SSO, AuthN/Z, and Cloud Storage for Common Fund DCCs

SSO, AuthN/Z, and Cloud Storage for Common Fund DCCs

Authors and Contributors

Author - The primary report author is Brian O'Connor
Research and contributions - The following individuals provided research and contributions to this report: Teresa Barsanti and Samir Faci
Comments and feedback - The following individuals provided helpful comments and feedback on this report: Kristin Ardlie, Amanda Charbonneau, Ian Foster, Allison Heath, Ronald Liming, Jared Nedzel, and Rick Wagner
Reference Materials - The following individuals provided valuable reference materials that were useful in preparing this report: Zachary Flamig, Robert Grossman, Kurt Rodarmer, Steve Sherry, and Eugene Yaschenko

Versions

Version	Date	Description
V1	July 31st, 2019	Initial report version with profiles of GMKF and GTEx DCCs.
V2 (upcoming)	TBD August, 2019	The next, upcoming release of this report that will include information on the MoTrPAC DCC and profiles of additional infrastructure including Globus Auth.
V3 (upcoming)	Sept 19th, 2019	An upcoming, final release of this report incorporating feedback, additional DCCs, and possibly additional infrastructure profiles.

Document Link

The latest version of this document can be found online at http://bit.ly/2Yuyyge. Anyone can leave comments or ask questions directly on the document.

Introduction

This report examines the needs of Common Fund Data Coordination Centers (DCCs) with regards to Single Sign On (SSO), authentication, authorization, and cloud storage systems. As part of this report, we profile multiple existing Commons Fund DCCs, identify their current functionality, ask what works well, what does not work well, and examine possible improvements to SSO, authentication, authorization, and cloud storage utilization. We then examine various existing solutions and how they relate to the work of the DCCs. Finally we explore emerging themes from multiple DCC interviews and propose efforts in the short term (September 2019 - October 2020) that will improve existing DCCs' use of SSO, authentication, authorization, and cloud storage systems as well as provide a template and starting point for future DCCs.

This report is a living document and subsequent releases will incorporate information from future DCC interviews, additional solution profiles, and refinements to the short term and longer term proposals.

Research on the Cloud

Over the last ten years dramatic technological advances have enabled the creation of enormous biomedical datasets. With the advent of next generation sequencing, the cost of producing genomic datasets has plummeted producing a deluge of data from both individual labs and large-scale consortium efforts (Figure 1). Examples of the latter include the TCGA project with ~13K exomes and ~4K whole genomes, the TOPMed project which has over 100K whole genomes, and GTEx which has RNAseq data on ~12K samples.

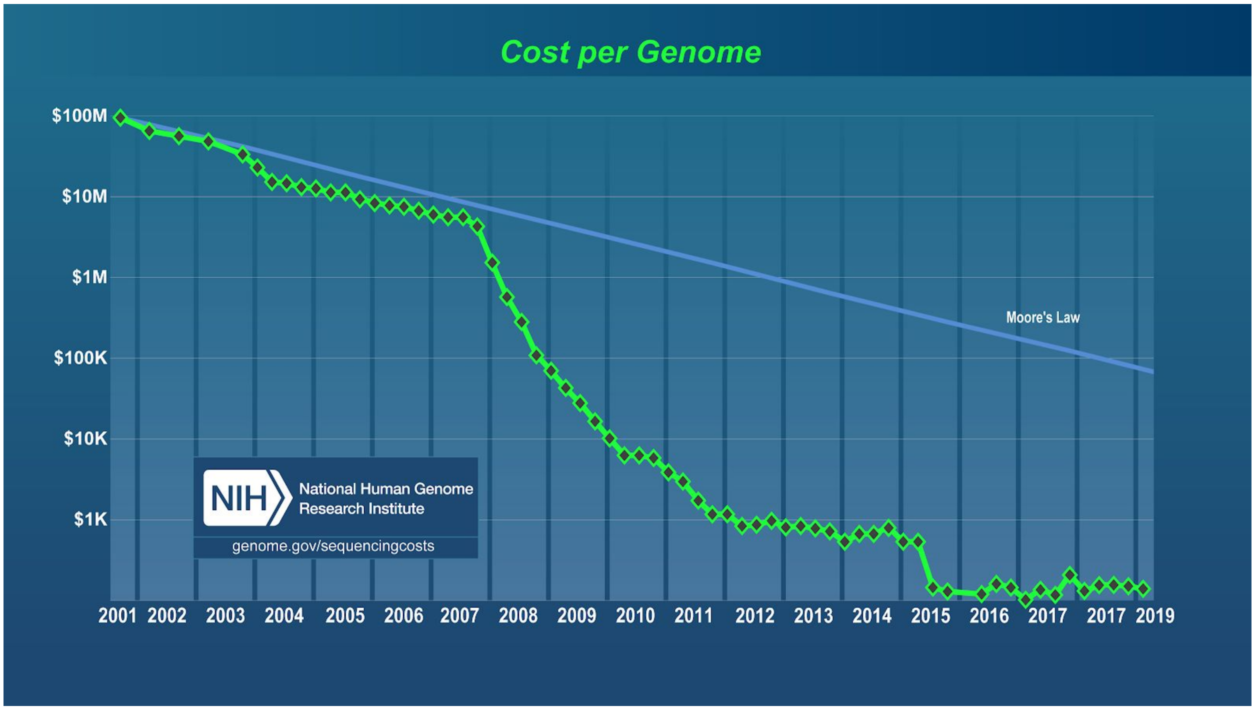


Figure 1: The dramatic decline in the cost per genome over time, source NHGRI (https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data)

Unfortunately, data production has stressed the informatics infrastructure of the field, with many groups struggling to keep up with hosting and computing on the datasets that have been made available. Infrastructure built for previous technologies is struggling to keep up and to provide adequate storage and compute to scientists eager to use these data. Repositories that previously acted as the archive of record for genomics datasets are struggling to accept the sequence data produced by large scale projects like GTEx or TOPMed. Over the next five years, we predict up to 50 petabytes of public genomic research datasets will be made available and this necessitates a rethink in the way data are stored and made accessible to researchers (Figure 2).

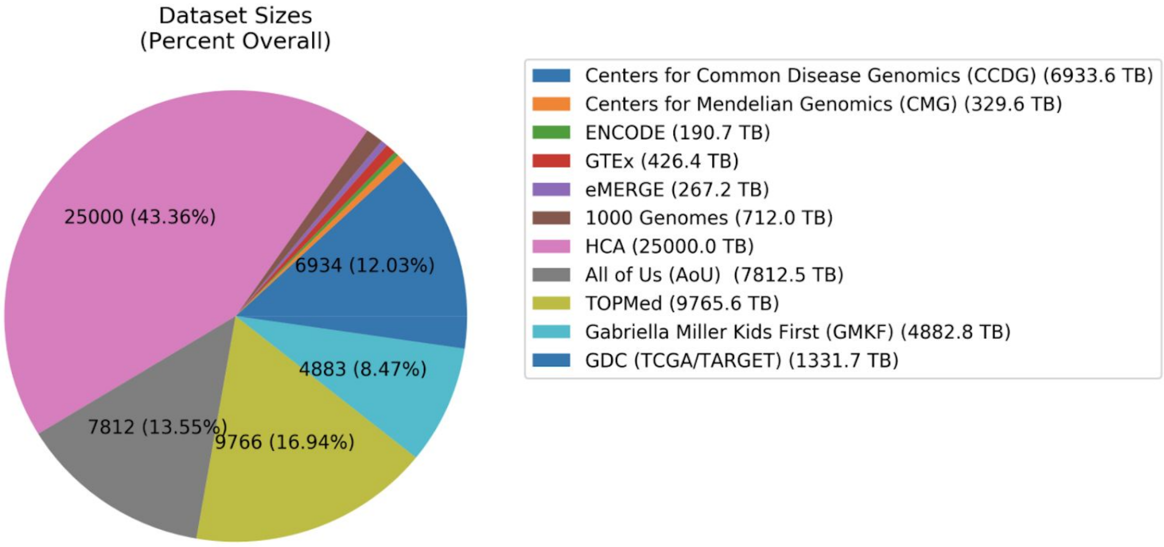


Figure 2: An approximate projection of cloud-based, genomic dataset sizes over the next 5 years based on stated data generation goals from project websites and FOAs for several notable projects.

The way scientists and organizations like the NIH have previously built archives assumes that scientists will apply for access and download datasets to their local infrastructure. This worked fine when datasets were small, megabytes to gigabytes in size, but the infrastructure breaks down when datasets are hundreds of terabytes to petabyte in scale. In this case, it can take researchers literally months to download datasets and requires expensive local infrastructure capable of handling these data (a compute cluster with adequate storage). From the archive perspective, it is incredibly costly to enable downloads from hundreds or thousands of users around the globe. Extremely significant network infrastructure is required to keep up with data transfer needs when using datasets of these sizes.

The commercial cloud represents a different way of thinking about data storage and compute. With a history of evolution that spans many decades, what we think of as the first commercial cloud emerged in 2007 with the release of the Amazon Web Services. This provided a way to “rent” storage and compute, and control that infrastructure through programmatic means. This simple concept caught on and in 2019 commercial cloud offerings by Amazon, Google, and Microsoft are projected to reach $206 billion[1].

This simple concept, that you can use a computer or storage for a per hour/per gigabyte cost and not be involved in its setup or maintenance, is transforming the way we do research. Probably most pressing for the biomedical research community is the ability to store and scale massive amounts of data on the cloud. The cloud providers have made a business of providing inexpensive and easily scaled storage to the petabyte range. While cost is a factor, the commercial cloud vendors have infrastructure that can support the extraordinary growth of data seen in the last ten years.

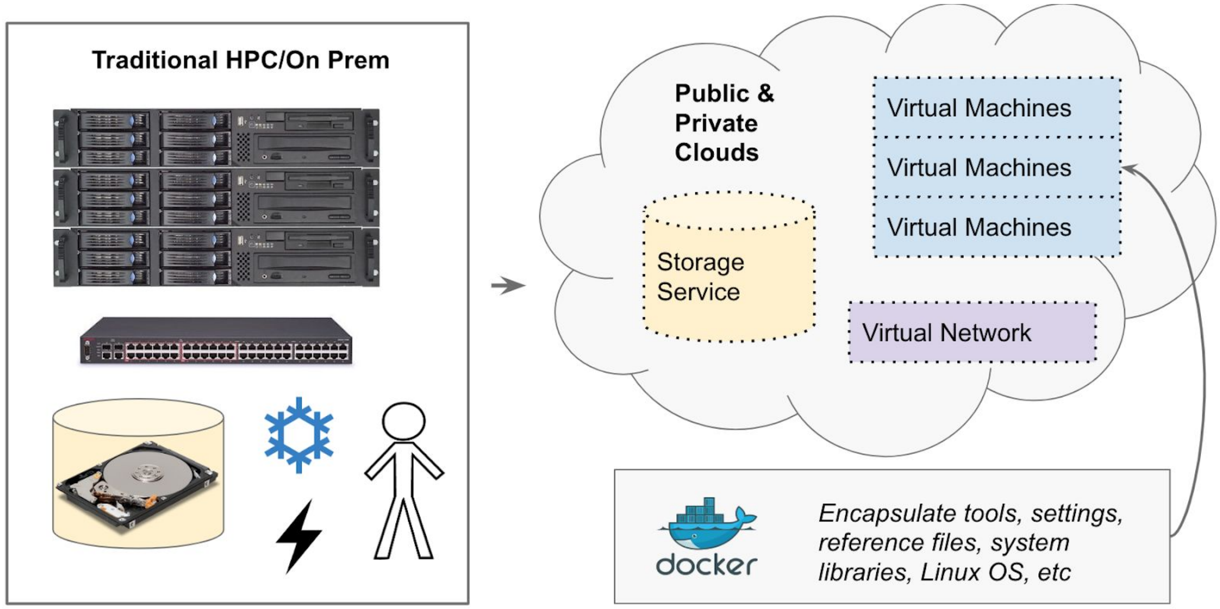


Figure 3: The cloud transition in biomedical research. Researchers are moving from local, self-hosted infrastructure that requires expensive upfront costs and maintenance to cloud systems offering compute and storage for rent. Coupled with datasets being stored and distributed on the cloud, researchers can leverage fast data access with elastic cloud compute that can grow and shrink as needed.

