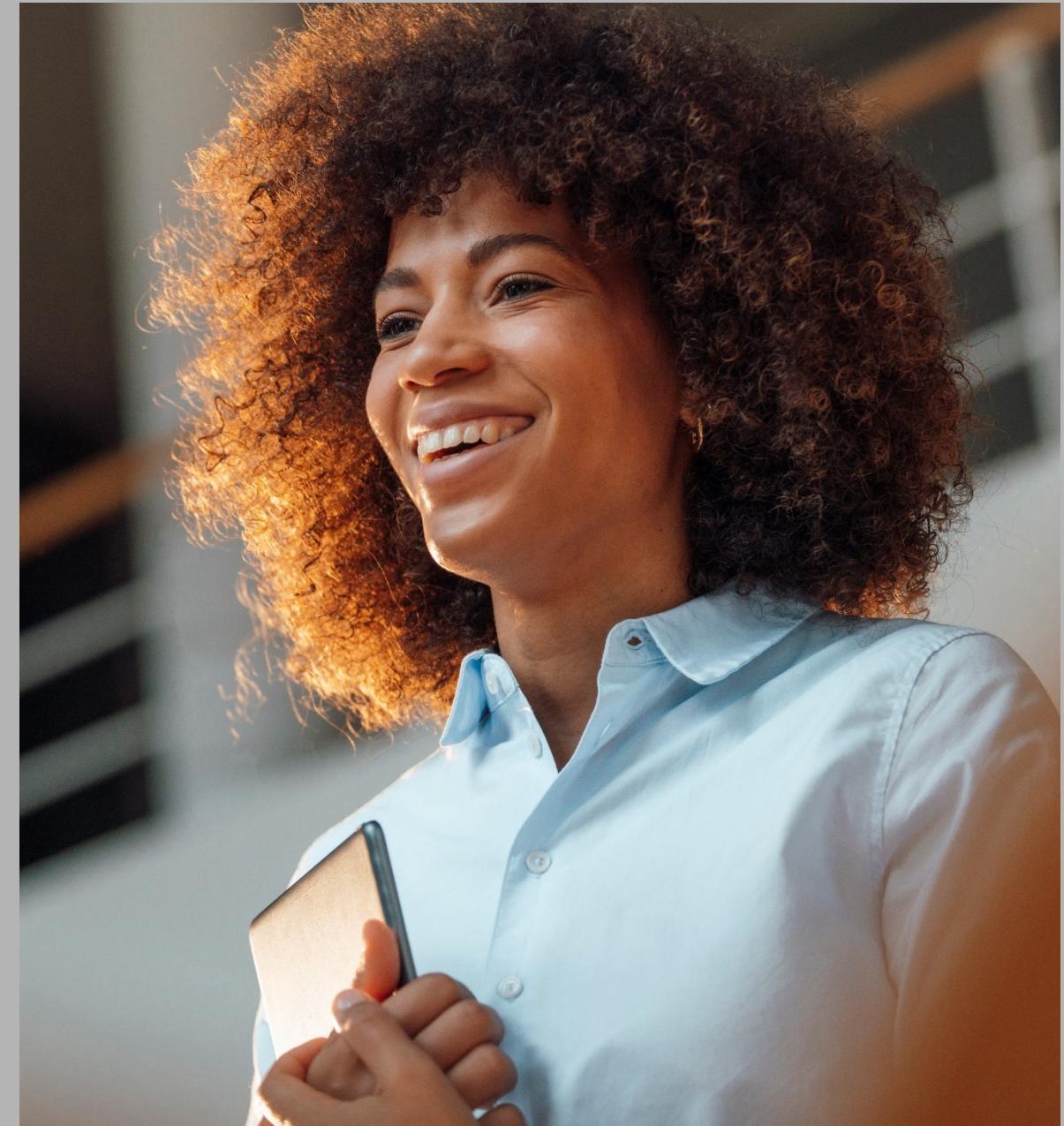


Introduction to Cloud Computing.

CAVATICA and Velsera





surs: Tuesdays 10 am ET, Thurs 2 pm ET. Click to join! ----- Office H

But First: Log On To CAVATICA

<https://www.cavatica.org/>

The screenshot shows the Cavatica homepage. At the top right is a purple and blue geometric logo. Below it, the word "CAVATICA" is written in large, bold, purple and blue letters. Underneath the title is the tagline "DEMOCRATIZING ACCESS TO GENOMICS DATA". A small paragraph explains what Cavatica is: "CAVATICA is a storage, sharing, and analysis platform designed to handle large volumes of pediatric tumor genomics data. It is produced in collaboration with [Seven Bridges](#) and based on the Seven Bridges Platform for cloud storage and bioinformatics analysis." To the right of this text is a photograph of three people sitting around a table, looking at laptops. To the right of the photo is a call-to-action box with the text "Stay Up To Date with the Latest News" and a link to the news page. Below the photo is another section titled "Access Data In Seconds" with a description of how users can integrate with various platforms like Kids First Portal and INCLUDE Data Hub. To the right of this text is a photograph of a laptop screen displaying a complex dashboard with various charts and data points.

CAVATICA

DEMOCRATIZING ACCESS
TO GENOMICS DATA

CAVATICA is a storage, sharing, and analysis platform designed to handle large volumes of pediatric tumor genomics data. It is produced in collaboration with [Seven Bridges](#) and based on the Seven Bridges Platform for cloud storage and bioinformatics analysis.

Stay Up To Date with the Latest News

We're constantly adding new data and tools. Learn more by visiting our [news page](#).

Access Data In Seconds

CAVATICA seamlessly integrates with the [Kids First Portal](#) and the [INCLUDE Data Hub](#), generating private workspaces with genomic data from selected cohorts in seconds. Users can also quickly identify and load genomic data from [Kids First](#), [INCLUDE DCC](#), [CBTRN](#), [TARGET](#), or [TUGA](#) into their workspace using the built-in [CAVATICA Data Browser](#). Finally, researchers can also bring their own files into CAVATICA for comparisons to petabytes of existing genomic data, increasing their sample size and maximizing their research impact.

Join Our Office Hours

Questions? Need help?

We hold sessions twice a week:

Tuesdays at 10 AM ET and

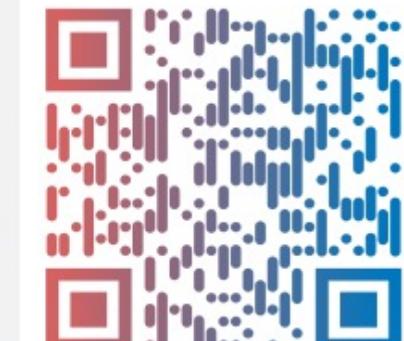
Thursdays at 2 PM ET

Come chat with us about your research!

Scan the QR Code or
Follow The Link



<https://meet.google.com/kbs-ojnjd-cg>



Introduction to Cloud Computing with **CAVATICA.**

Adding Concepts and Competencies to your Toolkit

What is Cloud Computing?



- **The cloud** - (kind of) just means “the internet”
- **Cloud computing** - connecting to computers (located elsewhere) over the internet to use their processing power and storage.
- **Cloud providers** - the people who own those computers. Amazon, Google, and Microsoft, plus others.

