Data Ark Data Commons Town Hall December 2022

Maria Julia Castro, Project Manager, GGS & Scientific Computing and Data
Jielin Yu, Computational Scientist, Scientific Computing and Data
Lili Gai, Director, High Performance Computing (HPC)
Patricia Kovatch, Dean for Scientific Computing and Data & Professor, GGS
Paul O'Reilly, Data Ark Science Director & Associate Professor, GGS
Bruce Gelb, Chair, Data Ark Advisory Committee

Mount Sinai

December 2nd, 2022

Outline

- Overview of Data Ark and Updates
- New Data Sets & Pending Data Sets
- 2022 Usage Summary
- Data Ark’s 2023 Goals

Data Ark Mission

The purpose of the Data Ark is to ensure that scientists at Mount Sinai have all the data they need to maximize the power of their science. These data should be easy-to-access, in analysis-ready form and be as large and diverse as possible.

What is Data Ark?

Data Ark Mount Sinai Data Commons
Funded by **GGS** and **Scientific Computing**

Increasing the power, pace and relevance of our science

Power: Our researchers could be using ~20x more data

Pace: Users will have rapid access to huge powerful data

Relevance: Sinai can be a world leader in biomedical science

- What is the Data Ark?
  - **Space on Minerva** to host all frequent-use research data sets (UK Biobank, GTEx, COVID Biobank…)
  - **A team** of data scientists/engineers to manage resource, process data, simplify access process
  - **An opportunity** for a step-change in the power and pace of Sinai research true 'big data science'

Data Ark is part of Mount Sinai's Computational and Data Ecosystem

- Co-locating images, genomic, EHR and other data sets with the compute enables large-scale, multi-modal and multi-scale analyses
- Utilizing high-performance computing accelerates analyses
- Enabling researchers to directly query data maximizes accessibility

Data Ark Data Sets Summary

Public Data Sets (Unrestricted)	
1,000 Genomes Project	Phase 3 individual-level called genotype data(VCF) on 2500 individuals of mixed ancestry.
GWAS Summary Stats	Genome Wide Association Studies results in a standardized format across thousands of outcomes
GTEx	Gene expression data collected from multiple tissue types(up to 54) from ~960 deceased donors
gnomAD	The Genome Aggregation Database-standardized WES/WGS processing from a wide variety of large-scale sequencing projects
Open Access TCGA	The Cancer Genome Atlas (TCGA) “ Open-access” RNA-seq counts and WXS data with pre-processing and QC.
eQTL Gen	The eQTLGen Consortium has been set up to identify the downstream consequences of trait-related genetic variants.
UKBB-LD	UKBB-LD is summary linkage disequilibrium (LD) matrices files computed from UK Biobank (UKBB) based on N=337K British-ancestry individuals.
LDSCORE	The baseline LD (linkage disequilibrium) scores information is generated by the Broad Institute in helping with LD Score regression analysis.
BLAST	You can find the nr database, the UniProt Reference Clusters (Uniref) including UniRef50, UniRef90, and UniRef100
Reference Genome	Data Ark hosts the most frequently used reference genome files and related annotations. We have fasta files--GRCh37 and GRCh38 for Human and GRCm38 and GRCm39 for Mouse

Data Ark Data Sets Summary

Public Data Sets (Restricted)	
UK Biobank	Genetic data (genotype/WES) from the UK Biobank data on 500,000 individuals
Mount Sinai Generated Data (Restricted)	
STOP COVID NYC Cohort	Symptom and behavior on COVID-19 on ~50,000 New York City residents survey data
MSDW COVID-19 EHR data	De-identified clinical data on patients from Caboodle with or suspected of COVID-19 containing 350 data elements and updated daily
The Mount Sinai COVID-19 Biobank	Blood samples from hundreds of COVID-19 patients hospitalized at Mount Sinai, with genotype/WGS data available
The Living Brain Project	The Living Brain Project (LBP) is a multiscale, data-driven investigation of the human brain wherein a single living population is being studied using the full human subject neuroscience toolkit, conducted by the Laboratory of Brain and Data Sciences.
Data Set Supplements	
The Imaging Research Warehouse	The IRW integrates clinical imaging with electronic health records, and as it expands it will give researchers new access to information about more than 1 million Mount Sinai patients
The BioMe BioBank Program	Bio*Me* has a wide array of phenotypic and genetic data available for use by researchers, and includes a diverse cohort of individuals from many ancestral and cultural backgrounds.
CIB (Cancer Institute Biorepository)	CIB, in combination with Freezerworks, supports the biorepository to annotate, manage, and search donor and sample (tissue and fluid) information, consent status, clinical annotations, and sample tracking.

