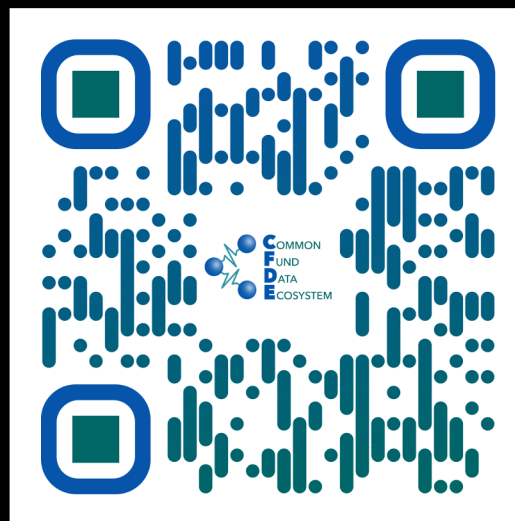


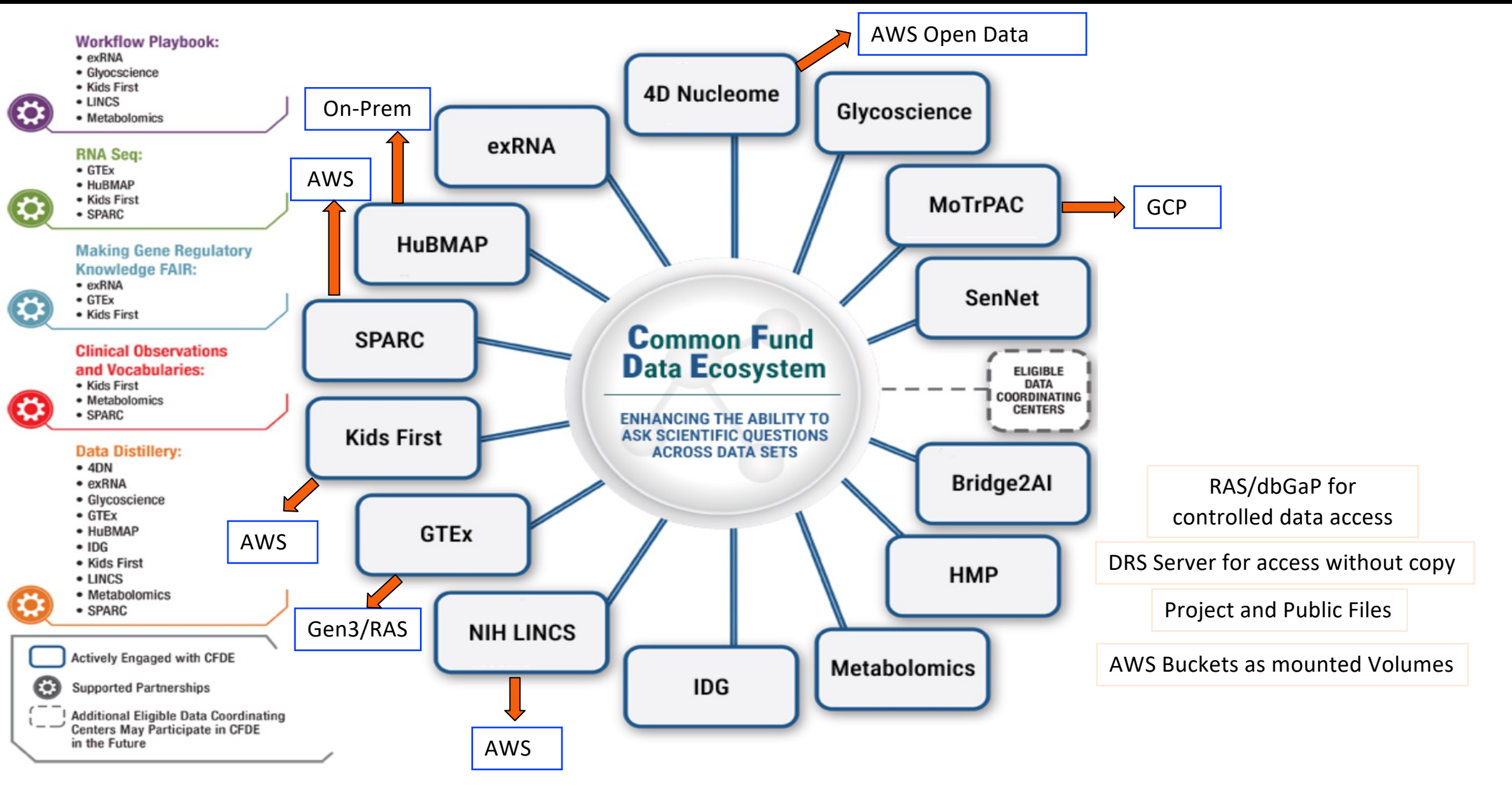
Cloud Workspace Pilot.