Explosion of 'omics data with ease of sequencing

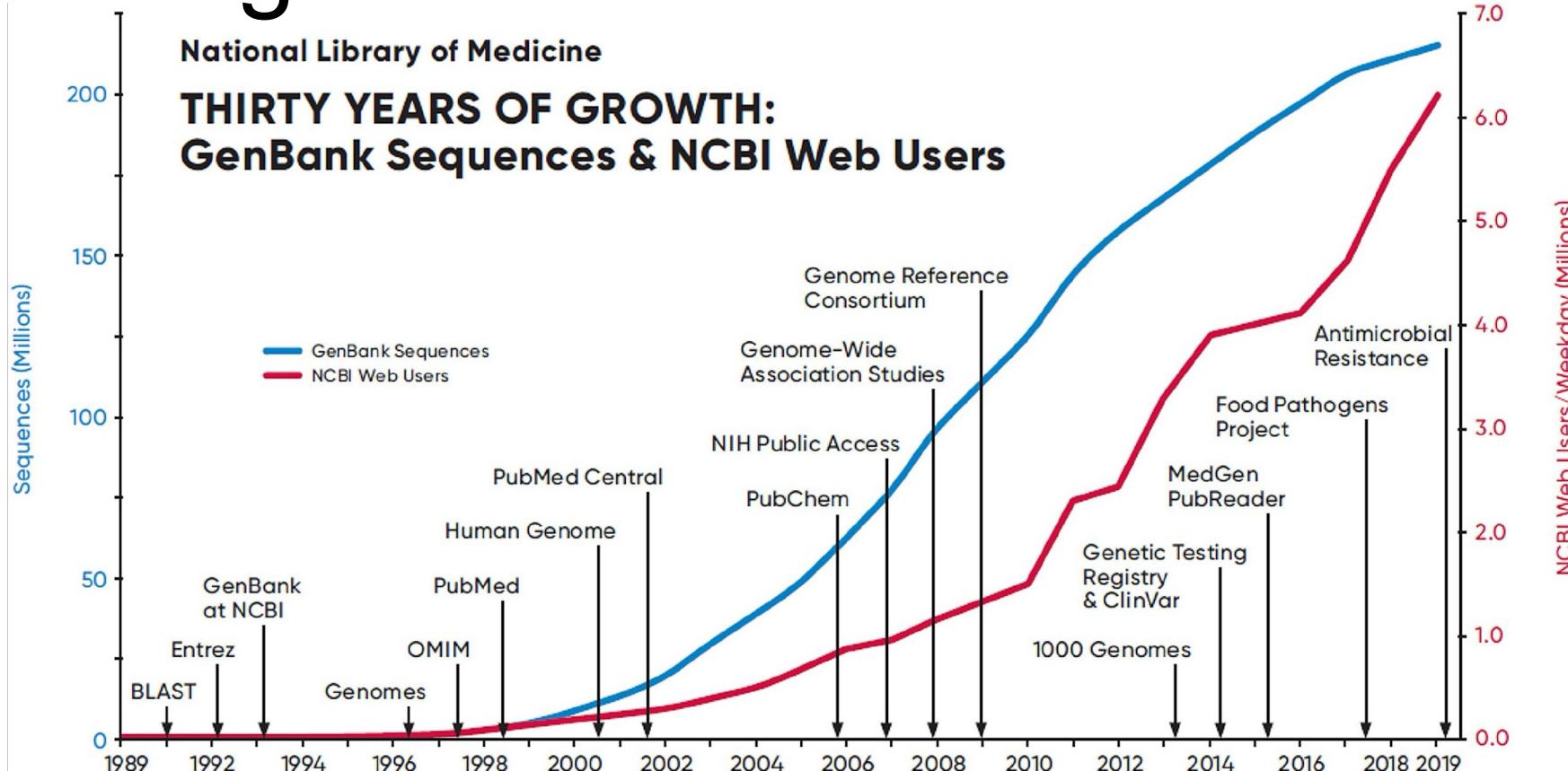
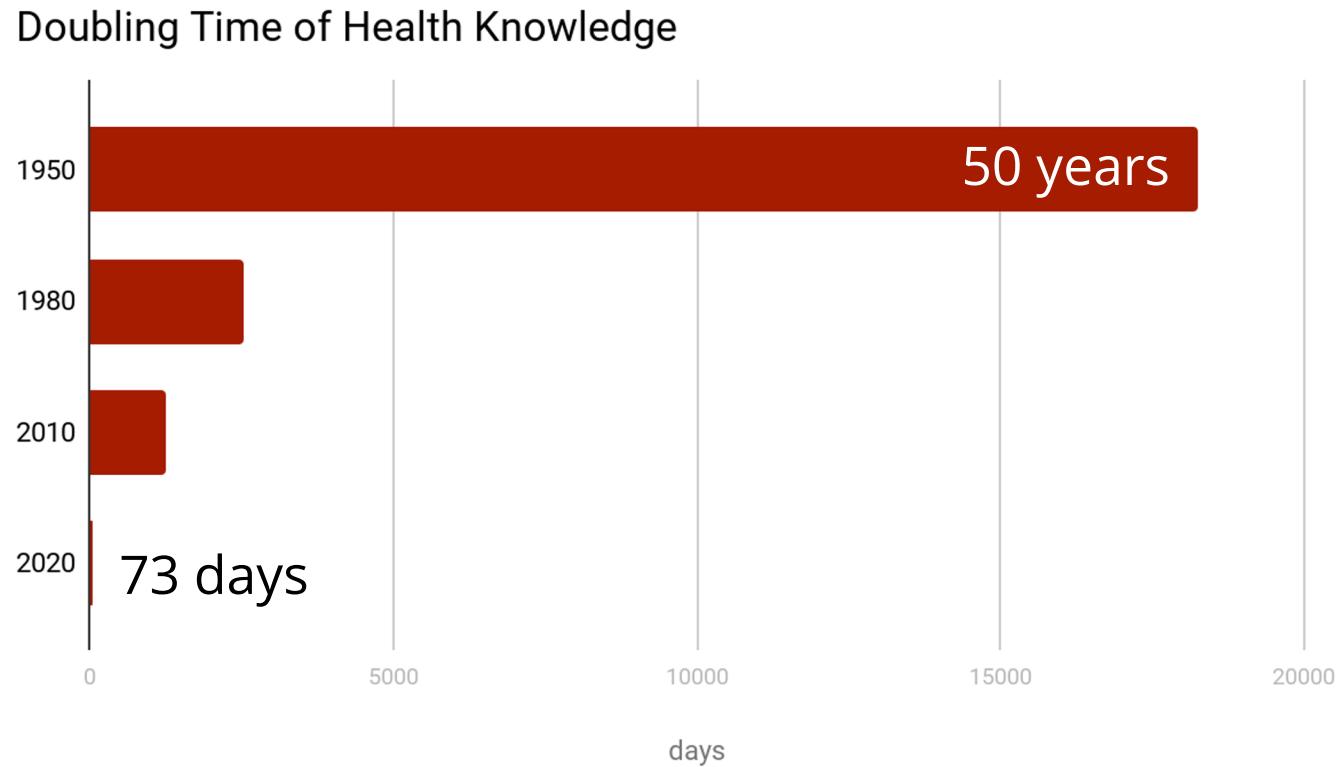


Fig. 1. Growth of GenBank sequences and NCBI web users through 2019. Figure from the Department of Health and Human Services National Institutes of Health. Jim Gaffney, et. al. Open access to genetic sequence data maximizes value to scientists, farmers, and society, Global Food Security, Volume 26, 2020.

The rate of data generation is accelerating rapidly



- More biomedical data will be generated this year than all previous years **combined**
- More different **kinds** of data – gene and genome sequencing, imaging, sensors (i.e. fitbit), cellular measurement data, and more

Why do Cloud Computing?



Why *should* I do Cloud Computing?



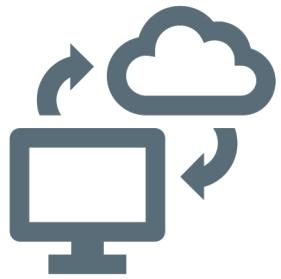
“Cloud is about how you do computing, not where you do computing.”



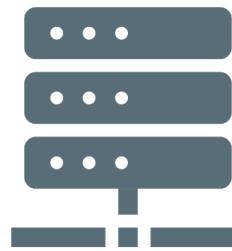
Important Concepts in Cloud Research.

- Cloud provider costs
- FAIR Data Principles
- Scalability, portability, and reproducibility
- Security and Compliance

Cloud Provider Costs...