This transition to the cloud offers new opportunities in addition to massively scalable storage. It presents an opportunity to shift the paradigm that has been popular in the biomedical research community for decades, where researchers download data to their local compute infrastructure. The model is shifting to one where data is already resident on the commercial cloud and users need to get access to it but they no longer need to download to their local infrastructure (Figure 3). This is coupled with scalable compute offered within the commercial clouds. From a researcher's perspective, it means she or he is no longer obligated to maintain a local compute

[1] Source: https://www.forbes.com/sites/louiscolumbus/2018/09/23/roundup-of-cloud-computing-forecasts-and-market-estimates-2018/#6db0e899507b

cluster for occasional large scale analysis runs. Instead, she or he can leverage the data already resident in the cloud and simply scale up compute exactly when needed. The researcher and her or his team can launch thousands of compute nodes to complete analysis on all samples in parallel. Researchers are no longer limited by slow downloads or how many compute nodes an institution has available in a shared system.

In addition to highly scalable storage and instantaneous and elastic compute, the cloud offers other advantages. Many of the cloud vendors offer services beyond just core compute virtual machines and data storage. Google and AWS offers advanced compute environments like Spark and sophisticated machine learning toolkits as well. These offer new opportunities for researchers that otherwise would not have access to such resources. For example, most research institutes and universities offer a shared compute environment, such as an HPC cluster, and these typically provide a fairly limited range of hardware options to run analysis on. Commercial clouds, however, allow for a wide range of systems with configurable RAM and storage components and access to GPUs or specialty hardware that can accelerate machine learning tasks (such as Tensor Processing Units (TPUs) on the Google cloud).

Data storage and compute are not the only challenge facing researchers working with genomic data. Convenient SSO, authentication, and authorization approaches are also needed to streamline data access. Currently, getting access to controlled datasets can be cumbersome and time consuming. Many datasets are managed through dbGaP which requires project by project applications to access data. These access requests are reviewed by Data Access Committees (DACs) and these individual requests are approved or rejected based on the research statement and identity of the researcher making the request. Often times, these data access requests need to be revised and resubmitted and the overall process can take weeks or even months to complete. Even if data storage and compute are fully moved to the commercial cloud environments, having a cumbersome process for requesting access dataset by dataset will still prove to be an impediment to research. Longer term, the use of systems like DUOS, which look to automate data access requests and approvals based on user identity and endorsed claims on a researcher's “passport”, may greatly enhance researchers' ability to access data quickly on the cloud, ultimately speeding up her or his research.

In this report we hope to profile how current Common Fund DCCs are adapting to the shift to the cloud and look at the approaches that work for them and where there are pain points. We then extrapolate some short and long term plans for improving authentication, authorization, SSO, and cloud storage for existing DCCs while laying the groundwork for a clear set of recommendations and components for building future, cloud-ready DCCs.

Report Goals

Despite the challenges outlined in the introduction, there are many opportunities for improving researchers' access to data and compute which will ultimately facilitate scientific discovery and progress. In this report our goals are centered around four primary areas to help streamline the use of SSO, authentication, authorization, and cloud storage of data for Common Fund DCCs:

- Our primary goal in this report is to first document the needs of the Common Fund DCCs.

- We then want to examine basic technologies and standards that help support SSO, authentication, authorization, and cloud data storage.

- We further want to examine solutions currently being used by the community.

- Finally, by documenting the needs of the Common Fund DCCs and examining standards and solutions in the community, we want to identify common approaches that can be used across multiple DCCs (both in the short term and the long term).

Report Anti-Goals

In addition to clarifying our goals it is important to establish what this report is not. The topics of SSO, authentication, authorization, and cloud storage are deep, complex, and very technical. It is impossible to thoroughly cover all possible technologies and solutions in this report. So we will limit the report in the following ways:

- This report is not a definitive guide of all possible solutions.
    - There are many solutions out there, both commercial and open source, far too many to be evaluated here.
    - We are looking at general classes of technologies and examining specific implementations used in the community.

- This report is not a recommendation for a single approach.
    - Much needs to be done to fully understand the needs of all DCCs, more interviews are being scheduled.
    - It is premature to recommend particular solutions at this time.
    - This report will evolve over time as we talk with more DCCs and explore additional solutions.

- This guide will not dictate solutions for all Common Fund DCCs.
    - Likewise, as this report evolves we will learn more and document that here.
    - We are not trying to create a one-size-fits all recommendation but document a range of solutions that will be helpful to Common Fund DCCs, both current and future.

Profile of Common Fund Data Coordination Centers

Overview

The core of this report is to understand the needs of, and current pain points for, Common Fund DCCs. To that end, we are performing a series of interviews with multiple DCCs to characterize their work and understand their needs with regards to SSO, authentication, authorization, and cloud data storage. The findings presented below are an attempt to catalog and summarize these interviews.

Genotype-Tissue Expression (GTEx)

Overview

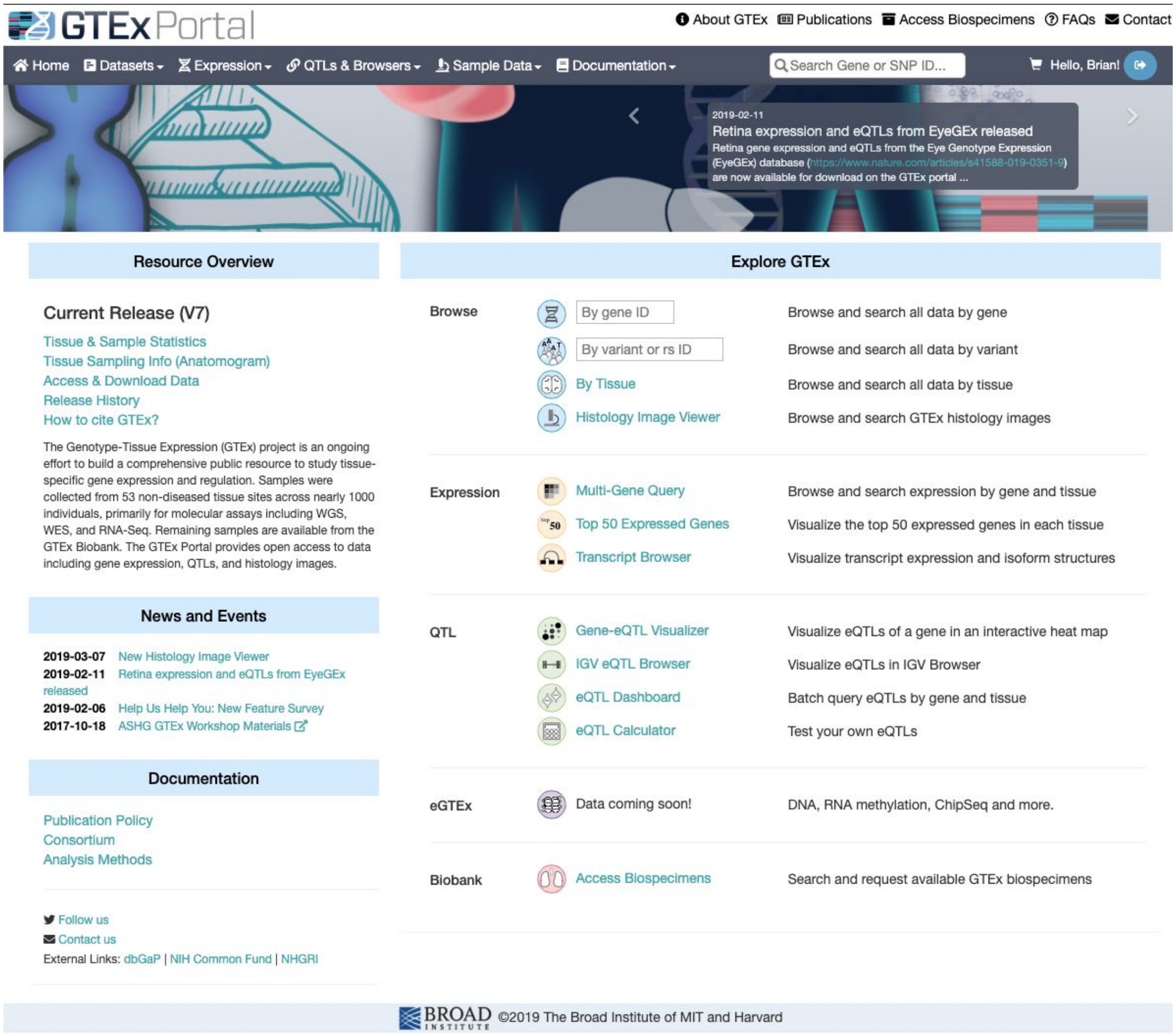
The Genotype-Tissue Expression program was created to explore the relationship between genetic variation and gene expression across many different normal tissues. The program started in 2010 and has collected samples from 53 non-disease tissues sites for nearly 1000 deceased individuals. The assays used include WGS, WES, and RNA-Seq with the project producing whole-genome sequence, RNA-seq, and eQTL analysis for over 600 adult donors (948 for the upcoming V8). Data are available through the GTEx portal, launched in 2013, with controlled access data available in dbGaP, and samples available from the GTEx Biobank.

Current DCC Functionality

The GTEx portal (https://gtexportal.org/home/) is a sophisticated and well-developed portal with rich analytical and visualization options for researchers. In addition to extensive documentation and background information, the portal allows researchers to explore the current release of GTEx data (V7, with V8 to be available in August). They can browse data by gene ID, variant, or tissue and use the incorporated histology image view to browse and search images. Expression data can be searched using a multi-gene query across genes and tissues, top expressed genes can be visualized, and transcript expression and isoform structures can be explored in their transcript browser. For QTLs, gene-eQTLs can be visualized in their interactive heat map browser or through IGV, queried by gene or tissue, and researchers can test their own eQTLs with the eQTL Calculator. While these are the major features, there are several additional visualization and analysis components available on the site, for example expression PCA. Finally, researchers can search and request biospecimens through the biospecimens browser.

The portal has a very active user community with 15K monthly users. In addition to the users of the web portal, approximately 10% of the users access data programmatically through the GTEx web API: https://gtexportal.org/home/api-docs/.

For data preparation and analysis, GTEx synchronizes pipelines with the MoTrPAC, ENCODE, and TOPMed projects whenever possible. This facilitates researchers leveraging GTEx data in comparison to other related datasets.



2019-02-11
Retina expression and eQTLs from EyeGEx released
Retina gene expression and eQTLs from the Eye Genotype Expression (EyeGEx) database (https://www.nature.com/articles/s41588-019-0351-9) are now available for download on the GTEx portal ...