CFDE All Hands Training


Sangeeta Shukla
Eric Tobin
20 March 2024

VELSERA





Big Data Sourcing and Management within CAVATICA

- Address per project storage limitations
 - Remove redundancy in ELT performance
 - Projects can handle files found in different storage space
 - Easier retrieval of Public Files
 - Enable re-use by multiple projects/pipelines running simultaneously
 - Controlled access at multiple points
 - Processed result files can be reused
 - Data Studio to allow data manipulation
- 



CAVATICA

CAVATICA is a **data analysis** and **sharing platform** designed to accelerate **discovery** in a **scalable, cloud-based** compute environment where **data, results**, and **workflows** are **shared** among the world's research community. Developed by Seven Bridges and funded in-part by a grant from the National Institutes of Health (NIH) Common Fund, CAVATICA is continuously updated with new **tools** and **datasets**.

Our goal is to help researchers **collaborate, share, interoperate**, and **connect** with any and all other data ecosystems in order to empower data analysis across diseases, ages and geography. By connecting previously disconnected datasets, CAVATICA supports researchers and patients across the United States and throughout the world to generate new insights into pediatric diseases.

The Cavatica Platform has been funded in whole or in part with Federal funds from the National Institutes of Health, Contract No. U2C HL138346 and funding from the Center for Data-Driven Discovery in Biomedicine (D³b)



CAVATICA

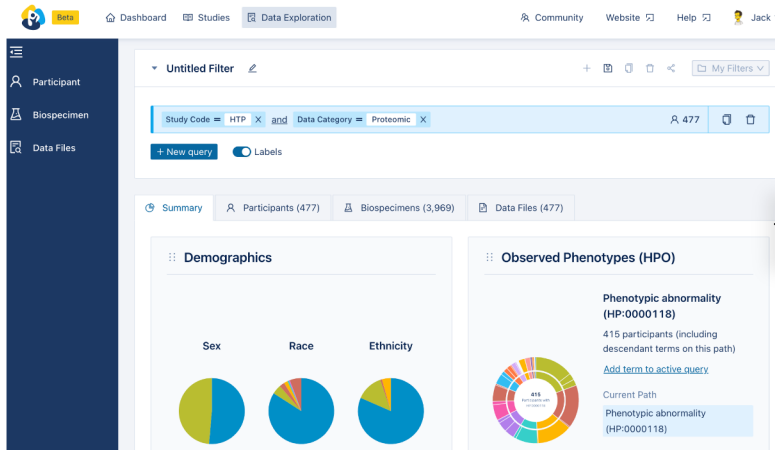
CAVATICA is a **data analysis** and **sharing platform** designed to accelerate **discovery** in a **scalable, cloud-based** compute environment where **data, results,** and **workflows** are **shared** among the world's research community. Developed by Seven Bridges and funded in-part by a grant from the National Institutes of Health (NIH) Common Fund, CAVATICA is continuously updated with new **tools** and **datasets**.

Our goal is to help researchers **collaborate, share, interoperate,** and **connect** with any and all other data ecosystems in order to empower data analysis across diseases, ages and geography. By connecting previously disconnected datasets, CAVATICA supports researchers and patients across the United States and throughout the world to generate new insights into pediatric diseases.