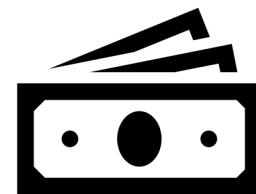
Compute Costs



File Storage Costs



Data Transfer Costs



... and Cloud Credit Support

- For Down Syndrome research
- \$100 INCLUDE Pilot Credits
- Up to \$1,000 Limited Credits
- Up to \$20,000 Extended Credits



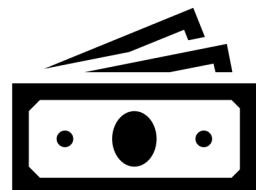
For every user - Email Support@Velsera.com

One page description of intended use

Budget and description

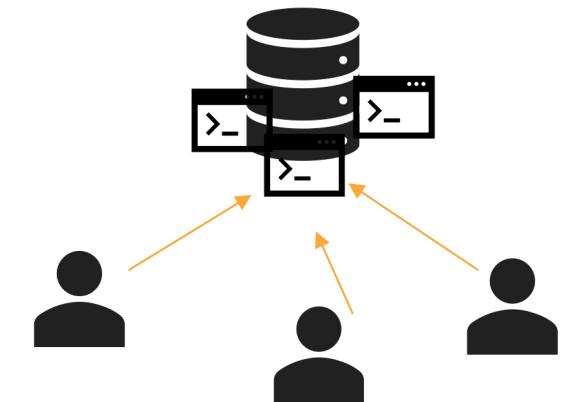
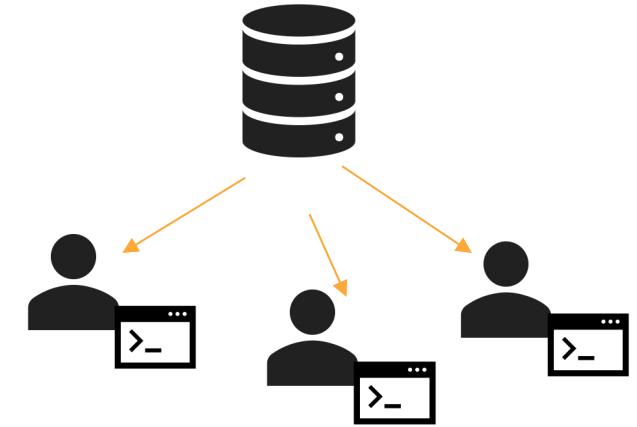


Other sources of support may be available depending on the research proposed – see <https://docs.cavatica.org/docs/cloud-credits-on-cavatica> for details

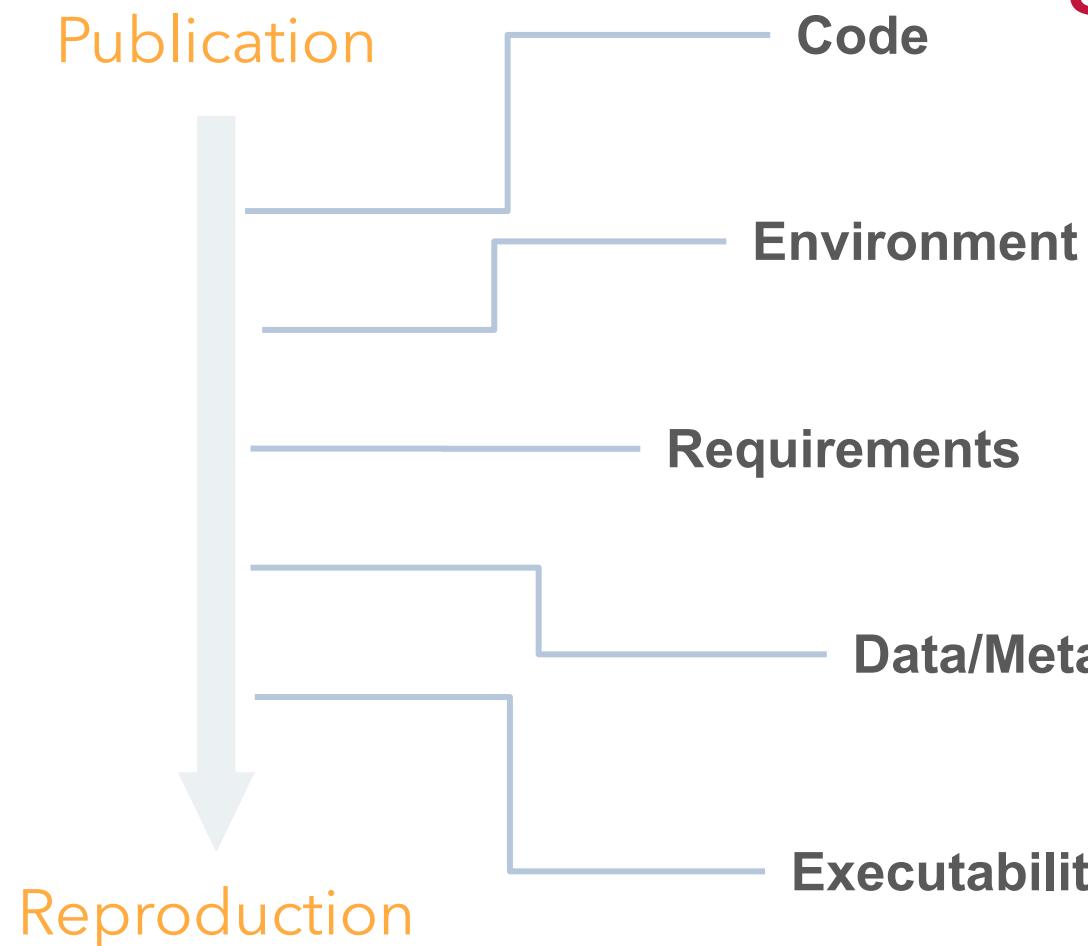


Using the Cloud to Store and Analyze Growing Health Data

- Immediate scaling. (“Rent” vs “Buy”)
- Levels the playing field
 - Researchers at institutions that lack high performance computation systems can access data and resources
- Many researchers can access the data without needing to make duplicate copies
- Data and analysis methods kept in the same repository



Portability via containerization and workflow language



What is FAIR?

Findable, Accessible, Interoperable, Reusable

“The FAIR Guiding Principles for scientific data management and stewardship”



- Wilkinson et al. (2016) *Scientific Data* 3, 160018. doi: 10.1038/sdata.2016.18
- **Findable:** Registered in a searchable resource; the data is described with rich **metadata**; both humans and machines can *find the data they want if it exists*.
- **Accessible:** Retrievable using standard protocols; includes **authorization** (is the requester allowed to have this data?) and **authentication** (is this the data that the requester means to get?)
- **Interoperable:** Uses standard, agreed-upon terminology; humans and machines can reliably understand what to do with the data.
- **Reusable:** Give rich and thorough detail about the data so that a researcher can reuse it in a new analysis; it is very clear where the data come from and how it was created (**provenance**).

Velsera Security & Compliance



NIH Genomic Data Sharing Guidelines 2025

Starting **January 26, 2025**, two major changes will affect how authorized users can download and analyze this data:

- Repositories storing genomic data must meet stricter security guidelines, like NIST SP 800-53 Moderate or equivalent standards (e.g., FedRAMP or FISMA Moderate).
- Approved users must confirm that their IT systems or third-party cloud services comply with NIST SP 800-171, covering all computers used to work with this data.
- This applies to all computers that are used to **download and analyze controlled-access data**.