Resource Overview

Current Release (V7)

Tissue & Sample Statistics
Tissue Sampling Info (Anatomogram)
Access & Download Data
Release History
How to cite GTEx?

The Genotype-Tissue Expression (GTEx) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. Samples were collected from 53 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including WGS, WES, and RNA-Seq. Remaining samples are available from the GTEx Biobank. The GTEx Portal provides open access to data including gene expression, QTLs, and histology images.

News and Events

2019-03-07	New Histology Image Viewer
2019-02-11	Retina expression and eQTLs from EyeGEx released
2019-02-06	Help Us Help You: New Feature Survey
2017-10-18	ASHG GTEx Workshop Materials

Documentation

Publication Policy
Consortium
Analysis Methods

Follow us
Contact us
External Links: dbGaP | NIH Common Fund | NHGRI

Explore GTEx

Browse	By gene ID	Browse and search all data by gene
	By variant or rs ID	Browse and search all data by variant
	By Tissue	Browse and search all data by tissue
	Histology Image Viewer	Browse and search GTEx histology images
Expression	Multi-Gene Query	Browse and search expression by gene and tissue
	Top 50 Expressed Genes	Visualize the top 50 expressed genes in each tissue
	Transcript Browser	Visualize transcript expression and isoform structures
QTL	Gene-eQTL Visualizer	Visualize eQTLs of a gene in an interactive heat map
	IGV eQTL Browser	Visualize eQTLs in IGV Browser
	eQTL Dashboard	Batch query eQTLs by gene and tissue
	eQTL Calculator	Test your own eQTLs
eGTEx	Data coming soon!	DNA, RNA methylation, ChipSeq and more.
Biobank	Access Biospecimens	Search and request available GTEx biospecimens

BROAD INSTITUTE ©2019 The Broad Institute of MIT and Harvard

Figure 4: GTExPortal provides access instructions for data and biosample ordering in addition to several easy to use analysis and visualization components.

Datasets

V7

Given the impact of the project, with over 650 papers citing the project, making the GTEx data accessible, both open access and controlled, is of the utmost importance.   V7 data includes

714 donors, 635 of which have genotyping data available for 10,361 samples. Non-controlled access and summary data can be explored directly in the portal. Controlled access data is accessible in dbGaP under accession phs000424.v7.p2 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v7.p2). These include:

- BAM files for RNA-Seq, Whole Exome Seq, and Whole Genome Seq
- Genotype Calls (.vcf) for OMNI SNP Arrays, WES, and WGS
- OMNI SNP Array Intensity files (.idat and .gtc)
- Affymetrix Expression Array Intensity files (.cel)
- Allele Specific Expression (ASE) tables
- All expression matrices from the Portal, including samples that did not pass the Analysis Freeze QC
- Sample Attributes
- Subject Phenotypes

A complete breakdown of V7 and how to access the data in dbGaP can be found at https://gtexportal.org/home/datasets.

In an effort to make data accessible on the cloud, MITRE and the NCBI teams developed a solution for providing signed URLs for data in the Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra) mirrored in the AWS and Google clouds. This provides the ability to sign URLs using a dbGaP repository key access from “My Research Projects” in dbGaP. While this solution is appealing there are multiple issues. First and foremost, not every cloud system is designed to work with signed URLs. Systems like the Broad’s Terra (https://app.terra.bio) require direct read only access to Google cloud buckets to access data. Furthermore, Google support for this signed URL service is currently considered beta and inaccessible to the research community while AWS access can only be obtained within that cloud environment, meaning it is not possible to access these data from another cloud or to download files directly. More information can be found at: https://www.ncbi.nlm.nih.gov/sra/docs/dbgap-cloud-access/.

V8

The V8 data set was finalized as a release almost 2 years ago. While it was finally released on July 19, 2019 (see https://gtexportal.org/home/v8ReleasePage), it remained inaccessible to GTEx users for a long period because of difficulties over how to host V8 (180TB) since dbGaP/SRA are no longer accepting large projects’ BAM files. As part of the NIH Data Commons Pilot Phase Consortium (DCPPC) these data were uploaded to a cloud bucket on both an AWS bucket (owned by NHLBI) and a Google bucket (owned by GTEx) as part of that pilot effort. However this pilot phase terminated and, with the difficulties in using the dbGaP signed URL approach and inability to upload new data, a new plan was needed for V8. GTEx V8 has now been onboarded into cloud bucket locations on Google through the NHGRI AnVIL project with access to those data via dbGaP application.

Current SSO, AuthN/Z, and Clouds Storage Strategies

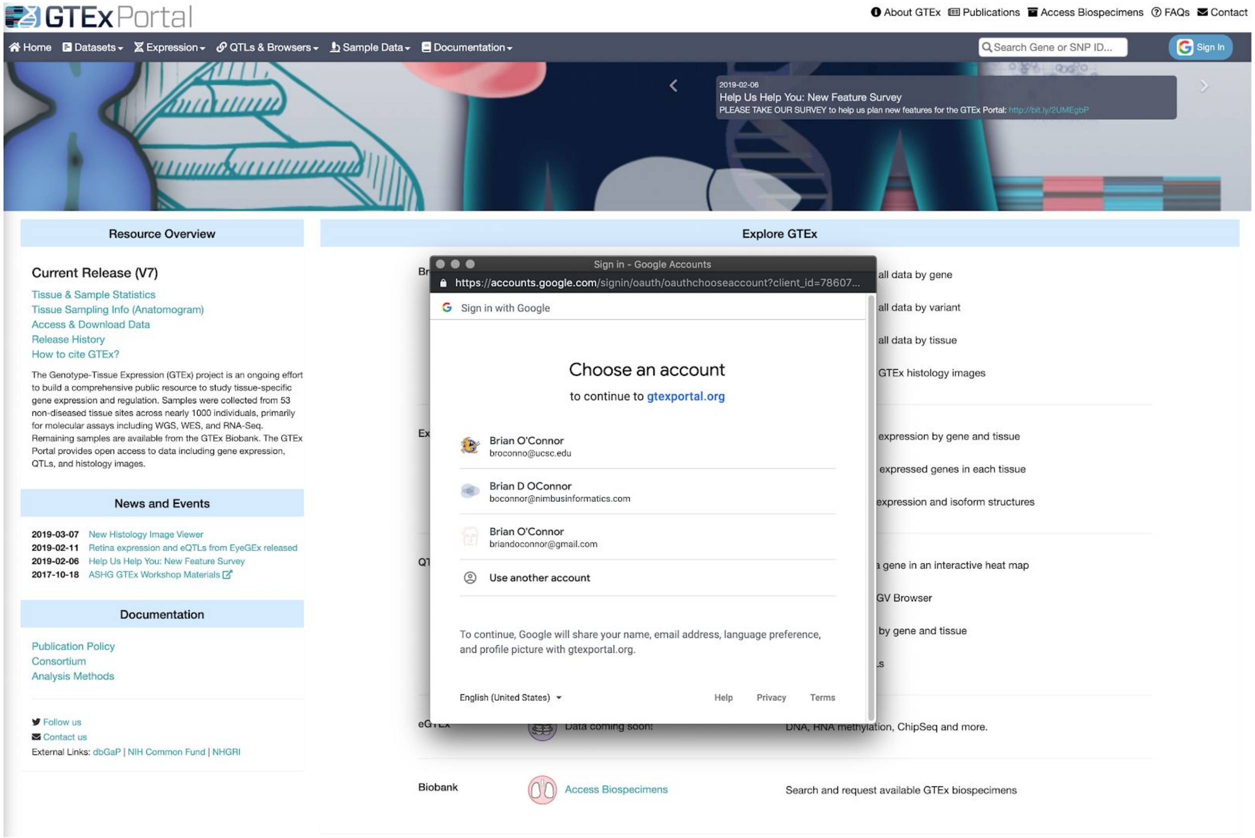
So far we have examined the functionality of the GTEx portal and the availability of their datasets. In this section we will examine their current approaches to single sign on, authentication and authorization, and cloud storage.

Single Sign On

The project effectively uses two single-sign on solutions: Google OpenID Connect (OIDC) and eRA Commons. Each are used for a different purpose and are not linked together.

GTEx Portal via Google OIDC

Because the GTEx Portal only provides access to public data, users are allowed to access the portal without logging in. The GTEx Portal only requires users to login in order to create GTEx Biobank requests. The GTEx portal uses the Google OIDC protocol to initiate a login using any Google account (GSuite, Gmail, Institutional emails powered by Google, etc). This is a standard authentication flow that allows the GTEx portal to identify the user and save state and preference for her or him. The only information shared from Google is the user name, email address, language preference, and profile picture. The token returned allows GTEx to verify this information but does not include a scope that would allow the GTEx portal to access cloud resources on behalf of the user.



2019-02-06
Help Us Help You: New Feature Survey
PLEASE TAKE OUR SURVEY to help us plan new features for the GTEx Portal: http://bit.ly/2UMEgbP

Current Release (V7)

Tissue & Sample Statistics
Tissue Sampling Info (Anatomogram)
Access & Download Data
Release History
How to cite GTEx?

The Genotype-Tissue Expression (GTEx) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. Samples were collected from 53 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including WGS, WES, and RNA-Seq. Remaining samples are available from the GTEx Biobank. The GTEx Portal provides open access to data including gene expression, QTLs, and histology images.

News and Events

2019-03-07	New Histology Image Viewer
2019-02-11	Retina expression and eQTLs from EyeGEx released
2019-02-06	Help Us Help You: New Feature Survey
2017-10-18	ASHG GTEx Workshop Materials

Documentation

Publication Policy
Consortium
Analysis Methods

Follow us
Contact us
External Links: dbGaP | NIH Common Fund | NHGRI

https://accounts.google.com/signin/oauth/oauthchooseaccount?client_id=78607...

Choose an account

to continue to gtexportal.org

Brian O'Connor
broconno@ucsc.edu

Brian D OConnor
boconnor@nimbusinformatics.com

Brian O'Connor
briandoconnor@gmail.com

Use another account

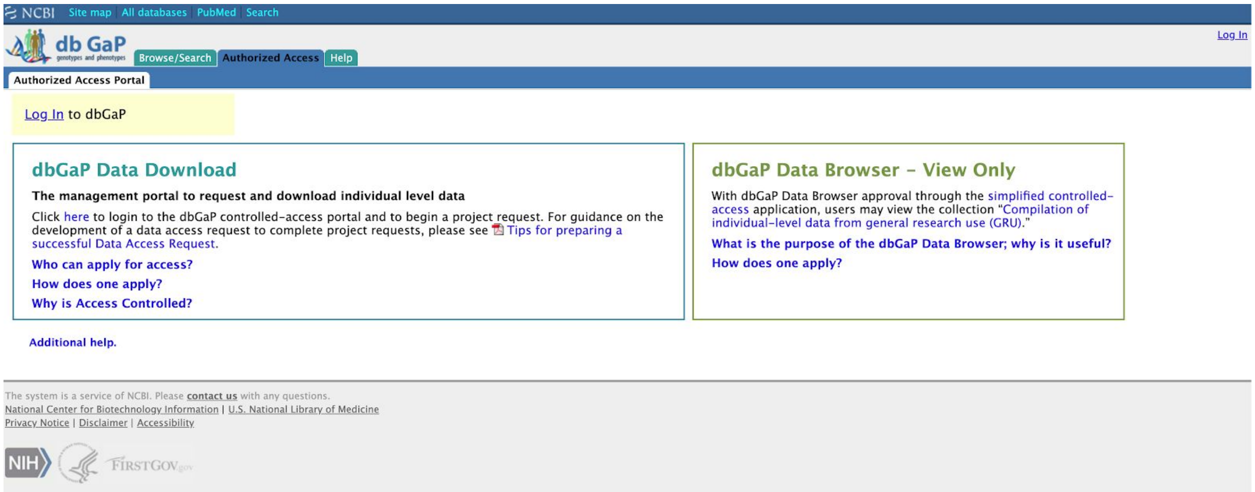
To continue, Google will share your name, email address, language preference, and profile picture with gtexportal.org.

all data by gene
all data by variant
all data by tissue
GTEx histology images
expression by gene and tissue
expressed genes in each tissue
expression and isoform structures
a gene in an interactive heat map
GV Browser
by gene and tissue
Data coming soon!
DNA, RNA methylation, ChipSeq and more.
Search and request available GTEx biospecimens

Figure 5: Clicking login takes a user to a Google to perform a login using the OIDC flow.

dbGaP via eRA Commons

Since the GTEx portal does not host controlled access data the site links to the dbGaP page for GTEx V7 (phs000424.v7.p2, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v7.p2). This page includes a link to "request access" which begins a user login flow that eventually takes the user to the eRA Commons login page.



