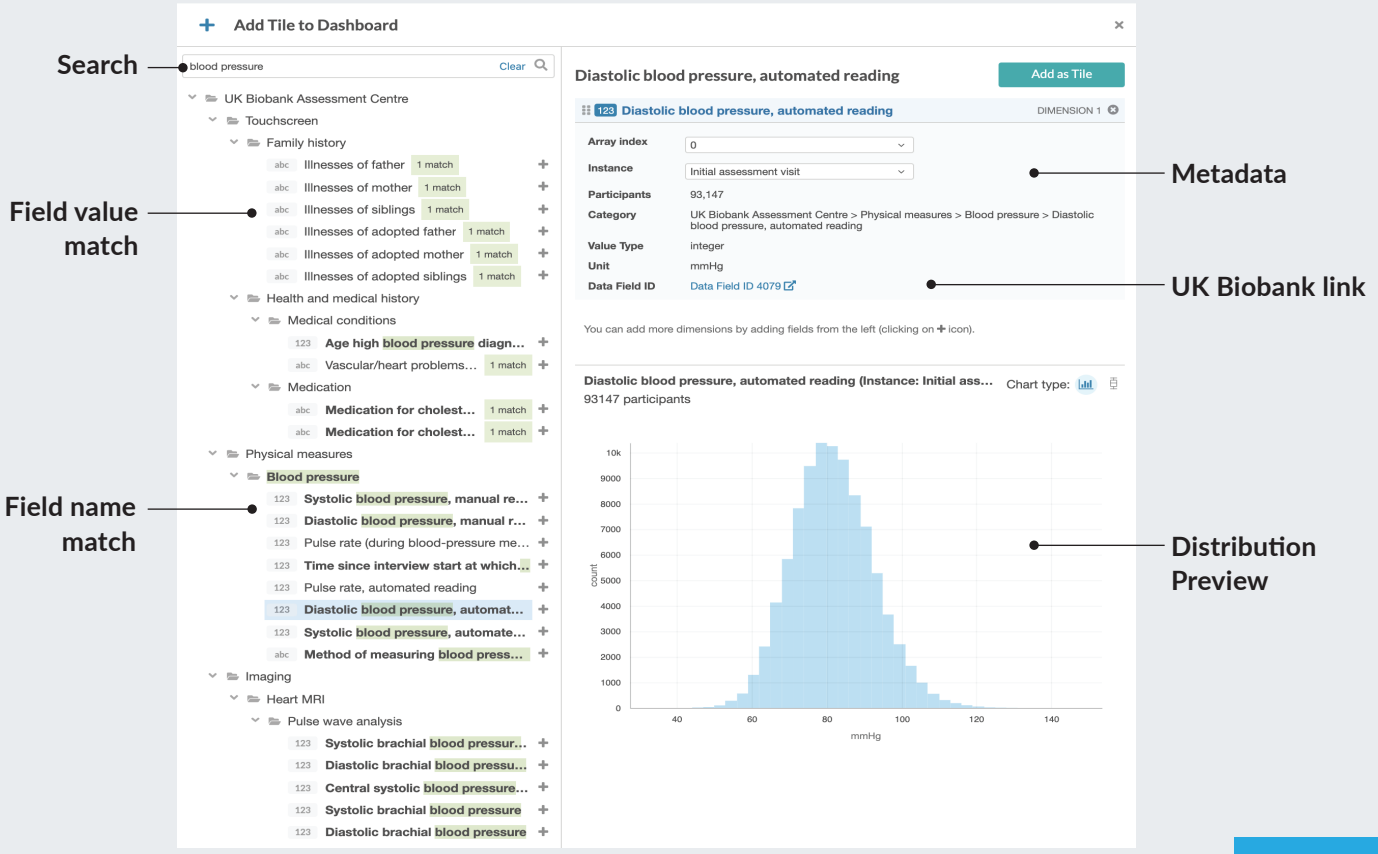
Case Study: A Top 20 Pharmaceutical Company

The Apollo Field Browser

One particular part of the Cohort Browser that has supported this team is the unique Apollo Field Browser (Shown below). The Field Browser shows available fields organized in a folder hierarchy that's similar to the one used in the UK Biobank Showcase. The research team can run keyword searches across folders, field names, and field values.

When a researcher selects a field, the Field Browser displays associated metadata, such as the type of data it contains and number of participants covered by the data, as well as a link to a page giving more details on the field, in the UKB Showcase. Where applicable, researchers can also see a graph showing the distribution of values.

In its effort to contribute to a greater understanding of how genetics, lifestyle, and disease are associated, the R&D team at this leading pharmaceutical company has found Apollo for UK Biobank to be an essential tool.