CAVATICA already meets and exceeds these guidelines

<https://velsera.com/nih-genomic-data-sharing-policy-updates/>

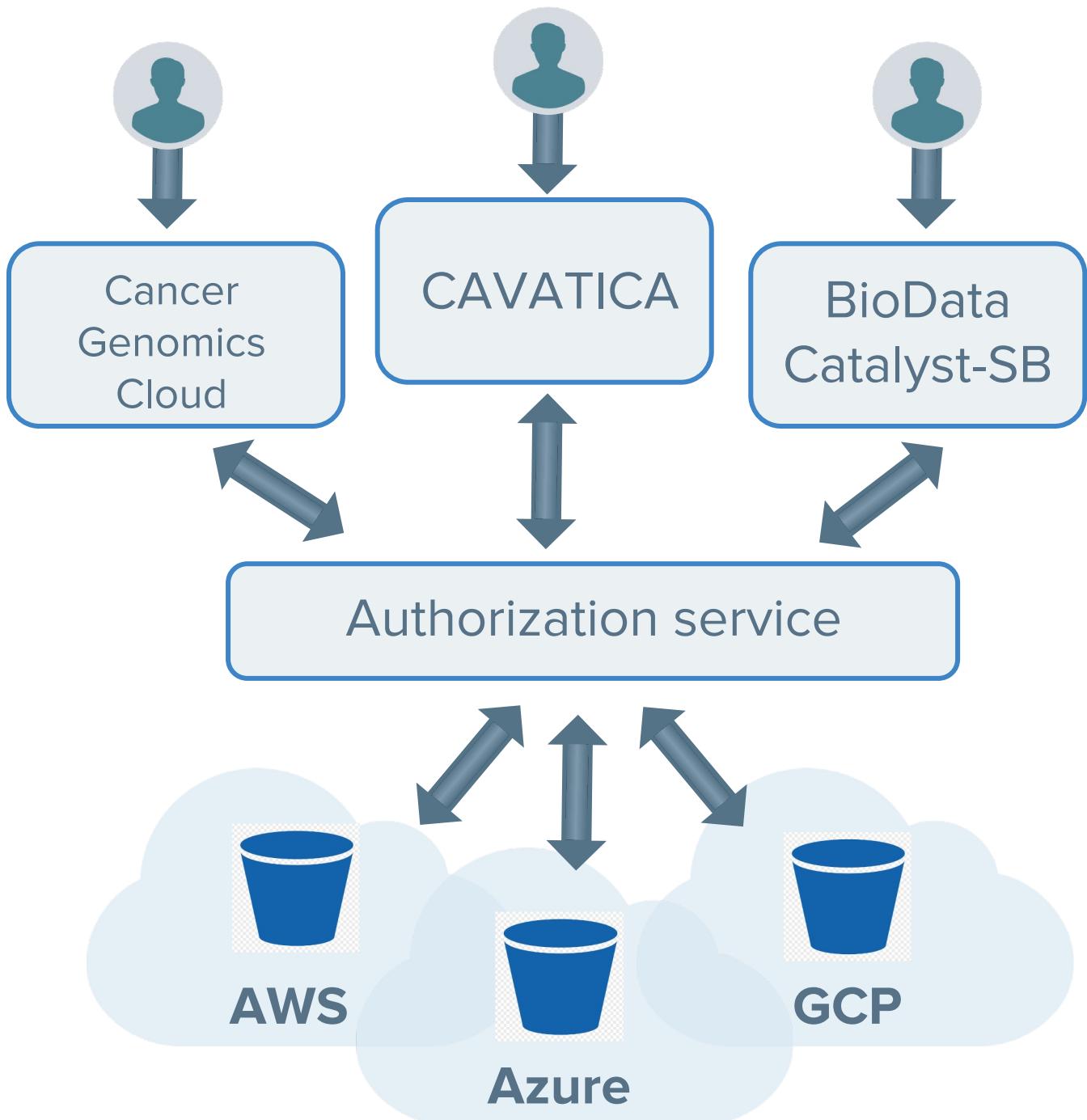


Access hosted data

Stored in cloud buckets

Available on all Velsera academic platforms

Access controlled programmatically by authorization service. User permissions are read from dbGaP using iTrust and eRA Commons login



Example: Data portability and data location depends on cost, policy, & other factors



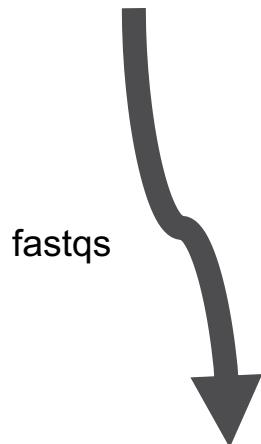
RNA-Seq & WGS (n ~ 9k) samples uniformly processed on CAVATICA with the Kids First RNA-Seq pipelines



RNA-Seq & **WGS** (n = 112) samples available in National Compute Infrastructure Australia (traditional HPC)

Moving or copying data can be expensive

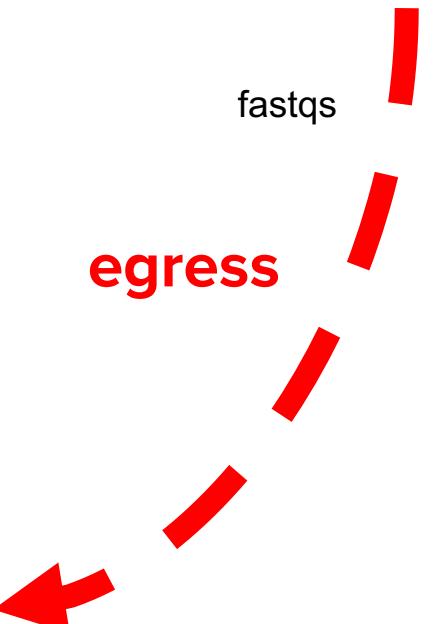
**ZERO RNA-Seq
dataset**



**CBTN RNA-Seq
dataset**

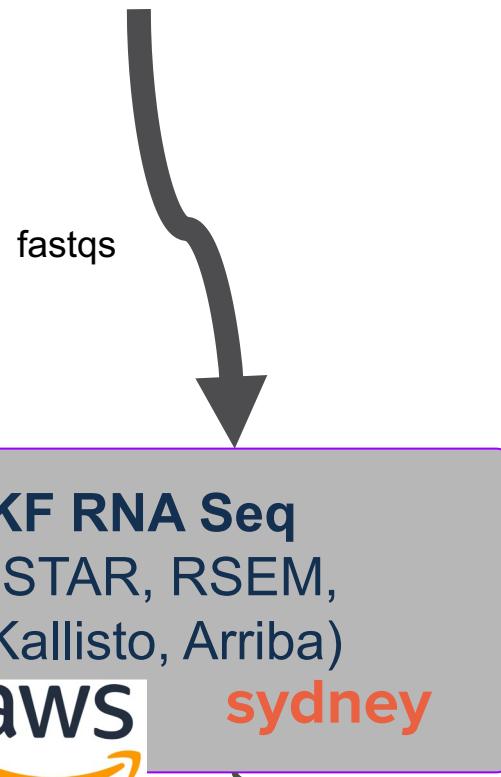
fastqs

egress

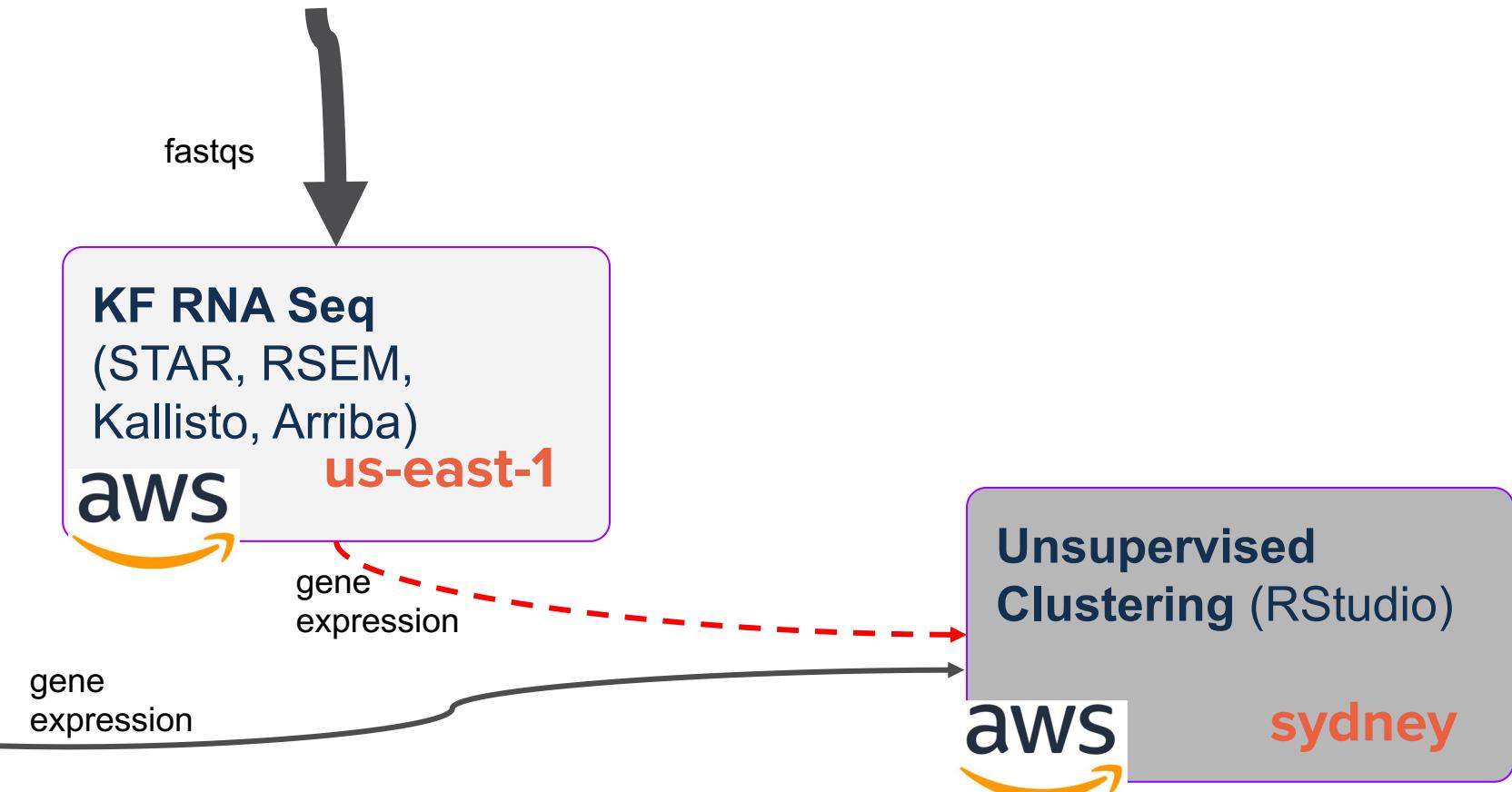


Design your analysis to leverage the cloud efficiently

ZERO RNA-Seq
dataset



CBTN RNA-Seq
dataset



Institutional level: Design Data for Cloud

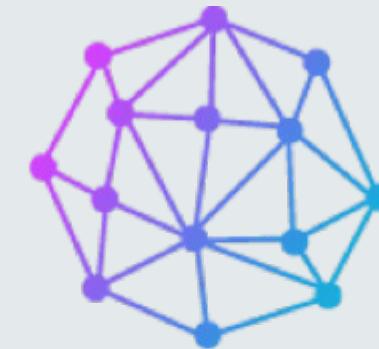
The INCLUDE Data is Cloud Native



- INCLUDE data processing is done on CAVATICA
- FAIR data by design:
 - Stored on a Data Repository Service (DRS) server
 - Researcher Auth Service (RAS) built in
 - Easy to access (as you will see!)

