Data Ark Data Commons Town Hall May 2022

Wen Huang, Bioinformatician, Scientific Computing and Data
Maria Julia Castro, Project Manager, Genetics and Genomic Sciences (GGS)
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

May 25, 2022

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.

Outline

- Overview of Data Ark and Updates
- 2022 Data Ark Usage
- 2022 User Survey Results
- Data Ark Future Plans

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.
Public Data Sets (Restricted)	
UK Biobank	Genetic data (genotype/WES) from the UK Biobank data on 500,000 individuals
Mount Sinai Generated Data	
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

New Data Sets Added in 2022

The Genome Aggregation Database (gnomAD) V2.1.1 and v3.1.2

- The Gold standard resource for variant interpretation
- Harmonize large-scale whole exome and whole genome sequencing data
- Aligned against both GRCh37/hg19 and GRCh38/hg38 reference genome
- v3.1.2 includes genomes from Human Genome Diversity Project (HGDP) and 1000 Genome Project (1KG)

The Cancer Genome Atlas (TCGA) Open Access

- Landmark cancer genomics program that characterized over 20,000 primary cancer and normal samples spanning 33 cancer types
- Preprocessed, QCed RNA-seq counts data(over 11,000 samples) into 33 different outcomes
- Open-access biospecimens, clinical, WXS (Mutation Annotation Format MAF files), cBioPortal data within the TCGA database are also available in Data Ark

Other data sets in the Data Ark ecosystem

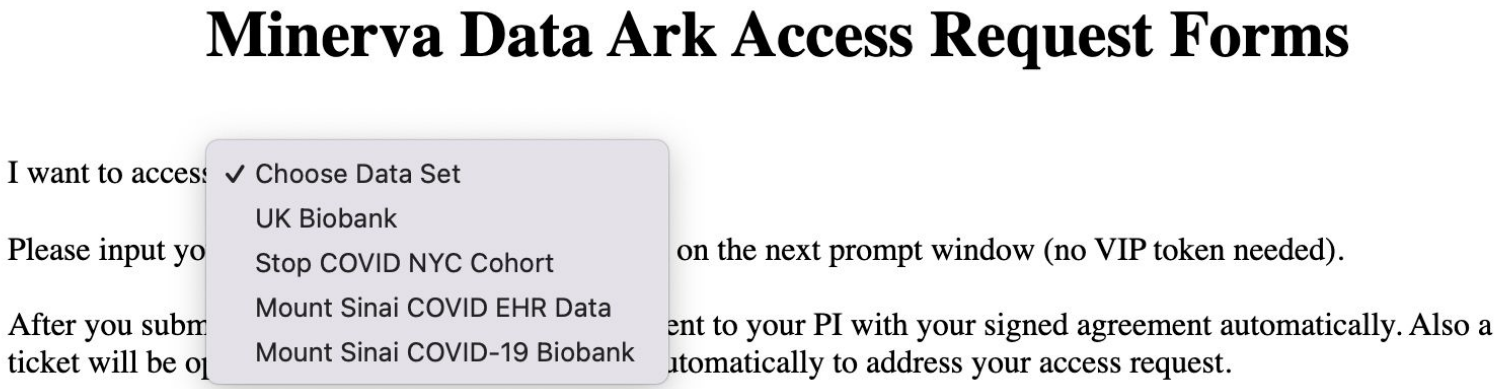
- The Mount Sinai **Bio*Me*® BioBank** Program
  - An electronic medical record-linked blood/serum biobank with over 50,000 enrolled participants with genetic, epidemiologic and molecular data are available on these participants, including whole exome sequencing data (WES) for a diverse cohort of individuals from over 30,000 with diverse ancestral and cultural backgrounds

- The Mount Sinai Cancer Institute Biorepository (CIB)
  - Pathology tissue specimen database with donor and sample (tissue and fluid) information, consent status, clinical annotations, and sample tracking