Data Ark Data Sets Summary

Cloud-Hosted Data Sets (Restricted)	
All of Us	The All of Us Research Hub stores health data from a diverse group of participants from across the United States.
School-Acquired Data Sets (Restricted)	
IBM MarketScan	MarketScan® Research Databases from IBM® provides one of the longest-running and largest collections of proprietary de-identified claims data for privately and publicly insured people in the U.S.

New Data Sets Added in Q3

eQTLGen

- The eQTLGen Consortium has been set up to identify the downstream consequences of trait-related genetic variants.
- To investigate the genetics of gene expression, the researcher group performed cis- and trans-expression quantitative trait locus (eQTL) analyses using a blood-derived expression in a total of 31,684 individuals. You can find the cis-eQTL, trans-eQTL, eQTS, replication, and single-cell eQTLGen data folders on the Data Ark.

UKBB-LD

- UKBB-LD is summary (LD) matrices files computed from UK Biobank (UKBB) based on N=337K British-ancestry individuals.
- The LD information is stored as 2,763 3Mb-long regions spanning the entire genome. This data set can be used for post-Genome-wide association studies (GWAS) analysis such as fine-mapping. The data is generated from Alkes Price's group at Harvard.

New Data Sets Added in Q3

Reference Genome

- Data Ark hosts the most frequently used reference genome files and related annotations. We have fasta files--GRCh37 and GRCh38 for Human and GRCm38 and GRCm39 for Mouse.
- In addition, we host Ensembl, Gencode, and Refseq for different annotation purposes.

LDSCORE

- The baseline LD (linkage disequilibrium) scores information is generated by the Broad Institute in helping with LD Score regression analysis.

BLAST

- Data Ark hosts the National Center for Biotechnology Information (NCBI) Basic Local Alignment Search Tool (BLAST) database at the users' request.
- You can find the nr database, the UniProt Reference Clusters (Uniref) including UniRef50, UniRef90, and UniRef100.

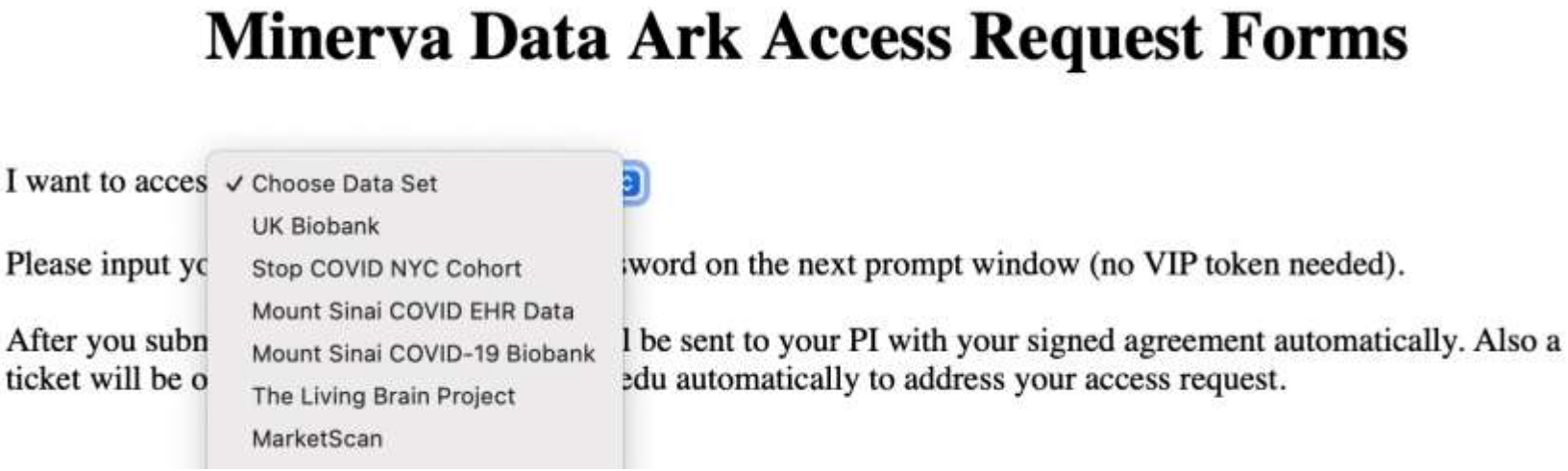
Other data sets in the Data Ark ecosystem