Introduction to CAVATICA

A cloud computing platform for biomedical researchers



CAVATICA

Who are the CAVATICA Users?

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



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



ADMINISTRATORS

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



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

What is it and what can it do?



User friendly portal

Access petabytes of publicly available data, and analyze it alongside private data.



Unprecedented Collaboration Features

Collaborate within and across organizations while keeping control of your assets with precise permission levels.



Industry Standard Tools: Reproducibility

Execute, build, and customize analysis pipelines using popular tools such as CWL, WDL, Nextflow, Docker, RESTful APIs, and CLI.



Connected Cloud

Keep data in your own buckets with storage support for AWS, and Google Cloud Platform.

Connect to the data & use it wisely



Federated Public Data Sets

We partner with a wide range of public organizations to facilitate access to large federated public data sets.



National Institutes
of Health

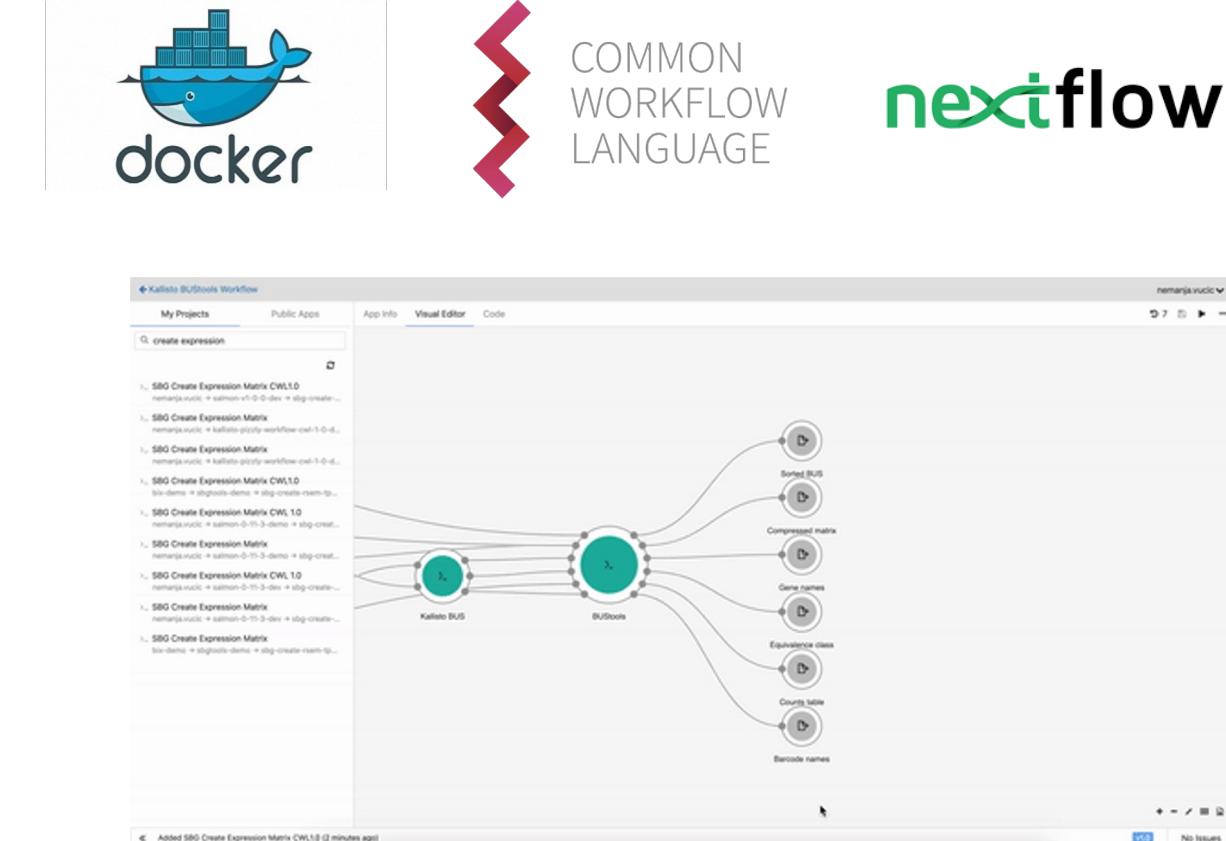
Trusted Partner

Only commercial "NIH Trusted Partner," and NIH Researcher Authentication Service partner; allowing our users with access rights to leverage large NIH datasets.



Use or Create Reproducible Workflows

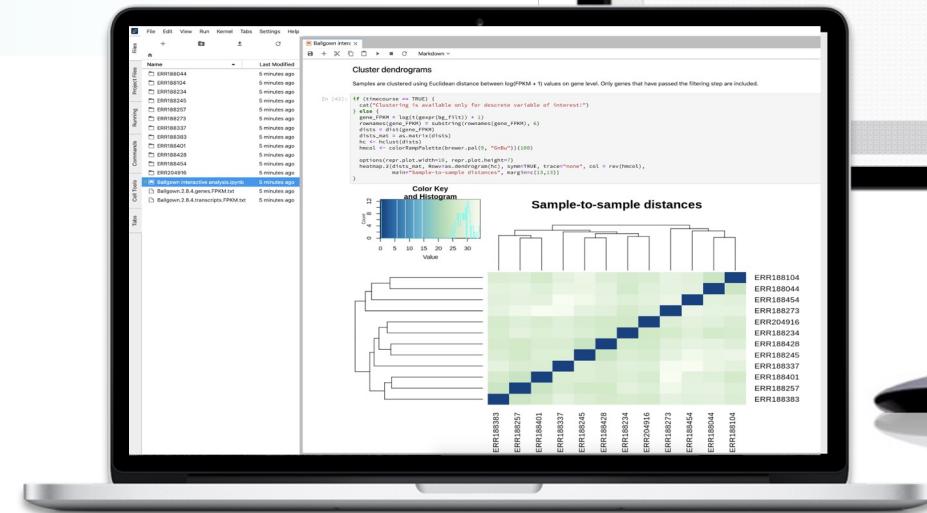
- ◆ Common Workflow Language enables **portability; reproducibility; & scalability**.
- ◆ Use or combine **800+ optimized tools** and workflows to construct your analysis.
- ◆ Seamlessly **import workflows** from external public repos (e.g. Dockstore).
- ◆ **Create your own apps** with our diagrammatic Web Composer.



Integrated Interactive Analysis Tools

Data Science Workbench

Derive new insights using interactive analysis environments with JupyterLab, and RStudio environments. Code in Python and R, and create Jupyter Notebooks to record and share your analyses.

A screenshot of a JupyterLab interface. It features a sidebar with tabs for 'File', 'Notebook', 'Files', 'Running', 'Commands', and 'Project Folders'. A central area displays two open notebooks: 'Landing' and 'R_demo.ipynb'. The 'R_demo.ipynb' notebook contains several code cells with R code for data manipulation and visualization, such as reading tables and creating plots like boxplots and heatmaps.