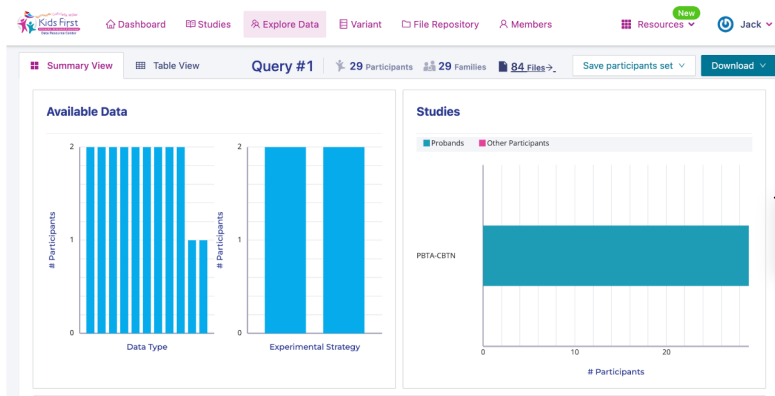
The Cavatica Platform has been funded in whole or in part with Federal funds from the National Institutes of Health, Contract No. U2C HL138346 and funding from the Center for Data-Driven Discovery in Biomedicine (D³b)

Empowering a one-to-many integration in the cloud

INCLUDE



Kids First



VELSERA



CAVATICA

TOPMed

AnVIL

TCGA

TARGET

PDC

ICDC

UDN

NIH National Heart, Lung, and Blood Institute | BioData CATALYST
Powered by Seven Bridges



Interoperability stack has provided the ability to have datasets from different "places" together in single platform

Interoperable data on CAVATICA via DRS

The screenshot shows the CAVATICA interface for a project named "CFDE DRS manifest testbed". The interface includes a navigation bar with "Files" selected, a search bar, and filter options for Extension, Sample ID, Task ID, and Tags. A table lists several DRS files with columns for Name, Created on, Extension, and Size. Three red arrows point from the "Name" column to logos for Kids First, HMP, and GTEX. The Kids First logo is for the Gabriella Miller Kids First Pediatric Research Program Data Resource Center. The HMP logo is for the Human Microbiome Project. The GTEX logo is for the Genotype-Tissue Expression project.

