Data Ark Data Commons Town Hall October 2023

Yiyuan Liu, Bioinformatician, Scientific Computing and Data

Lili Gai, Director, High Performance Computing

Patricia Kovatch, Dean for Scientific Computing & Data

Bruce Gelb, Chair, Data Ark Committee

Cyrus Hedvat, Data Ark Clinical Science Director

Paul O’Reilly, Data Ark Science Director

October 24, 2023

Icahn School of Medicine at **Mount Sinai**

Outline

Brief Introduction to Data Ark

Accomplishments & Updates

- Usage summary and support tickets
- Research engagement updates
- New dataset onboarded - MSDW OMOP data mart
- Submitted IRB for Data Ark for BioMe onboarding
- Policy review on public datasets
- A user group model built for the MarketScan dataset

What's next?

- 2024 Roadmap

Q&A

Meet the Data Ark Operations Team

Patricia Kovatch

Professor and Dean for Scientific Computing and Data

Ranjini Kottaiyan, MBA

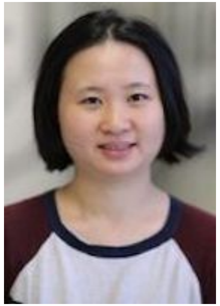
Senior Director

Maria Julia Castro, MS

Project Manager

Lili Gai, PhD

HPC Director



Yiyuan Liu, PhD

Bioinformatician

Introduction to Data Ark

Data Ark Data Commons Increases the Power, Pace and Relevance of Our Science

Challenges

- Exhaustive searches for relevant datasets
- Repeated downloads of the same files across groups
- Difficulty in understanding opaque data structures

How Data Ark Helps

- Storage space for frequent-use research datasets
- A team managing the resources, simplifying access, training and user support

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

The Data Ark Data Commons Is on Minerva HPC and Growing

There are 18 datasets hosted under Data Ark currently

- Immediate access to public-unrestricted data sets - **may subject to changes according to MSIP guidance**
- Access within 24 hours to Mount Sinai-generated data sets

Immediate Access

Public Data Sets

- GTEx
- GWAS Summary Stats
- gnomAD
- eQTLGen
- BLAST
- Reference Genome
- Genebass
- 1,000 Genomes Project
- UKBB-LD
- Partial of the Cancer Genome Atlas (TCGA)
- LDSCORE

Access within 24 hours

Mount Sinai Generated Data

- The CBIPM-BioMe Data Set (coming soon)
- MSDW De-identified OMOP Data set
- MSDW COVID-19 EHR Data Set
- Mount Sinai COVID-19 Biobank
- The Living Brain Project
- STOP COVID NYC Cohort

Restricted Access

Public Data Sets

- UK Biobank **(changes to be released soon)**

User Group-Acquired Data Sets

- MarketScan®

Responsibilities for Data Ark vs. Users in the Terms of Service Agreement

What Data Ark does

- **Dataset management,**
  - a. data ascertainment, updates, backup and archiving for offboarding.
- **User support,**
  - a. Q&A through email tickets and digital concierge;
  - b. Data Ark website and broadcast information;
  - c. present at outreach events;
  - d. provide bi-annual user training;
  - e. present at bi-annual town hall meetings.
  - f. data analysis currently **NOT** supported
- **Access and permission control,**
- **Data usage metrics analysis,**
- **User satisfaction survey.**

What users do

- **For data access**,
  - a. register an Minerva account.
  - b. review all policies and sign as needed
- **For data onboarding**,
  - a. fill in the survey to indicate other Sinai user groups;
  - b. provide data source link and data description
  - c. provide/help on Data Use Agreement (DUA)
- **For data use**,
  - a. read through documentations on the data and usage;
  - b. manage project-specific tasks;
  - c. inquire the data owner with project-specific questions;
  - d. respond and submit data use report if required;
  - e. acknowledge Data Ark in publications.

Accomplishment & Updates

Data Ark Usage Summary (Jan 2023 - Sep 2023)

# of active users: **52**
# of support tickets: **107**

Dataset	Size (GB)	# of unique users	# of times data accessed
TCGA	154	10	8,421,194
Genebass	903	9	291,566
gnomAD	8,628	13	247,245
UK Biobank LD	2,866	15	212,175
UK Biobank	12,618	17	135,154
GWAS Summary Statistics	6,826	14	86,675
1000 Genome	143	17	45,492
GTEx	1,888	15	10,702
Reference Genome	142	12	4,872
Blast	1,116	8	2,811
LD Scoree Regression	173	10	2,296
MarketScan	2,265	15	678
eQTLGen	39	8	306

Researcher Engagement Highlights

Events	Date	Attendees	Delivered Service
Sinai CTSA Research Day	6/09/23	121 trainees, junior faculties, and invited speakers	Awareness of Data Ark offerings, services and expertise
TCI Basic and Translational Committee Meeting	6/12/23	30 TCI faculties	Mapping out collaborative opportunities between TCI and Data Ark
Data Ark Advisory Board Meeting	10/16/23	8 Mount Sinai faculties	Engaging in constructive dialogue about current operations, challenges and emerging opportunities
Data Ark Training	10/18/23	32 Institute-wide researchers	Data Ark datasets; UK Biobank data access

One New Dataset Added

Mount Sinai Data Warehouse (MSDW) OMOP (Observational Medical Outcomes Partnership) Data:

- The MSDW de-identified OMOP dataset contains over 11 million patient records and over 87 million patient encounters of Epic data.
- Plain text files directly ingestible by programming tools.