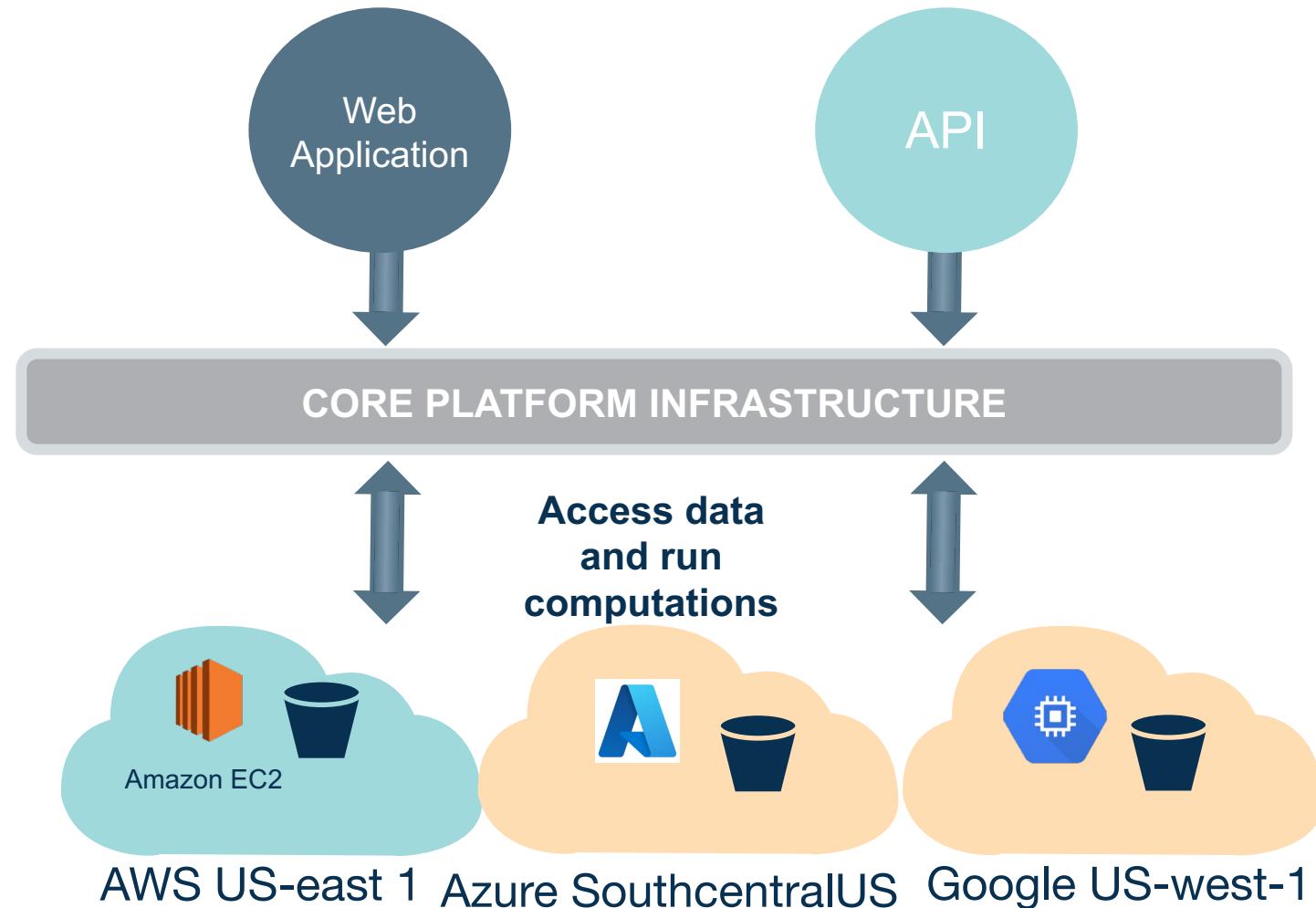
Compute on AWS, Google Cloud, or Azure, based on where data lives

Access and compute on data stored on distributed cloud locations from one user interface or API

Avoid egress costs

Current cloud location options:

- AWS US N. Virginia (us-east-1)
- Google US Oregon (us-west-1)
- Azure *coming soon* (southcentralus)



Data Type-Agnostic Flexible Analysis Ecosystem

Hundreds of tools for different research questions

Easy to share with collaborators

Reproducible analysis

Flexible to match your analysis needs

Add your own tools

Contact us at any time with questions!



DNA-seq



DNA Methylation



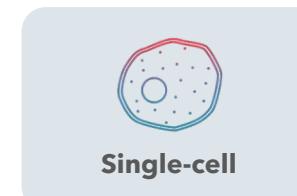
ATAC-seq



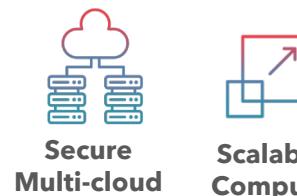
ChIP-seq



RNA-seq



Single-cell



Secure
Multi-cloud



Scalable
Compute



Tools and
Workflows



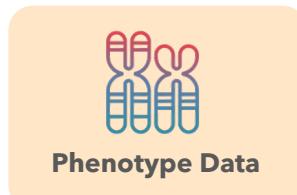
Analysis &
Visualization



Imaging



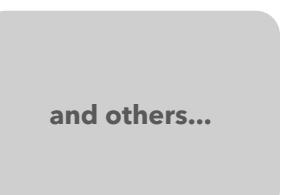
GWAS/pheWAS



Phenotype Data



EHR/RWE



and others...

Success stories using CAVATICA



Neoplasia
Volume 35, January 2023, 100846



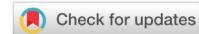
The children's brain tumor network (CBTN) - Accelerating research in pediatric central nervous system tumors through collaboration and open science

Jena V. Lilly^a ¹, Jo Lynne Rokita^a ¹, Jennifer L. Mason^a, Tatiana Patton^a, Stephanie Stefanekewitz^a, David Higgins^a, Gerri Trooskin^a, Carina A. Larouci^a, Kamnaa Arya^a, Elizabeth Appert^a, Allison P. Heath^a, Yuankun Zhu^a, Miguel A. Brown^a, Bo Zhang^a, Bailey K. Farrow^a, Shannon Robins^a, Alison M. Morgan^a, Thinh Q. Nguyen^a, Elizabeth Frenkel^a, Kaitlin Lehmann^a ...Angela J. Waanders^P

POSTER PRESENTATIONS - PROFFERED ABSTRACTS | APRIL 04 2023

Abstract 6576: Gabriella Miller Kids First Data Resource Center (KFDRC): Empowering discovery across germline and somatic variation in pediatric cancer

David Higgins; Jean-Philippe Thibert; Michele Mattioni; Jack DiGiovanna; Robert L. Grossman; Bailey K. Farrow; Eric Wenger; Samuel Volchenboum; Robert J. Carroll; Melissa A. Haendel; Deanne M. Taylor; Yuankun Zhu; Vincent Ferretti; Adam C. Resnick; Alison P. Heath



+ Author & Article Information

Cancer Res (2023) 83 (7_Supplement): 6576.