- The Living Brain Project
  - The Living Brain Project (LBP) is a multiscale, data-driven investigation of the human brain wherein a single living population is being studied using the full human subject neuroscience toolkit, conducted by the Laboratory of Brain and Data Sciences.
  - The LBP cohort consists of prefrontal cortex (PFC) samples from a living brain cohort (LIV) matched for age and sex to a post-mortem brain cohort (PM).
  - Data Ark hosts several RNAseq data files that have passed the quality control test. RNA sequencing was performed on 289 living brain samples from 172 subjects and 248 post-mortem samples from 248 individuals.

- IBM MarketScan
  - MarketScan® Research Databases from IBM® provides one of the longest-running and largest collections of de-identified claims data for privately and publicly insured people in the U.S.
  - MarketScan Commercial Claims and Encounters Database (CCAE) – contains data from active employees, early retirees, COBRA continuees, and dependents insured by employer sponsored plans (i.e., individuals not eligible for Medicare).
  - MarketScan Medicare Supplemental and Coordination of Benefits Database (MDCR) – created for Medicare-eligible retirees with employer-sponsored Medicare Supplemental plans. This database contains predominantly fee-for-service plan data.

Accessing Data Ark

https://labs.icahn.mssm.edu/minervalab/resources/data-ark/



- **Login to the form with your Minerva ID** within Mount Sinai campus network or school VPN . If you haven't used Minerva before, follow this link https://acctreq.hpc.mssm.edu/ to request for a Minerva user account

Accessing Data Ark on Minerva

- For **Public Unrestricted** data sets, **NO Data Use Agreement Form required**

  you can access the data from the following path on Minerva:

  /sc/arion/projects/data-ark/Public_Unrestricted

  Or you can load module **$module load data ark** to see the path variables.

- For any other data sets, you must read and agree the **Agreement** specific to each data set that you want to access

Data Ark Usage

Data Ark Usage Summary

Dataset	Size (GB)	# of Users Q1	# of Users Q2	# of Users Q3	# of Users Q4	# of times accessed
1000 Genomes	143	21	34	9	36	1,213
GTEx	1,888	17	30	4	32	19
GWAS	936	18	30	7	30	42,826
UK Biobank	11,239	40	54	16	64	3,121,363
STOP COVID NYC Cohort	1	1	6	3	9	23
Mount Sinai Data Warehouse COVID-19 EHR	1.7	8	10	5	10	23
COVID-19 Biobank	25	0	8	4	11	55
gnomAD	8,628			6	6	1496
TCGA	155			3	3	11
eQTLGen	39			4	4	29
UKBB-LD	2,867			5	5	945
Reference genome	143			3	3	9
The Living Brain Project	1			0	0	0