Log In to dbGaP

dbGaP Data Download

The management portal to request and download individual level data

Click here to login to the dbGaP controlled-access portal and to begin a project request. For guidance on the development of a data access request to complete project requests, please see Tips for preparing a successful Data Access Request.

Who can apply for access?
How does one apply?
Why is Access Controlled?

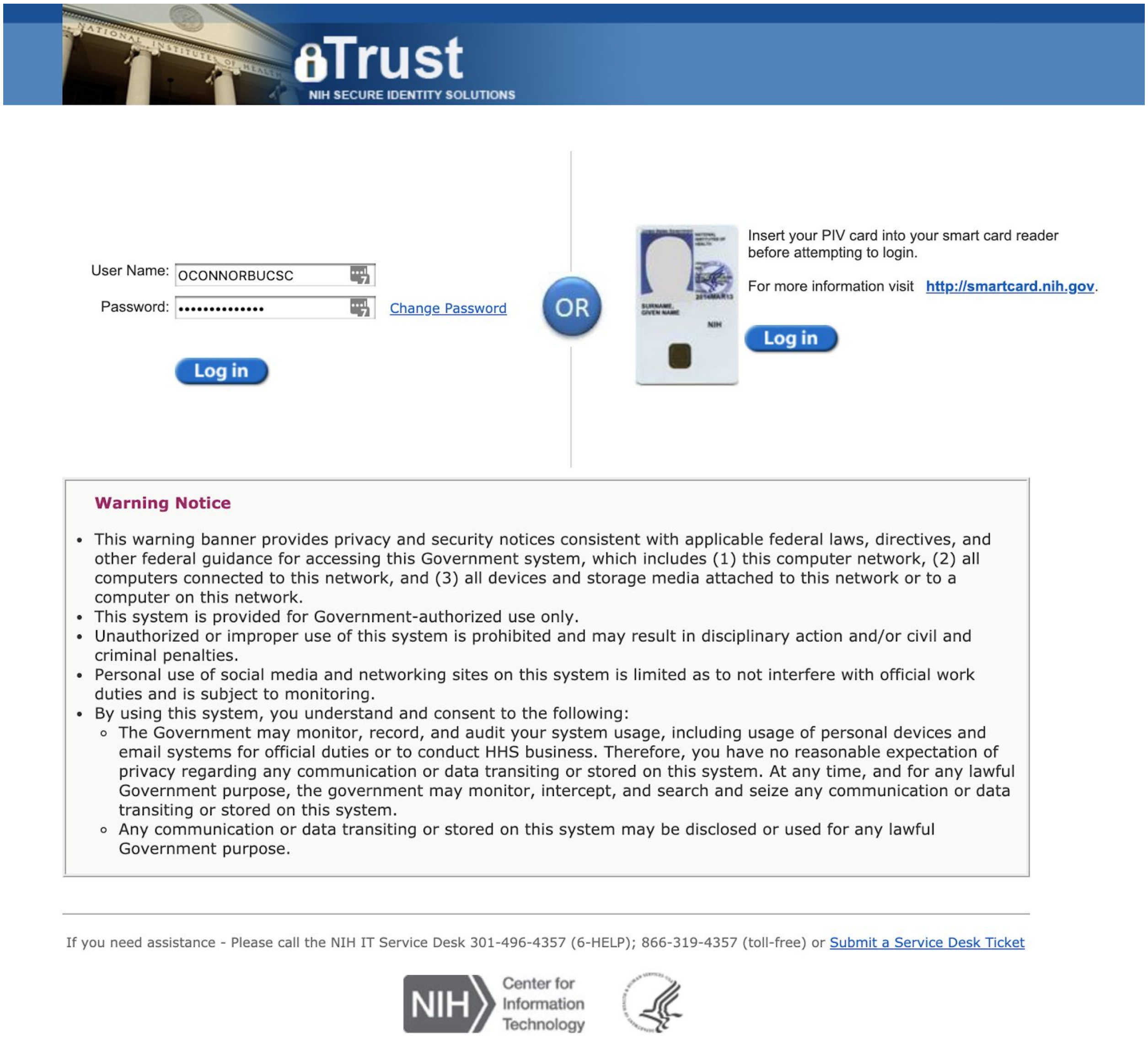
dbGaP Data Browser – View Only

With dbGaP Data Browser approval through the simplified controlled-access application, users may view the collection "Compilation of individual-level data from general research use (GRU)."

What is the purpose of the dbGaP Data Browser; why is it useful?
How does one apply?

Additional help.

The system is a service of NCBI. Please contact us with any questions.
National Center for Biotechnology Information | U.S. National Library of Medicine
Privacy Notice | Disclaimer | Accessibility


iTrust
NIH SECURE IDENTITY SOLUTIONS
Change Password
OR
Log in
Log in
SURNAME, GIVEN NAME
NIH
Center for Information Technology

User Name: OCONNORBUCSC

Password: ••••••••••••••

Insert your PIV card into your smart card reader before attempting to login.

For more information visit http://smartcard.nih.gov.

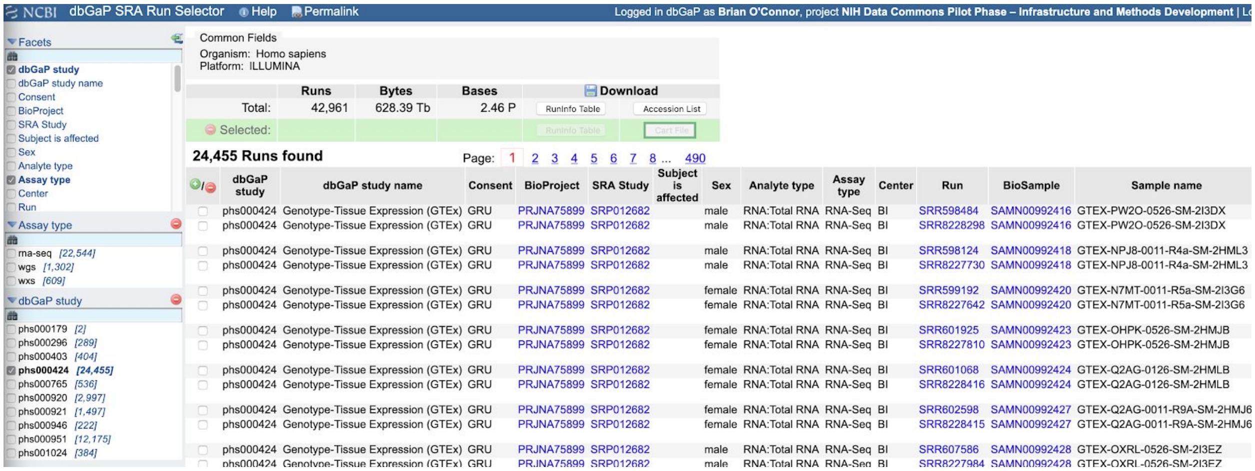
Warning Notice

- This warning banner provides privacy and security notices consistent with applicable federal laws, directives, and other federal guidance for accessing this Government system, which includes (1) this computer network, (2) all computers connected to this network, and (3) all devices and storage media attached to this network or to a computer on this network.
- This system is provided for Government-authorized use only.
- Unauthorized or improper use of this system is prohibited and may result in disciplinary action and/or civil and criminal penalties.
- Personal use of social media and networking sites on this system is limited as to not interfere with official work duties and is subject to monitoring.
- By using this system, you understand and consent to the following:
  - The Government may monitor, record, and audit your system usage, including usage of personal devices and email systems for official duties or to conduct HHS business. Therefore, you have no reasonable expectation of privacy regarding any communication or data transiting or stored on this system. At any time, and for any lawful Government purpose, the government may monitor, intercept, and search and seize any communication or data transiting or stored on this system.
  - Any communication or data transiting or stored on this system may be disclosed or used for any lawful Government purpose.

If you need assistance - Please call the NIH IT Service Desk 301-496-4357 (6-HELP); 866-319-4357 (toll-free) or Submit a Service Desk Ticket

Figure 6: eRA Commons login from the dbGaP page for V7 of the GTEx project page.

Once authenticated with eRA Commons, the researcher is taken to a dbGaP page that summarizes the datasets he or she is authorized to download from dbGaP. The researcher can then choose to download files directly from dbGaP or use the MITRE signed URL approach described in more detail in the Cloud Storage section.



Common Fields
Organism: Homo sapiens
Platform: ILLUMINA

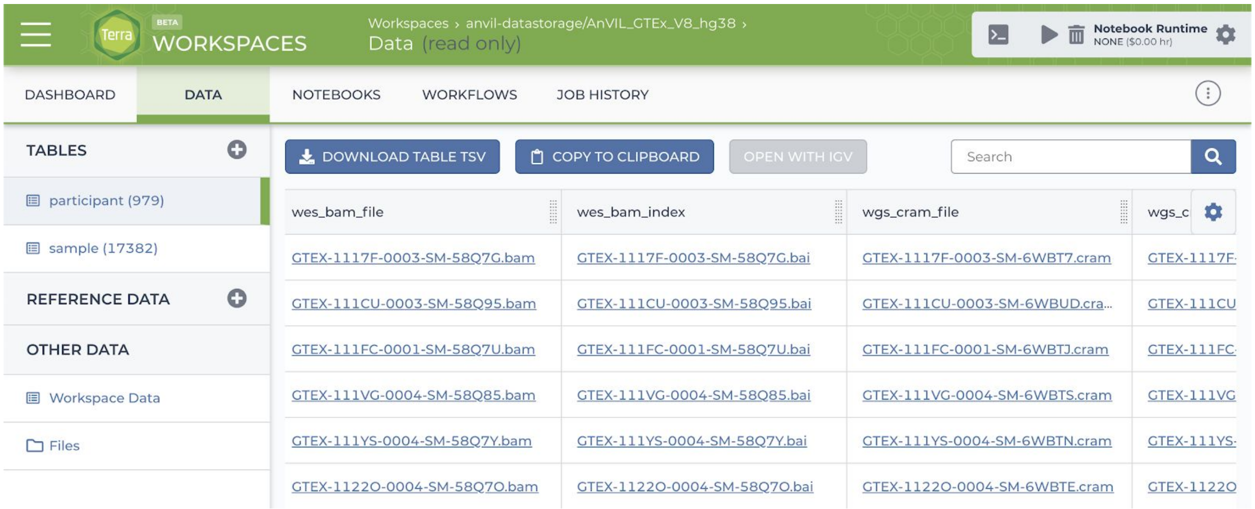
	Runs	Bytes	Bases	Download
Total:	42,961	628.39 Tb	2.46 P	RunInfo Table / Accession List
Selected:				RunInfo Table / Cart File