<https://doi.org/10.1158/1538-7445.AM2023-6576>

Open access

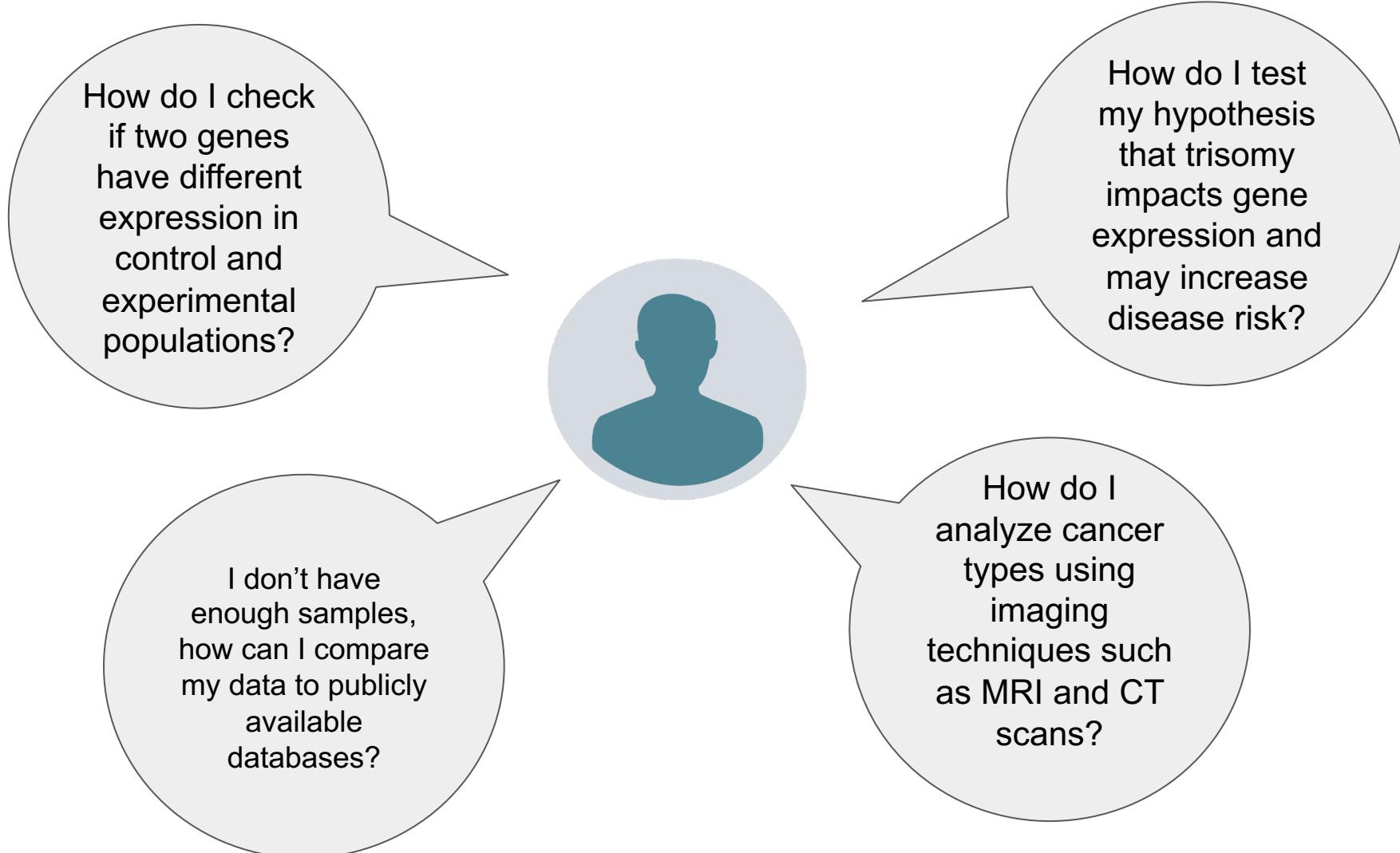
Protocol

BMJ Open Protocol for a comprehensive prospective cohort study of trio-based whole-genome sequencing for underlying cancer predisposition in paediatric and adolescent patients newly diagnosed with cancer: the PREDICT study

Noemi Auxiliadora Fuentes Bolanos ^{1,2}, Bhavna Padhye ^{3,4}, Macabe Daley ², Jacqueline Hunter ^{5,6}, Kate Hetherington ^{6,7}, Meera Warby ^{2,8}, Eliza Courtney ^{1,2}, Judy Kirk ^{9,10}, Sarah Josephi-Taylor ^{11,12}, Yuyan Chen ⁴, Frank Alvaro ¹³, Kristine Barlow-Stewart ^{2,5}, Marie Wong-Erasmus ^{2,6}, Paulette Barahona ^{2,6}, Pamela Ajuyah ^{2,6}, Ann-Kristin Altekoester ^{2,6}, Vanessa J Tyrrell ², Loretta M S Lau ^{1,2}, Claire Wakefield ^{5,6}, Dianne Sylvester ⁴, Katherine Tucker ^{1,8}, Mark Pinese ^{2,6}, Luciano Dalla Pozza ³, Tracey A O'Brien ¹

To cite: Fuentes Bolanos NA, Padhye B, Daley M, et al. Protocol for a comprehensive

What's the research question you want to answer?



Prototypical User Flow

Create a Project

Select/access data

Select/create tools

Create and run analysis

Organizational unit within platform

Many ways to find and bring in data:

- Data Browser
- Desktop uploader
- Command line uploader
- Volumes

Tools, workflows, and software packages

- Public Apps
- Gallery
- Tools or workflows wrapped in CWL
- R packages
- Python libraries

Specify how an analysis will be run

- Task page
- Notebooks in RStudio or JupyterLab

Access and search petabytes public datasets on CAVATICA

The screenshot shows the CAVATICA search interface. At the top, there is a navigation bar with links for Projects, Data, Public Apps, Public Projects, Developer, Controlled projects, and a notification bell. Below the navigation bar, a search bar contains the placeholder "Search by ID". A button labeled "New query" is visible. The main content area is titled "Select dataset(s)" and features a list of datasets under the "Kids First" project. Each dataset entry includes a checkbox, the dataset name, a brief description, and a "Details" button. The datasets listed are:

- Kids First:
 - OpenDIPG: ICR London RE_M7P6CTHV · DM 2021.6.0 [Details](#)
 - Kids First: Intersections of Cancer & SBD RE_QNTQV8J6 · DM 2021.6.0 [Details](#)
 - Kid First: Hemangiomas (PHACE) RE_KQ4KFPWA · DM 2021.6.0 [Details](#)
 - Kids First: Microtia - Hispanic RE_EJ814TNN · DM 2021.6.0 [Details](#)
 - Kids First: Disorders of Sex Development RE_3TVKNGXZ · DM 2021.6.0 [Details](#)

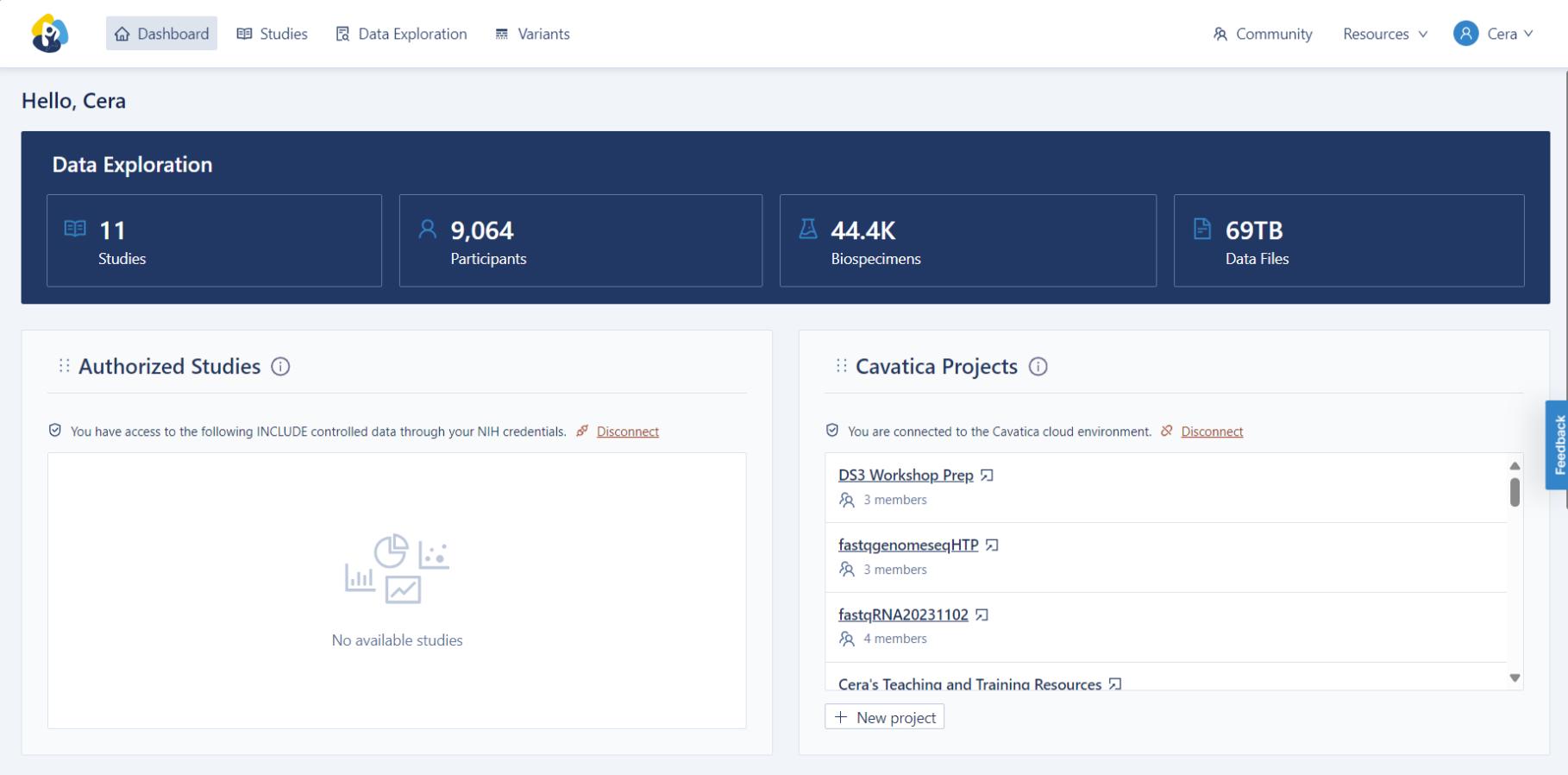
At the bottom of the page, there are links for Terms, Privacy, and Copyright, along with a copyright notice for Seven Bridges Genomics.