- The Mount Sinai Imaging Research Warehouse 1.0
  - Contains over 217 million de-identified image slices for over 700,000 MSHS studies from 2017-2021 in image modalities including DX, CT, CR, MR, MG and NM

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 User 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
(Jan 2022 - May 2022)

Data Ark Usage Summary – Year to date 2022

Dataset	Data set size (Gigabytes)	Total # of users given access by Q1	Total # of users given access in Q2	# of unique users accessed
1000 Genomes	137	21	34	14
GTEx	1,900	17	30	12
GWAS	506	18	30	10
UK Biobank	11,000	40	54	18
STOP COVID NYC Cohort	1	1	6	5
Mount Sinai Data Warehouse (MSDW) COVID-19 EHR	1.7	8	10	9
COVID-19 Biobank	25	0	8	4
gnomAD	8,500			
TCGA	155			

Data Ark user support tickets received: 52 since Jan 2022

	Q1 2022	Q2 2022
Access requests via DUA	24	6
Questions about using the data sets	13	9

Data Use Agreement (DUA) is no longer required for public unrestricted data sets (April 2022)

User Survey Results 2022

2022 Data Ark Survey Questions and Initial Results

We asked the following questions:

1. What data sets would you like to see added?
2. What features would you like to see added to the Data Ark?
3. What barriers exit preventing your usage of Data Ark?
4. Additional feedback or comments:

20 responses so far:

Used Data Ark	16%	**Yes**
	84%	**No**
Plan to Use DA	42%	**Maybe**
	32%	**Yes**
	5%	**No**
	21%	**No reply**

Thank you for your feedback! This is the motivation for our 2022 Data Ark Plans!

Data Ark User Survey feedback: Question 1

1. What data sets would you like to see added?

Data set list requested from the survey

- Epidemiology sets
- TCGA
- COVID-19 biobank
- Mount Sinai Brain Bank
- RNAseq and eQTL datasets
- Human Genomes Diversity Project
- MassIVE (UCSD)
- BioMe Biobank
- 1000 Genomes Project
- Gabriella Miller Kids First Pediatric Research Program WGS data
- PsychENCODE
- ROSMAP
- Full Mount Sinai Biobank
- UK Biobank
- Sleep Studies
- CommonMind
- gnomAD
- Molecular dynamics trajectories

Next Steps: Onboard new data sets as prioritized by the Data Ark Advisory Committee

Data Ark User Survey feedback: Question 2

2. What features would you like to see added?:

- A codebook to understand what variables are stored
- Neuroimaging data
- Streamlined pipelines for processing most commonly used data types (WGS, RNAseq)
- Better folder organization, with dataset release/versions/formats clearly separated

Next Steps:

- Set up environmental modules on Data Ark on Minerva ✔
- Create a more informative “readme” file under each data set on Minerva

Data Ark User Survey feedback: Question 3

3. What barriers prevent you from using Data Ark?

- Confusion on access and how to use resources
- Instructions are unclear
- Limited interactions with Data Ark team
- Lack of knowledge
- Unclear about permissions/IRB protocol

Next Steps:

- We will update the Data Ark website to be more clear that there are no user fees associated with Data Ark and there is a user support team available

Data Ark User Survey feedback: Question 4

4. Additional feedback/comments?

- More descriptive EHR information that could be uniformly processed
- Review what the most common data problems are and universal ISMMS data sets
- Instructions to navigate types of data and how to use the data to answer certain research questions
- Create a web app with searchable metadata index of the various datasets through which to access information about each dataset and dataset releases/versions/formats

Next Steps:

- We are updating our website with more descriptive EHR information
- We are working to help make the data dictionaries and metadata searchable in the new research data catalog

What’s next

Data Ark 2022 Goals

- Add the reference genome data set
- Work with the advisory committee to develop a policy to onboard and offboard data sets
- Onboard new data sets as prioritized by the Data Ark Advisory Committee
- Hold more seminars on Data Ark to increase awareness and usage

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

https://redcap.mountsinai.org/redcap/surveys/?s=LLTARNCPP7HYYT9X

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!