24,455 Runs found

dbGaP study	dbGaP study name	Consent	BioProject	SRA Study	Subject is affected	Sex	Analyte type	Assay type	Center	Run	BioSample	Sample name
phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		male	RNA:Total RNA	RNA-Seq	BI	SRR598484	SAMN00992416	GTEX-PW2O-0526-SM-2I3DX
phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		male	RNA:Total RNA	RNA-Seq	BI	SRR8228298	SAMN00992416	GTEX-PW2O-0526-SM-2I3DX
phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		male	RNA:Total RNA	RNA-Seq	BI	SRR598124	SAMN00992418	GTEX-NPJ8-0011-R4a-SM-2HML3
phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		male	RNA:Total RNA	RNA-Seq	BI	SRR8227730	SAMN00992418	GTEX-NPJ8-0011-R4a-SM-2HML3
phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		female	RNA:Total RNA	RNA-Seq	BI	SRR599192	SAMN00992420	GTEX-N7MT-0011-R5a-SM-2I3G6
phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		female	RNA:Total RNA	RNA-Seq	BI	SRR8227642	SAMN00992420	GTEX-N7MT-0011-R5a-SM-2I3G6
phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		female	RNA:Total RNA	RNA-Seq	BI	SRR601925	SAMN00992423	GTEX-OHPK-0526-SM-2HMJB
phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		female	RNA:Total RNA	RNA-Seq	BI	SRR8227810	SAMN00992423	GTEX-OHPK-0526-SM-2HMJB
phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		female	RNA:Total RNA	RNA-Seq	BI	SRR601068	SAMN00992424	GTEX-Q2AG-0126-SM-2HMLB
phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		female	RNA:Total RNA	RNA-Seq	BI	SRR8228416	SAMN00992424	GTEX-Q2AG-0126-SM-2HMLB
phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		female	RNA:Total RNA	RNA-Seq	BI	SRR602598	SAMN00992427	GTEX-Q2AG-0011-R9A-SM-2HMJ6
phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		female	RNA:Total RNA	RNA-Seq	BI	SRR8228415	SAMN00992427	GTEX-Q2AG-0011-R9A-SM-2HMJ6
phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		male	RNA:Total RNA	RNA-Seq	BI	SRR607586	SAMN00992428	GTEX-OXRL-0526-SM-2I3EZ
phs000424	Genotype-Tissue Expression (GTEx)	GRU	PRJNA75899	SRP012682		male	RNA:Total RNA	RNA-Seq	BI	SRR8227984	SAMN00992428	GTEX-OXRL-0526-SM-2I3EZ

Figure 7: the dbGaP Run Selector allows browsing of V7 GTEx data that is available for download.

For V8, even though data is not stored in dbGaP, users still need to login via eRA Commons and ensure they have correctly applied for access to GTEx data to access these data on the cloud.



wes_bam_file	wes_bam_index	wgs_cram_file	wgs_c
GTEX-1117F-0003-SM-58Q7G.bam	GTEX-1117F-0003-SM-58Q7G.bai	GTEX-1117F-0003-SM-6WBT7.cram	GTEX-1117F-
GTEX-111CU-0003-SM-58Q95.bam	GTEX-111CU-0003-SM-58Q95.bai	GTEX-111CU-0003-SM-6WBUD.cra...	GTEX-111CU
GTEX-111FC-0001-SM-58Q7U.bam	GTEX-111FC-0001-SM-58Q7U.bai	GTEX-111FC-0001-SM-6WBTJ.cram	GTEX-111FC-
GTEX-111VG-0004-SM-58Q85.bam	GTEX-111VG-0004-SM-58Q85.bai	GTEX-111VG-0004-SM-6WBTS.cram	GTEX-111VG
GTEX-111YS-0004-SM-58Q7Y.bam	GTEX-111YS-0004-SM-58Q7Y.bai	GTEX-111YS-0004-SM-6WBTN.cram	GTEX-111YS-
GTEX-1122O-0004-SM-58Q7O.bam	GTEX-1122O-0004-SM-58Q7O.bai	GTEX-1122O-0004-SM-6WBTE.cram	GTEX-1122O

Figure 8: The Terra workspace for AnVIL displays available data files stored in the Google cloud for users authorized to access these data in dbGaP.

Authentication and Authorization

For the GTEx portal the authentication flow is provided by the OpenID Connect (OIDC) sign in process. The user is identified uniquely by Google with a verifiable OAuth2 token returned to the portal. The portal does not require authorization to see its content.

Similar to the discussion in Single Sign On, dbGaP uses eRA Commons which uses the SAML protocol specification. With a token establishing the user identify, dbGaP provides web based access to controlled access files the user has approved access to for download or inspection. Furthermore, the user can create project specific access tokens for use with the URL signing service created by MITRE.

Direct access to Google buckets via the Terra workspace (https://app.terra.bio/#workspaces/anvil-datastorage/AnVIL_GTEx_V8_hg38/data) uses Google OIDC for user authentication and cross-checks with a whitelist of known users provided by dbGaP following account linking with eRA Commons via the Terra application.

Cloud Storage

The dbGaP V7 data was uploaded to cloud buckets as part of the DCPPC pilot and is used by the MITRE solution to sign URLs.

In an independent effort, the GTEx DCC has uploaded ~180TB of data from V8 to a Google bucket and has made these accessible in Terra workspaces (https://app.terra.bio/#workspaces/anvil-datastorage/AnVIL_GTEx_V8_hg38/data) for use by the NHGRI AnVIL project.

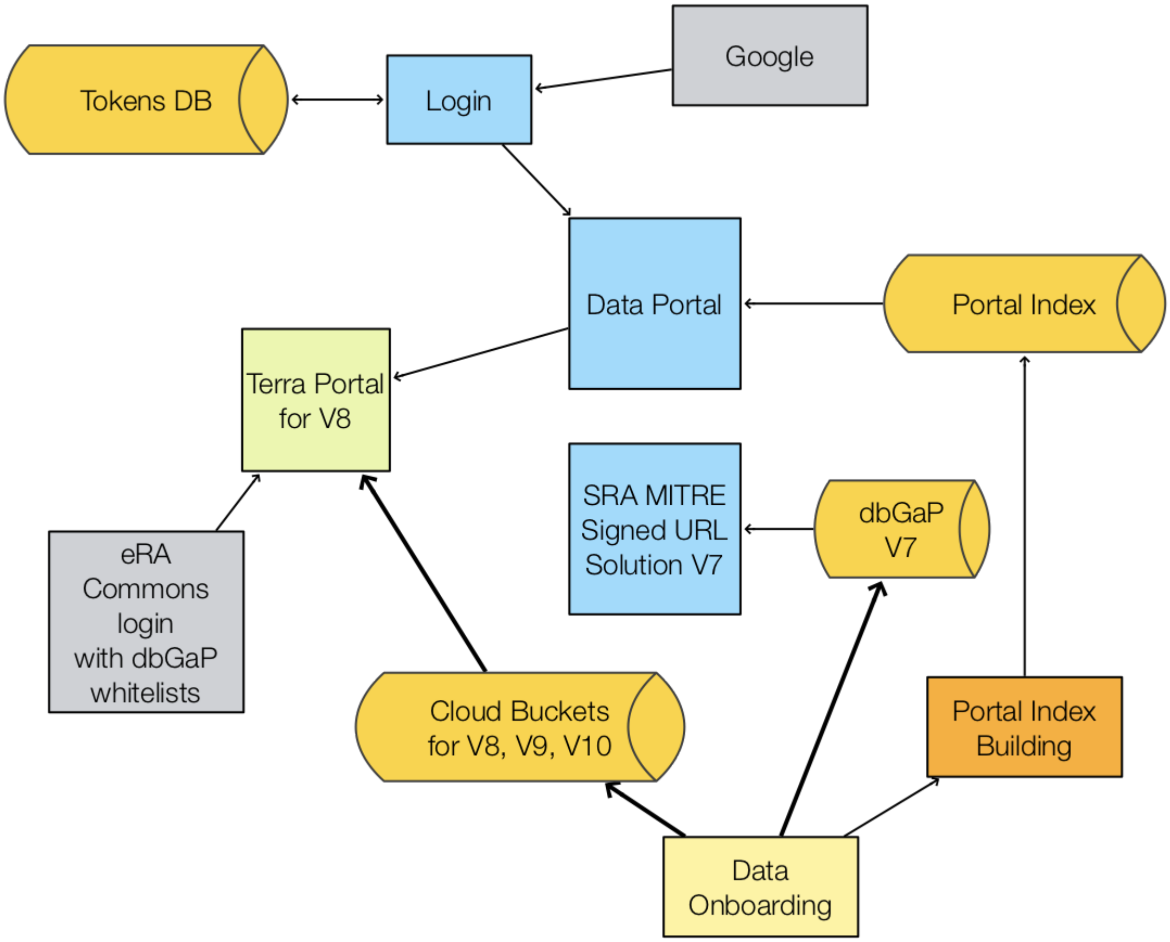


Figure 9: The current GTEx portal architecture. V8 data is supported on Terra directly through access on Google Storage buckets and linked eRA Commons/dbGaP whitelists. V7 is present in dbGaP and cloud copies are shared via signed URLs from the SRA/MITRE solution.

What Works Well?

Despite challenges of leveraging cloud technology there is so much functionality and usability in the GTEx portal, and research value in the dataset, that many researchers find it to be an invaluable resource in their work. In terms of infrastructure, these are the areas the project team thinks are going well and have been successfully leveraged.

Single Sign On

For eRA Commons the fact that this login account and mechanism is ubiquitous across NIH projects is a benefit.

For Google OIDC, the technology is standardized and widely adopted, allowing users to log in once for various Google services and quickly login to the GTEx portal without reentering their password.

Authentication and Authorization

For eRA Commons/dbGaP, the fact that dbGaP have a fully developed Data Access Committee (DAC) component, which forms the basis for authorization of users to access data, is a benefit.

For Google OIDC, it can also eventually be expanded to request cloud API access scopes which may be helpful in future cloud integrations. Regardless, these are the same identities associated with Google Cloud accounts, which makes integration with environments that leverage Google Cloud, such as Terra, much easier.

Cloud Storage

For dbGaP, the GTEx team noted how helpful the staff is for onboarding data into that system. They also mentioned the value in the basic QC work done on data submissions that, while simple, provides a nice safetynet through basic data integrity checking.

For the MITRE solution the GTEx staff were able to provide a bucket in the Google cloud where they onboarded their data and provided a manifest for describing it. They then added a service account from MITRE to the bucket to provide access. NHLBI then copied the data into an AWS S3 bucket in order to have data in both commercial clouds. Ultimately, the MITRE solution was successful in the sense that they provided signed URLs for data on the cloud. So there is a way to get access to data on the cloud, even though there are limitations and caveats to this process.