The screenshot shows the CAVATICA dataset details page for "PBTA-CBTN". The top navigation bar is identical to the one in the search interface. The main content area is titled "Datasets" and shows a list of datasets under the "PBTA-CBTN" project. The datasets listed are:

- PBTA-PNOC
- PBTA-CBTN
- MARIS_NB_XE_01
- MARIS_NB_CL_01
- SU2C_MB_PA_01
- MCGILL_DIPG_PA_01
- Chordoma Foundation Dataset

To the right of the dataset list, there is a detailed description of the "PBTA-CBTN" project. The title is "Children's Brain Tumor Network (CBTN)". The description states: "The Children's Brain Tumor Network is dedicated to driving innovative discovery, pioneering new treatments and accelerating open science to improve health for all children and young adults diagnosed with a brain tumor. By accelerating the pace of translational research and the discovery of new treatments, we are a global community with the shared goal to save children and young adults from brain tumors." A link to the "2019-2020 Annual Report" is provided. Below the description, there is a section titled "Accelerating Global Childhood Brain Tumor Research". A note at the bottom states: "Brain tumors are complex and difficult conditions to treat in children. This year around the world, more than 67,000 children and young adults will be diagnosed with a primary brain or central nervous system (CNS) tumor."

Connected data portals provide an easy way to find data



The screenshot shows the INCLUDE Data Portal dashboard for a user named Cera. At the top, there are navigation links for Dashboard, Studies, Data Exploration, and Variants. On the right, there are links for Community, Resources, and Cera. The main area starts with a greeting "Hello, Cera". Below it is a dark blue header with the title "Data Exploration" and four summary boxes: "11 Studies", "9,064 Participants", "44.4K Biospecimens", and "69TB Data Files". The left side features a section titled "Authorized Studies" with a note about NIH credentials and a message stating "No available studies". The right side features a section titled "Cavatica Projects" listing several projects: "DS3 Workshop Prep" (3 members), "fastqgenomeseqHTP" (3 members), "fastqRNA20231102" (4 members), and "Cera's Teaching and Training Resources". A "New project" button is at the bottom of this list. A vertical "Feedback" button is located on the far right.

<https://portal.includedcc.org/>

Connect directly to 30+ petabytes of data in the SRA

CAVATICA Projects ▾ Data ▾ Public Apps ▾ Public Projects Developer ▾ Controlled projects Staff ▾

cfisher92

Public apps

🔗 SRA to DRS converter

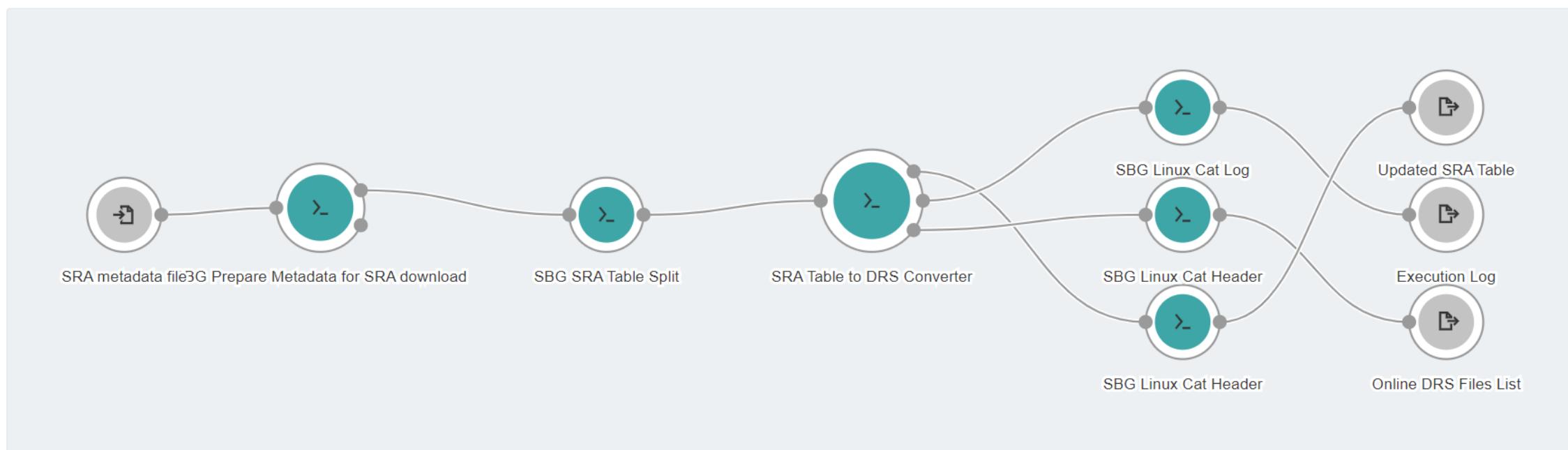
Created by admin on Mar. 1, 2024 05:56 • Last edited by admin on Apr. 12, 2024 08:26

Revision note: "App info rev 70"

Revision 22 ▾

▶ Run

...



Conveniently bring in your own data



Drag & drop files from your computer or

[Browse files](#)

This upload method is primarily intended for small-scale uploads.

To upload a [larger volume of files](#), please use our [Data Tools](#). Learn more about [uploading from your computer](#).

New folder

+ Add files ▾

...

Datasets

Public Files

Projects

Your Computer

FTP / HTTP

GA4GH Data

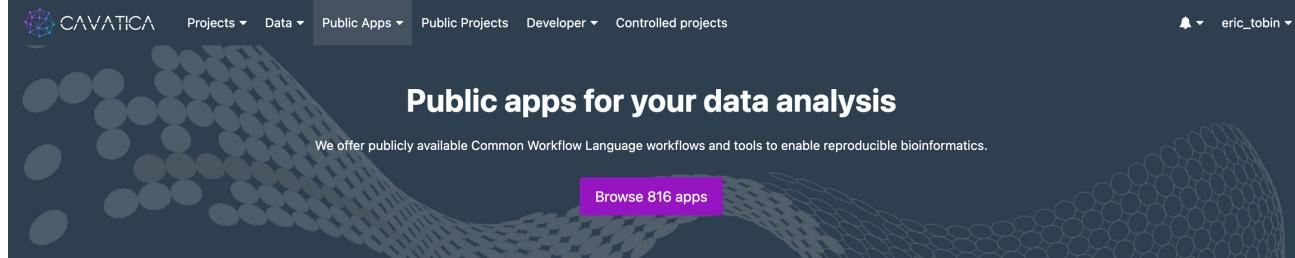
Repository Service (DRS)

Data Tools

Volumes

Import from a manifest file

Public Applications Gallery – tools and workflows for analysis



Platform Tool/Workflow Repository

800+

Curated tools & workflows



Secure



High quality apps &
documentation



Optimized to
run on cloud



Updated
regularly



Customized user
tools/workflows

The Common Workflow Language (CWL) logo, featuring a red zigzag icon followed by the text 'COMMON WORKFLOW LANGUAGE'.

Founded in 2014 by Seven Bridges and partners

An open standard adopted by 40+ organizations worldwide including Institut Pasteur, Wellcome Sanger Institute, CERN, UCSC, Harvard T.H. Chan School of Public Health, and many more

An ecosystem of tools and workflows

Reproducible and portable

Complete tooling and low-entry barrier for new users



Dockstore

Create, Share, Use



CWL
project

Why (*should I*) do Cloud Computing for Down Syndrome Research?

- Because, to do great science and help people, we have to be able to work together.

“Be positive – be curious. There are a lot of good people in the world doing amazing things. We have so much more useful information than we’ve ever had, and we can harness it to help the world. It’s only going to get better with all the amazing tools we have at our disposal. There is a lot of positivity in the world, and we should contribute as best we can. “

– Vaidhy Mahaganapathy



Dr. Vaidhyanathan Mahaganapathy

Thank you!