~8% genomic-related users used Data Ark on Minerva for the 1st year of Data Ark

Data Ark user support tickets received: 79 since Jan 2022

- Tickets for assistance on DUA form: 41
- Tickets for assistance on Data Ark sets: 38

What’s next

Data Ark 2023 Goals

- Adding new and frequently used data sets:
  - BioMe: adding anonymized phenotypic, genotype (microarray + WES data), and (aggregated) return data sets for BioMe
  - Human Cell Atlas
  - and others suggested by users and advisory board members

- Distribute Annual User Survey in Jan 2023

- Hire dedicated bioinformatician

- Marketing:
  - Hold more seminars and training sessions on Data Ark to increase awareness and usage
  - Promote Data Ark through GGS website, faculty meetings, retreats, and research engagement events.

Contacting the Data Ark Team

We want to hear from you!

Submit which data sets will be useful for your research:

Suggest a Data Set Survey

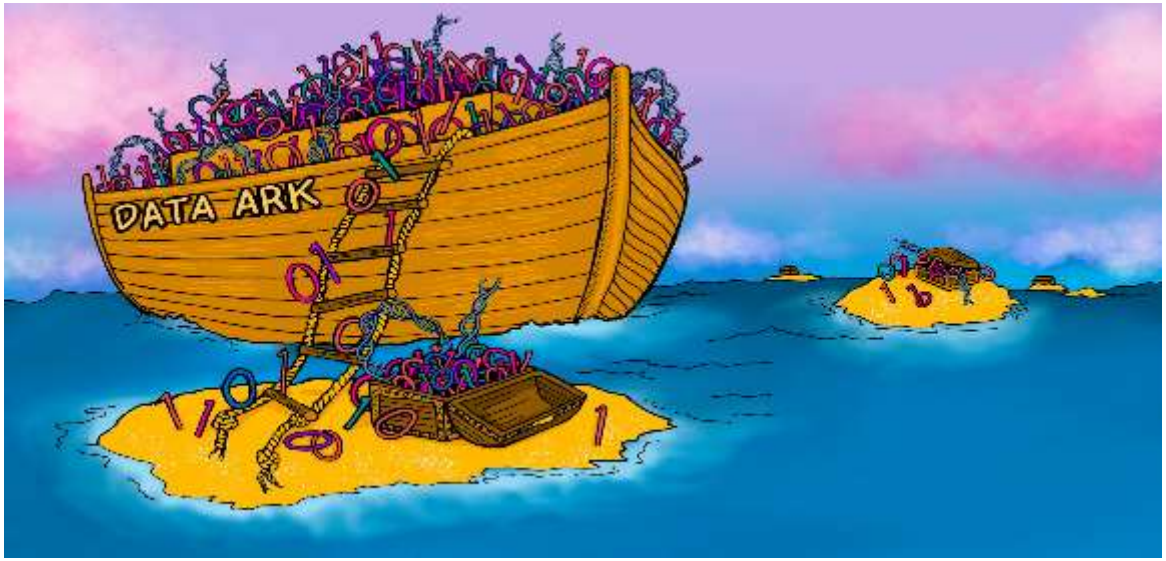
Also please refer to our Data Ark's data set onboarding & retention policy:

- Onboarding Data Sets:
  - PIs will fill out the onboarding form and list expected research groups
  - Approval process according to data set size:
    - =<1 TB: Data Ark operations team will approve
    - >1 TB: must be approved by the Data Ark Advisory Board

- Retention Period:
  - Expected usage after 1 year is commensurate
  - If usage is low, then the data sets will be removed from Data Ark
  - The original data owner will receive usage reports every quarter and be alerted when others have not used their sets

How to interact with the Data Ark team

- Anyone can contact the Data Ark team with questions or ticket submissions by writing to data-ark-team@list.mssm.edu.
- Data Ark Slack community ---we have channels for every common data set. To join the channel, navigate to https://join.slack.com/t/data-ark/signup and sign up using your Mount Sinai credentials. You'll be able to start interacting with other researchers on common data sets right away.
- Click here for more information!



Thank you!