The Terra solution for V8 data makes the data finally accessible to users which is a huge benefit to the community. The fact that the Terra platform is tightly integrated with workflow and notebook execution makes working with the data extremely easy and scientific analysis readily accessible.

What Does Not Work Well?

Given the desired release of V8, challenges of leveraging the cloud environment, and difficulty using systems provided by other groups, here are the current areas where GTEx is looking for improvement.

Single Sign On

Right now eRA Commons does not provide an ability to link to other accounts, such as Google cloud accounts. While this is something they can do in their portal, it does require the GTEx team to implement an account linking function in their site (and other portals to do the same) if they want to start integrating cloud access into their portal. For V8 they have used account linking in the Terra portal to accomplish this goal.

Authentication and Authorization

For authorization encoded in dbGaP the major concerns from the GTEx team are primarily focused on process and not technology per se. One initial concern, setting up a project in dbGaP is complicated and potentially impossible to do without help. The DAC process for researchers applying for authorization to access projects is lengthy and confusing. It could be made simpler. The GTEx team observed that it is hard to understand which datasets to apply to ahead of time since it is difficult to search datasets when you are not authorized.

Cloud Storage

For traditional storage on dbGap, the major blocker from the GTEx team's perspective is that the SRA does not accept new datasets so BAM, CRAM or other large genomic data cannot be stored there. This makes dbGaP not viable for storing and redistributing GTEx BAM files for example regardless of the fact that files uploaded to dbGaP/SRA are not stored on the cloud.

In addition to the fundamental issue of not being able to upload all genomic data to dbGaP (via SRA), the GTEx team also noted several issues with onboarding data into dbGaP. Notably, they commented that the data submission process is impossible without help, it is not scalable or automatable. The XML submission format is difficult to use and the associated schema not findable. Another impediment for adding data to dbGaP is the necessity to define a project's data dictionary, there are no metadata templates that can be used to facilitate reuse across projects. This means that GTEx and another project may be representing similar or identical metadata for their samples yet not be comparable since they structure and name fields in their dictionary differently. Finally, for files stored in the SRA, the storage of data is not necessarily lossless. BAM files, for example, are modified when retrieved from the SRA.

While these issues are not directly related to cloud storage, they still provide impediments to onboarding data in dbGaP and present challenges that will persist even as the cloud storage process is worked out.

The MITRE solution, while attempting to address the lack of dbGaP data access on commercial clouds, also presents challenges to the GTEx team. Namely, they are still paying for the storage (without a STRIDES discount) and the MITRE system only currently works on AWS with Google being a closed beta (V8 is now hosted on the AnVIL cloud bucket but GTEx is currently paying for V9 and V10 pre-releases). Furthermore, just providing signed URLs makes it extremely difficult for researchers to analyze these files in workflows since the signed URLs expire relatively quickly. Also, the signed URLs do not work outside of the AWS environment, meaning they can't be used to download data outside the cloud, a use case the GTEx team still thinks is important. Finally, there's a limitation in requesting BAM files one at a time using a specialty key only accessible to individual researchers. The GTEx team would like to sign URLs within the portal to enable point and click downloads and also would like to sign URLs in bulk for multiple BAM files at a time.

The use of the AnVIL system for storing data on the cloud is problematic in that the GTEx DCC is currently paying for the storage (without a STRIDES discounts, however V8 is now hosted on the AnVIL cloud bucket but GTEx is currently paying for V9 and V10 pre-releases). Also, just storing data in a Google bucket and providing access through the Terra workspace is limiting in the use and sharing of these files. *And ideal solution would provide both signed URLs and native cloud access on both the AWS and Google clouds.*

What Would a Shorter-Term Improvement Look Like?

The current key issue with GTEx seems to be determining a route of enabling the use of the V8 release data with a wide variety of cloud platforms. Without a reliable way to submit BAM and other large files to dbGaP/SRA, finding a viable alternative that allows researchers to access the raw data is of utmost importance. The GTEx team worked with the AnVIL project to onboard V8 data in a Google bucket, share it through Terra workspaces, and this facilitates the use of the V8 data on the cloud. Building on that, having a system such as U. Chicago's Gen3 (https://gen3.org) would allow additional functionality. A framework like this could allow GTEx to onboard their data into a Google bucket, provide both signed URLs and Google native paths, and link users eRA Commons and Google Cloud IDs. This would allow data to be leveraged by web users (through signed URLs) while other users with cloud credentials associated with their eRA Commons ID would be able to use native Google cloud APIs with these data. This would be a significant improvement over the current situation where researchers can only access V8 data on Terra.

Another area of improvement would be the incorporation of eRA Commons login directly into the GTEx portal. This would facilitate integrating controlled access metadata and data directly in the portal to greatly enhance visual exploration of the GTEx data for users authorized in dbGaP to access the GTEx project data.

Together these improvements would allow users to log into the GTEx portal and explore more metadata fields and critical information that will facilitate their research, generate signed URLs for direct file download for controlled access data in the portal, and work with controlled access files directly in the Google cloud environment and analysis platforms like Terra.

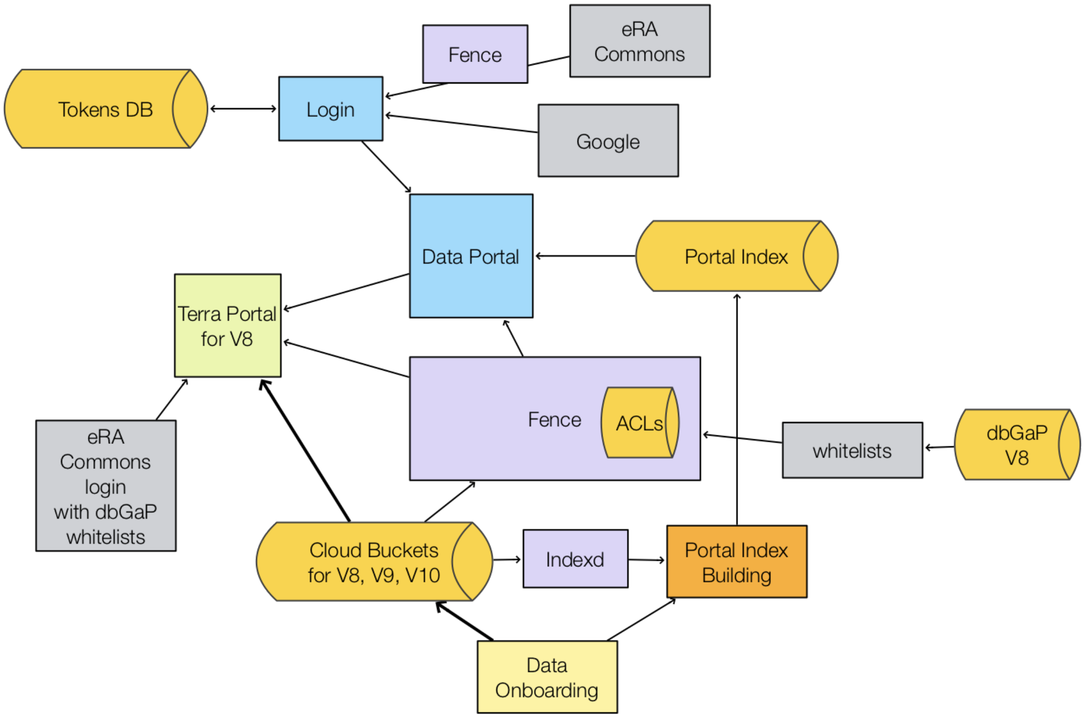


Figure 10: The DCC could use a system like Fence to provide both signed URL and native URI access to data in a cloud bucket. This would allow Terra to work with V8 data (as it does now) and also allow the portal to offer controlled access data via signed URLs for direct download for user logged in with eRA Commons.

What Would a Longer-Term Solution Look Like?

Longer term the GTEx team could benefit from a common way of linking eRA Commons IDs to IDs used in the cloud and potentially other IDs as well. This is in line with work by the GA4GH which is looking to create a series of standards that facilitate user passports that aggregate identities that can be used across multiple systems. To further help with interoperability, CIT has created and is testing an OIDC gateway for eRA Commons, this would further align that system with a more modern authentication standard.

Another area of enhancement could be based on the prototype work of NCBI who is developing a JWT-based system (see “NCBI FEDERATED IAM AND CLOUD DEPLOYMENT PROTOTYPE”) for representing authorization claims for a user from dbGaP. This would replace the inflexible and potentially stale whitelists used by many projects for providing access to controlled access data in portals and on the cloud. An enhancement suggested by the GTEx team to this prototype system would be reverse lookup. While the current JWT prototype allows a caller to ask the question “does a given user have access to this file” the GTEx team suggested the ability to, for a given resource, lookup the users who have access to it.

Finally, ensuring GTEx V8 and future releases are accessible in clouds beyond Google, such as AWS, would be a huge benefit for users on these alternative cloud environments, of which there are many. This could be accomplished by supporting both native access and signed URLs on AWS and potentially other clouds.

Gabriella Miller Kids First Pediatric Data Resource

Overview

The Gabriella Miller Kids First Pediatric Data Resource Center combines clinical and genomic data from a wide range of structural birth defects and childhood cancers. Currently the GMKF Data Resource Portal has made cohorts consisting of nearly nine thousand patients and their families available in the platform. Data types include WGS, RNA-Seq, WXS, miRNA-Seq and data has been harmonized with consistent pipelines while metadata conforms to a consistent schema. These two approaches ensure that data is compatible across studies and cohorts.

On top of the data stored in the cloud, the Kids First Data Resource provides a sophisticated data portal, allowing researchers to search across data and metadata for data files of interest. These search results, or synthetic cohorts, can then be exported to the Cavatica analysis environment which provides a platform for writing and executing analytical workflows. Researchers can further use this environment to collaborate with others on their research questions.

Current DCC Functionality

The GMKF Data Resource Portal (https://portal.kidsfirstdrc.org) provides an extremely simple to use, but powerful, set of features for finding data and performing cloud-based analysis. The site's entry point is the dashboard which provides various high-level views of the data accessible on the portal. For example, the Dashboard gives a summary of the datasets the logged in user currently has approval to access. It also summarizes Cavatica projects (see below) and saved queries from previous searches on the site. Finally, it summarizes datasets, participants, research interests of members, and diagnosis in a series of plots.

Functionality

The site also includes capabilities for building sophisticated searches through the Explore Data feature (currently in beta). This section of the site includes a series of dynamic plots that show overall survival, age at diagnosis, demographic information, diagnoses, study overview, and summary of the available data including data type and experimental strategy.

At the top of the page are a series of facets including some frequent quick filters (including data type and diagnosis categories), along with study, demographic, clinical, and biospecimen filters. As filters are applied the queries are built up in a query summary section of the page ("Combine Queries") which allows queries to be combined with "and" or "or". This is extremely powerful since most faceted browsers, made popular with shopping websites like Amazon, are limited in their ability to construct complicated queries. The “Combine Queries” feature allows users of the GMKF data portal to create more complicated queries by chaining them together with “and” or “or” relationships, the plots and summary statistics below are updated in real time as the queries are applied (Figure 11).