Submitted an IRB Protocol

- Why?
  - Onboarding BioMe data is contingent on the approval of Data Ark IRB

- The current status:
  - Submitted on Sep 27, 2023 and under review

Updates: Policy Review on Public Data Sets

- ✔ MSIP examined the Data Use Agreements for public datasets and provided us guidance on how to proceed.
- ✔ We requested approval from original data owners to redistribute within Mount Sinai.
- ➔ Data use agreements may be reinstated for all public datasets on Data Ark soon.

Datasets	Status
1000 Genome **LD Score Regression** **TCGA** **UK Biobank LD**	● Pending approval from the data owners ● Will update soon once confirmed
GWAS Summary Stats **(IEU & Psychiatric Genomics Consortium)**	● Removed from Data Ark ● Emailed the current GWAS users on 10/05/23

We Developed a User Group model for MarketScan Data: A Blueprint for Managing Datasets of a Small User Group

Data Ark provides **basic support** for small user groups

- For specialized, shared, easy-to-onboard datasets with a small user group, the Data Ark team will load the dataset and grant access (per the user group lead)
- All other duties (licensing, funding, storage costs, domain expertise, etc.) must be covered by the user group

What Data Ark does

- **Grant access permission**
- **Provide technical support**
  a. SAS license renewal;
  b. tasks common to managing other datasets.

What the user group does

- Negotiate license cost and terms negotiation with the data providers and among the members
- Hire analysts expert in a dataset
- Create a Data Use Agreement
- Manage the list of users with access authorization

Update on MarketScan Dataset

- Current school-acquired MarketScan license will be expiring 11/04/2023.
- License renewal is managed by the MarketScan User Group led by Dr. Parul Agarwal.
    - Users access list is provided by the User Group. The rest will be removed immediately after 11/04/2023.
    - We will send out an announcement soon.

2024 Roadmap

Expanding Data Ark: Proposed Dataset Additions

Dataset	Data Owner	Date	Requesting PI	Interested PI(s)	Data Type	Records	Data Size	Status
Digital Pathology Slides	Mount Sinai Health System	Since June 2019	Thomas Fuchs	Thomas Fuchs	Imaging	1.2 million	1.3 PB	On Minerva, to be linked with MSDW EHR data
Human Genome Diversity Project	Stanford University's Morrison Institute	1990s (established); 2020 (WGS released)	Samira Asgari	Kuan-lin Huang, Vikas Pejaver	Whole-genome sequencing (WGS)	929 study subjects	318 GB	Downloaded to Minerva; to be screened for data redistribution restrictions
Simon Genome Diversity Project	David Reich (PI), Broad Institute	2016 released	Samira Asgari	Vikas Pejaver	Whole-genome sequencing	279 study subjects	< 1 TB	To be downloaded

Boosting User Engagement Through Upcoming Training

Events	Date	Target Attendees	Delivered Service
GGS Faculty Meeting	10/26/23	GGS faculties	Latest updates on Data Ark and roadmap ahead
Lunch and Learn	To be confirmed	Masters and PhD students, postdoctoral researchers and CTSA funded assistant professors	Data Ark datasets; demonstration of analytical use of datasets

Coming Soon: Important Changes to UK Biobank Hosting on Data Ark

- **Why?** To align with the mission to provide access to shared data on Data Ark.
- **Current Status:**
  - 8 UK Biobank application datasets hosted on Data Ark, including application-specific clinical and common genetic data
- **Changes soon to be announced:**
  - Only a single copy of genetic data retained on Data Ark
    - The practice of sharing UK Biobank genetic data across applications by symlinking to a single application **has not been officially confirmed** by UK Biobank
  - The current application-specific datasets will be migrated to project folders
  - Users will need to maintain their own approved application

Submitting Your Request Through Our Ticket System Ensures a Streamlined Process

All service requests must come through the ticket system.

- This will streamline ticket tracking and ensure redundancy across team members
- Jielin Yu, Computational Scientist of the HPC team, provides service backup
- Ticket system: data-ark-team@lists.mssm.edu

Crediting CTSA in Your Publications Ensures Continued Access to Our Services

- Data Ark and Minerva funded by CTSA - crediting CTSA support is imperative

- **Annual user agreement required starting January 1, 2024 – new and existing users**
  - With logging into one site, you will see a list of form for you to review and sign annually

- Please acknowledge the support from Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai by including the following acknowledgement in a publication of any material, whether copyrighted or not, based on or developed with Data Ark resources: *"This work was supported in part through the computational resources and staff expertise provided by Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai and supported by the Clinical and Translational Science Awards (CTSA) grant UL1TR004419 from the National Center for Advancing Translational Sciences."*

CTSA Clinical & Translational Science Awards®

DATA ARK

Thank You!