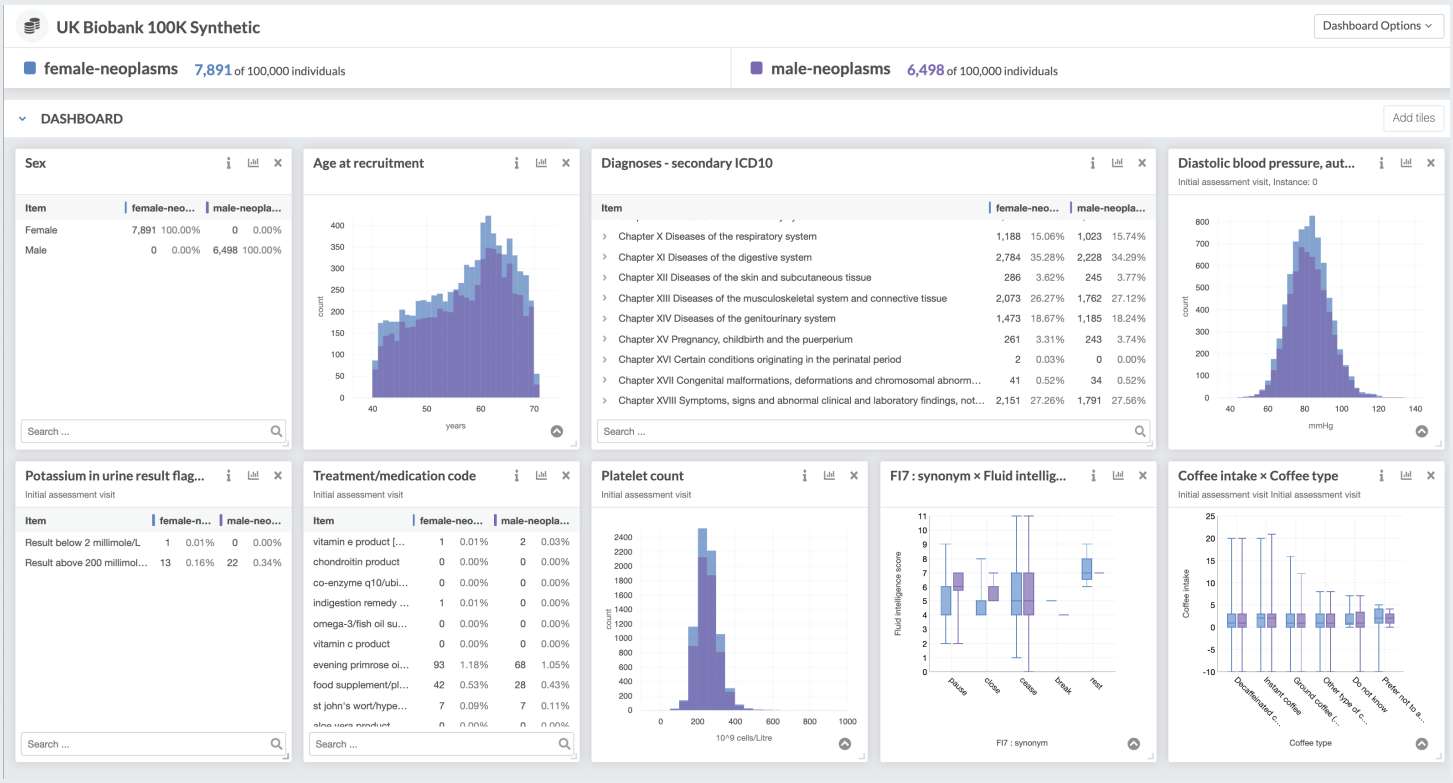


5 Democratizing Data Access & Speeding Time-to-Insights

Case Study: A Top 20 Pharmaceutical Company

Apollo has made the team more self-sufficient in its data exploration and analysis, particularly by enabling them to create and refine custom cohorts faster. Easier collaboration with their data science team has also been a benefit. Having created their custom cohorts within Apollo, the team was able to share them with the data science team to conduct advanced analyses in an embedded JupyterLab Notebook. The data science team then used Apollo to deliver GWAS and other analysis results back to the larger organization.

> ***By enabling a wide range of users of different specialties to work together on this project, Apollo has democratized the UKB dataset across the organization.***
> “



Conclusion: Don't Drown in UK Biobank Data

Explore, Analyze, & Discover with DNAnexus Apollo for UK Biobank

The UK Biobank is the single most useful resource for researchers working to glean insights from population-scale clinico-genomic dataset. But its sheer scale and richness can make it unwieldy to use. And having the right bioinformatics solution is essential to success.

Your solution should, above all, be able to scale to enable you to explore and do complex analyses on a dataset that incorporates trillions of data points. Many legacy systems can't do this, and so are inadequate to work with the UKB dataset. Moreover, these systems often can't meet the stringent privacy and security requirements, for even getting access to this data.

Your solution also needs to enable you to work effectively with UKB data. This entails support for the full range of tasks you'll engage in, over the course of a project – from initial exploration, building and comparing, to running complex analyses and visualizing the results. Specialized navigation and visualization tools are essential throughout, speeding your progress, and enabling you to structure your work and carry it out in such a way, as to get to maximally useful insights.

When looking for a bioinformatics solution for your work with UKB data, choose one that can support everything you and your team need to do, over the course of a study that draws on this incredibly rich resource. Look also for a solution that will scale effortlessly, as your analyses become more complex. A trusted cloud-based solution, such as DNAnexus Apollo for UK Biobank, will help you get the most out of working with the UK Biobank dataset.

DNAnexus® Apollo for UK Biobank

Get to Insights Faster

DNAnexus has been at the forefront of scalable informatics. DNAnexus Apollo for UK Biobank was purpose-built to support researcher's important breakthroughs by providing a fast and easy way to explore and analyze this rich dataset.

DNAnexus was used by scientists as part of their effort to sequence, analyze and deliver the WES data of the UK Biobank study. As part of this delivery, scientists successfully deployed the cohort browser on a collection of thousands of phenotypic fields extracted from the UKB and millions of genetic variants computed through its scientific pipelines.

Apollo for UK Biobank Empowers Researchers to:

- Explore possible correlations across population-wide genomic and phenotypic data
- Easily create and compare the genetic and phenotypic features of custom cohorts
- Quickly conduct sophisticated, iterative analyses of cohort data, using the embedded JupyterLab environment

Dedicated to Your Success.

DNAnexus® has built the world's most secure platform and global network for scientific collaboration and accelerated discovery. We embrace challenge and partnership to tackle the world's most exciting opportunities in human health.

Contact us at:
info@dnanexus.com

For more information:
www.dnanexus.com/product-overview/apollo/apollo-for-ukb