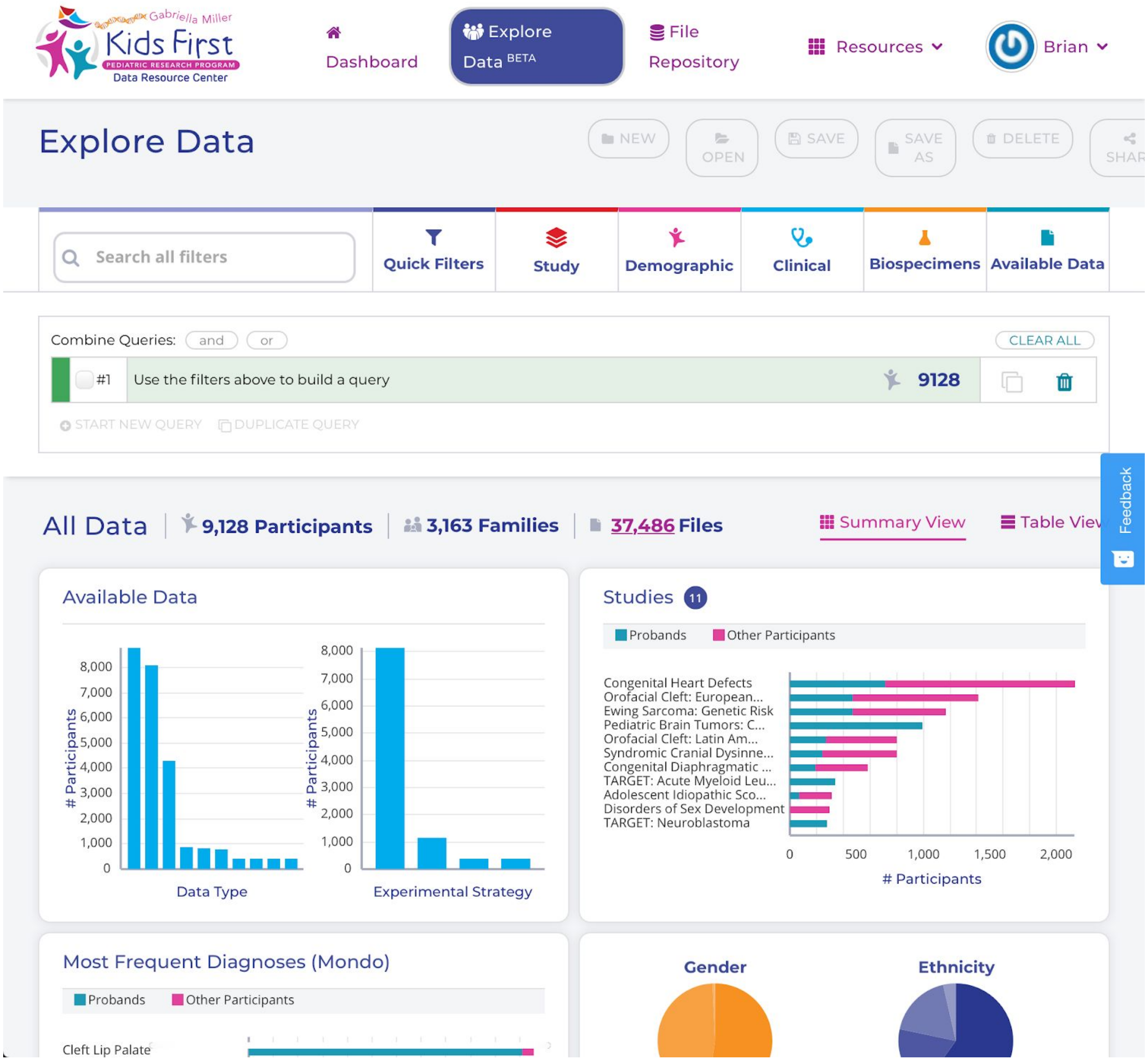
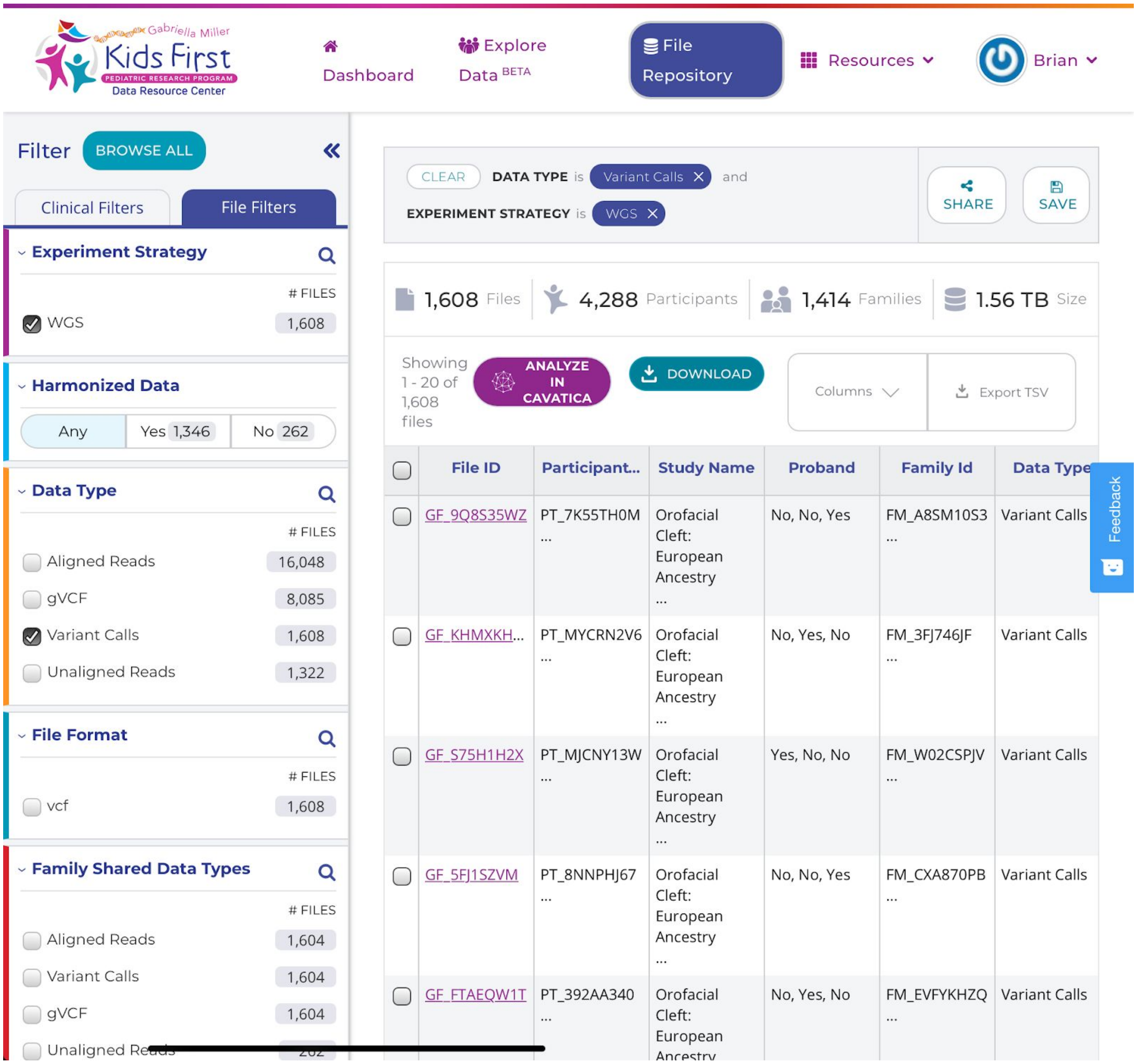


Figure 11: The Explore Data interface.

Once a user has completed their query building, they have a choice of downloads. They can use the Download button to retrieve a summary file with clinical data for participants, participants plus family members, or biospecimen data.

The query can also be saved on the site and shared via a URL. This allows researchers to come back over time, re-run queries for updated results, modify queries, and share the results with collaborators.

The File Repository section of the portal allows users to interact with similar data compared to the Explore Data section of the site but with a focus on file access (File 12). Researchers can filter files on a variety of clinical filters such as study name, observed phenotype, and tissue type. There are a variety of file filters as well, including filters for files from particular experimental design, data type, and file format.



Experiment Strategy

	# FILES
☑ WGS	1,608

Harmonized Data

Any | Yes 1,346 | No 262

Data Type

	# FILES
□ Aligned Reads	16,048
□ gVCF	8,085
☑ Variant Calls	1,608
□ Unaligned Reads	1,322

File Format

	# FILES
□ vcf	1,608

Family Shared Data Types

	# FILES
□ Aligned Reads	1,604
□ Variant Calls	1,604
□ gVCF	1,604
□ Unaligned Reads	202

CLEAR DATA TYPE is Variant Calls and EXPERIMENT STRATEGY is WGS

1,608 Files | 4,288 Participants | 1,414 Families | 1.56 TB Size

Showing 1 - 20 of 1,608 files

□	File ID	Participant...	Study Name	Proband	Family Id	Data Type
□	GF_9Q8S35WZ	PT_7K55TH0M ...	Orofacial Cleft: European Ancestry ...	No, No, Yes	FM_A8SM10S3 ...	Variant Calls
□	GF_KHMXKH...	PT_MYCRN2V6 ...	Orofacial Cleft: European Ancestry ...	No, Yes, No	FM_3FJ746JF ...	Variant Calls
□	GF_S75H1H2X	PT_MJCNY13W ...	Orofacial Cleft: European Ancestry ...	Yes, No, No	FM_W02CSPJV ...	Variant Calls
□	GF_5FJ1SZVM	PT_8NNPHJ67 ...	Orofacial Cleft: European Ancestry ...	No, No, Yes	FM_CXA870PB ...	Variant Calls
□	GF_FTAEQW1T	PT_392AA340 ...	Orofacial Cleft: European Ancestry	No, Yes, No	FM_EVFYKHZQ ...	Variant Calls

Figure 12: The File Repository interface.

Once the user has completed her or his query, they have several options. First, they can share the query, much like they can with Explore Data, using a URL or save the search query for use later. They can explore the search results as a TSV, in this case it includes all the columns displayed on the search result table and these can further be adjusted, adding or removing columns of metadata as needed.

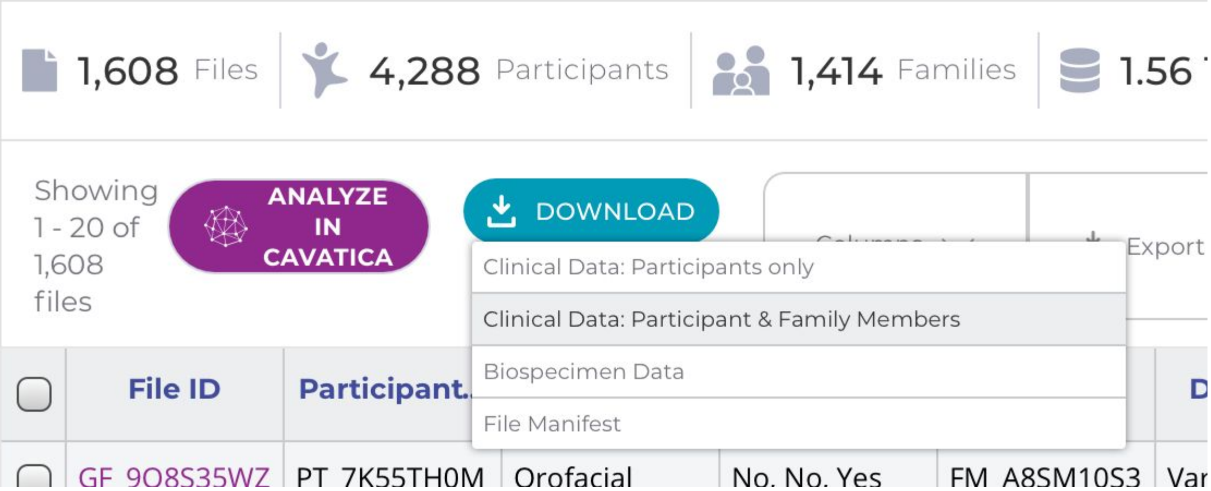


Figure 13: Download options available in the portal.