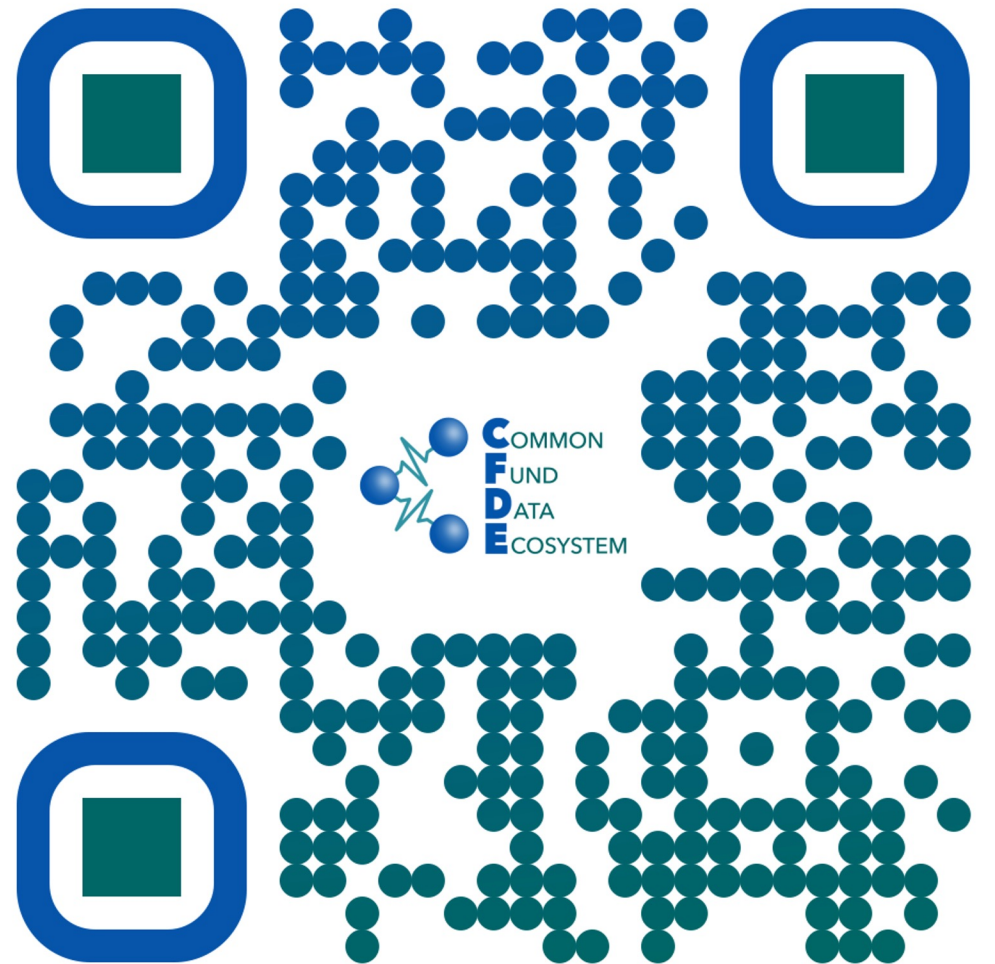
Name	Created on	Extension	Size
DRS HLKHCXX-7.hgv.bam <small>WGS KIDS FIRST SD_46SK55A3</small>	Feb. 24, 2023 08:06	BAM	62.4 GiB
DRS CDH4-80f.cram <small>WGS KIDS FIRST SD_46SK55A3</small>	Feb. 24, 2023 08:06	CRAM	26.5 GiB
DRS SRS1346527_qc.fastq.bz2 <small>METAPHLAN HMP METAGENOME</small>	Feb. 24, 2023 08:04	FASTQ.BZ2	144.2 MiB
DRS SRS1346527_metaphlan_summary_stats.json <small>METAPHLAN HMP METAGENOME</small>	Feb. 24, 2023 08:04	JSON	0.4 KIB
DRS GTEX-N7MS-2526-SM-2D7W3.Aligned.sortedByCoord.out.patched.md.bam <small>GTEX ANVIL RNASEQ</small>	Feb. 24, 2023 08:02	BAM	4.7 GiB
DRS GTEX-N7MT-1226-SM-2TC6K.Aligned.sortedByCoord.out.patched.bam <small>GTEX ANVIL RNASEQ</small>	Feb. 24, 2023 08:02	BAM	7.3 GiB

- The DRS client on CAVATICA resolves authorizations
- Researchers can focus on the science and use them as input to a task

Office Hours



Collaborative Project Interest Form



Q&A and ... Thank You!



Sangeeta Shukla,
Kids First / Children's Hospital of Philadelphia



Eric Tobin,
Velsera / Seven Bridges CAVATICA

Appendix: Other References

Who are the CAVATICA Users?

The platform is designed to serve a wide range of scientists and users with varying skill sets



ADMINISTRATORS

- Manage and Control Users
- Monitor and Control Institutional Assets
- Manage and Monitor Projects
- Monitor and Control Costs
- Create Reports



BIOINFORMATICIANS

- Store, Manage, and Share Data
- Access Public and Proprietary Datasets
- Query, Build, and Investigate Cohorts of Interest
- Access Optimized Tools and Workflows
- Create, Optimize, Maintain, and Distribute New Tools and Workflows
- Create Push-button Automation Solutions
- Analyze Data at Scale with Tools and Workflows
- Conduct Interactive Exploratory Analyses
- Explore/Visualize Results and Gather Insights
- Easily Collaborate with Other Stakeholders
- Integrate with External Systems



BENCH SCIENTISTS

- Store, Manage, and Share Data
- Run Optimized Tools/Workflows at Scale
- Conduct Defined Analyses via Push-button Solutions
- Investigate/Visualize Results
- Easily Collaborate with Other Stakeholders



CLINICIANS

- Conduct Validated Analyses via Push-button Solutions
- Query, Build, and Investigate Cohorts of Interest
- Create Reports
- Investigate/Visualize Results
- Easily Collaborate with Other Stakeholders



DEVELOPERS

- Create, Optimize, and Maintain New Tools and Workflows
- Create Push-button Automation Solutions
- Create Custom Interfaces for Specific Use Cases
- Distribute Proprietary Tools/Workflows
- Integrate with Upstream/Downstream Systems