As with the Explore Data section of the site, a user can select Download which gives the same options to download clinical or biospecimen data (Figure 13). Unlike the Explore Data section, however, the File Repository allows users to download a File Manifest, a list of files that match the search specified by the user.

Probably the most significant feature of the GMKF data portal is the ability to send search results from the File Repository to Cavatica. Unlike many other DCCs, the GMKF portal presents a cloud-first approach. Rather than focusing on data downloads, the portal facilitates users finding data and taking these data to an analysis environment. This is an incredibly powerful model since users can search with a data browser to identify datasets of interest and hand those off to the Cavatica work space environment for running arbitrary batch analysis. These can be workflows written by the researcher or from the community and can contain steps with most types of analysis tools. Since data is not copied, instead referenced by ID, the handoff process to Cavatica takes just a few seconds and the search results can immediately be analyzed in that platform. In comparison to DCCs that focus on download, which can take days or weeks for large datasets, this model is extremely efficient and enables immediate productivity for researchers willing to use the Cavatica analysis platform. To facilitate researchers' transition to the cloud, the GMKF project offers cloud credits for use on the Cavatica platform. In terms of user adoption, the GMKF Data Portal and Cavatica integration have approximately 500 registered users with approximately 200 being regular, active users of the platform.

Infrastructure

The current GMKF Data Portal is the product of about a year's worth of development effort with a modestly sized team of approximately eight software developers. The portal team was able to achieve this significant accomplishment through the clever reuse of existing systems originally created as part of the NCI Cloud Pilot program and evolved into the Gen3 platform. The system built includes data storage on the commercial cloud in Amazon Web Services using the Indexd and Fence services provided by U. Chicago's Gen3 software running in the Bionimbus Protected Data Cloud FISMA moderate compliant environment (https://bionimbus.opensciencedatacloud.org/). The University of Chicago hosts, monitors, and audits the Bionimbus system which previously achieved NCI Trusted Partner status. All components run on the AWS cloud including the portal, Gen3 and files stored for GMKF, and the Cavatica analysis environment.

The use of Bionimbus allowed the GMKF portal and associated data to be hosted within the existing FISMA moderate Bionimbus environment, greatly simplifying the process of receiving an Authority to Operation (ATO) and speeding up the development and availability of the portal. This meant within the first year they had deployed the portal, made raw data accessible on the cloud through the partnership with Bionimbus, and linked to Cavatica as the data analysis environment.

Datasets

The GMKF genomic data files come from 4 sequencing centers with clinical data coming from many different locations. The Data Resource Center as of July 2019 has 8,809 participants corresponding to 3,098 families available in the Data Resource Portal. This corresponds to 37,490 files and almost 1PB of data (927TB). Harmonized CRAM alignment files using a consistent pipeline are available for 8,134 participants with much smaller numbers of participants with WXS, RNA-Seq, and miRNA-Seq harmonized data available (240, 256, and 249 respectively). Harmonized pipelines are based on the GATK best practices workflows maintained and distributed by the Broad. Unharmonized data for RNA-Seq is available for 921 participants. In terms of data growth, the Data Resource Center anticipates approximately 6-10K genomes per year with a footprint of approximately 20GB per genome. The current total storage footprint of the Data Resource Center is ~3 PB accounting for alternative workflow outputs stored per sample.

The data managed by the Data Resource Center and available in the portal is hosted on the AWS cloud environment. Approximately 1.5 years ago the Short Read Archive at NCBI stopped accepting WGS data for large-scale projects including the GMKF project. As a result the Data Resource Center had to find an alternative and embarked on creating their own, cloud-based solution. As a result the GMKF datasets are all hosted on Gen3 services running in the Bionimbus environment on AWS. The system supports users accessing these data in Cavatica without delays for file transfer since both systems leverage the same cloud. The system also supports embargo, allowing the project to redistribute data after a 6 month time period.

Current SSO, AuthN/Z, and Cloud Storage Strategies

So far we have examined the functionality of the GMKF Data Resource Center and associated portal. In this section we will examine the current approaches to single sign on, authentication and authorization, and cloud storage.

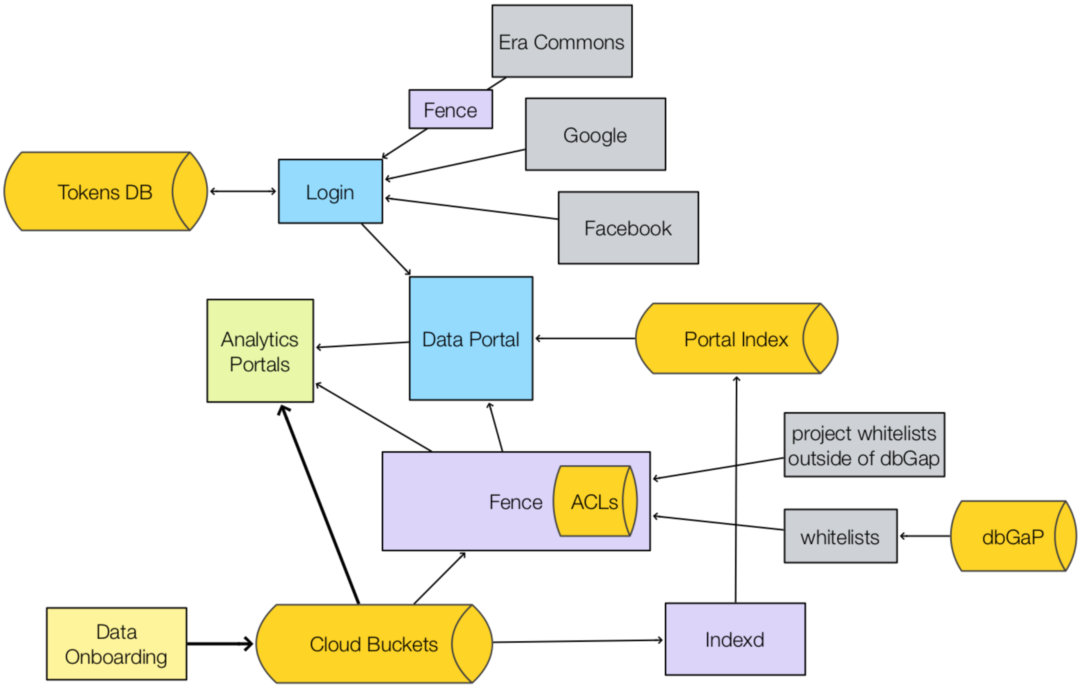
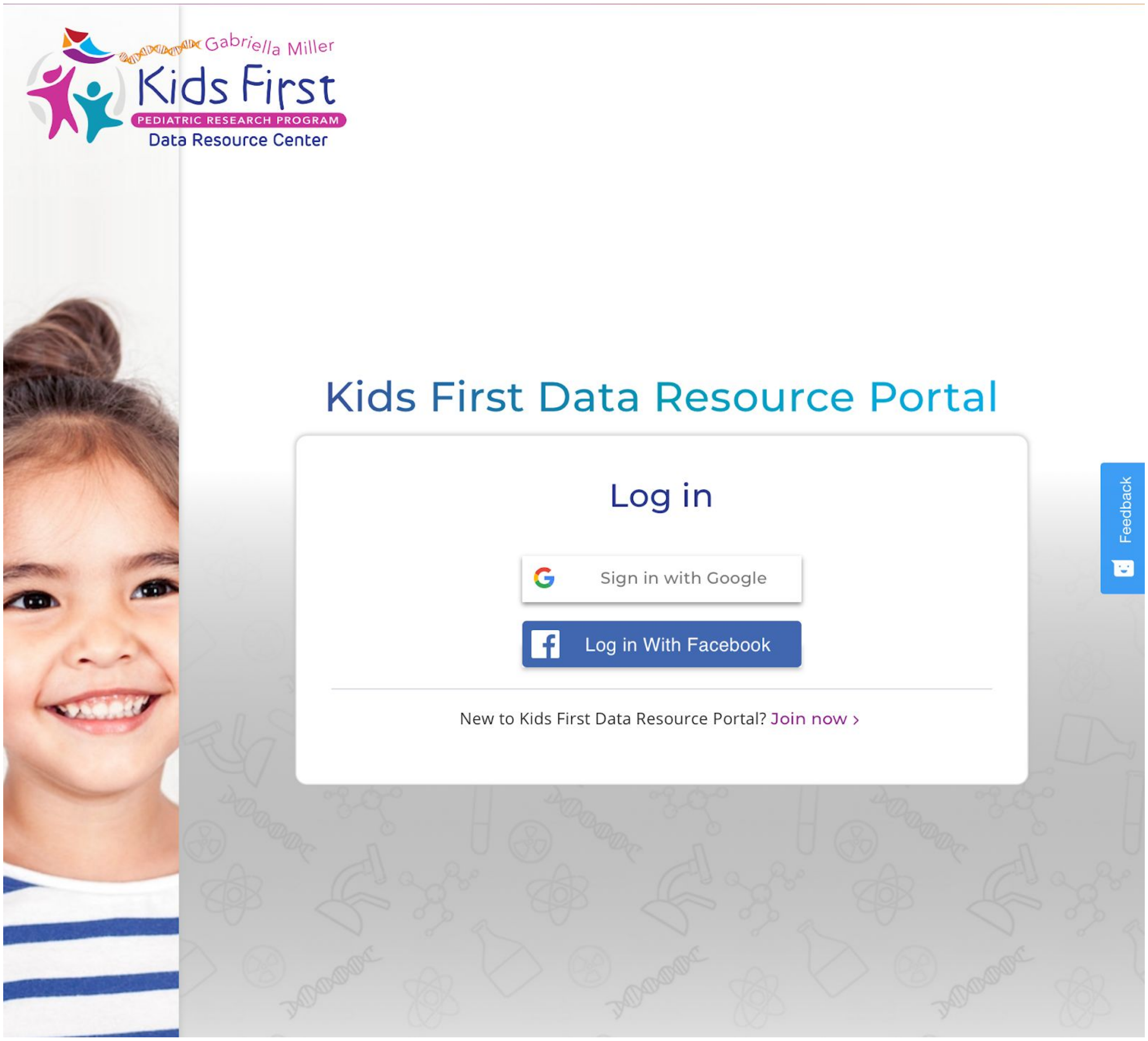


Figure 14: The current architecture of the GMKF Data Resource Center. Arrows represent information/data flow with darker lines indicating large data file access.

Single Sign On

The GMKF Data Resource Portal supports login using the Google OIDC service (https://developers.google.com/identity/protocols/OpenIDConnect) as well as the Facebook Login (https://developers.facebook.com/docs/facebook-login/manually-build-a-login-flow). Google login uses the standardized OIDC flow while Facebook uses a custom process following an OAuth 2.0-style flow. Each allow the Data Resource Portal to log in users and verify their identities, associating portal users with an identity from a trusted provider.



Kids First Data Resource Portal

Log in

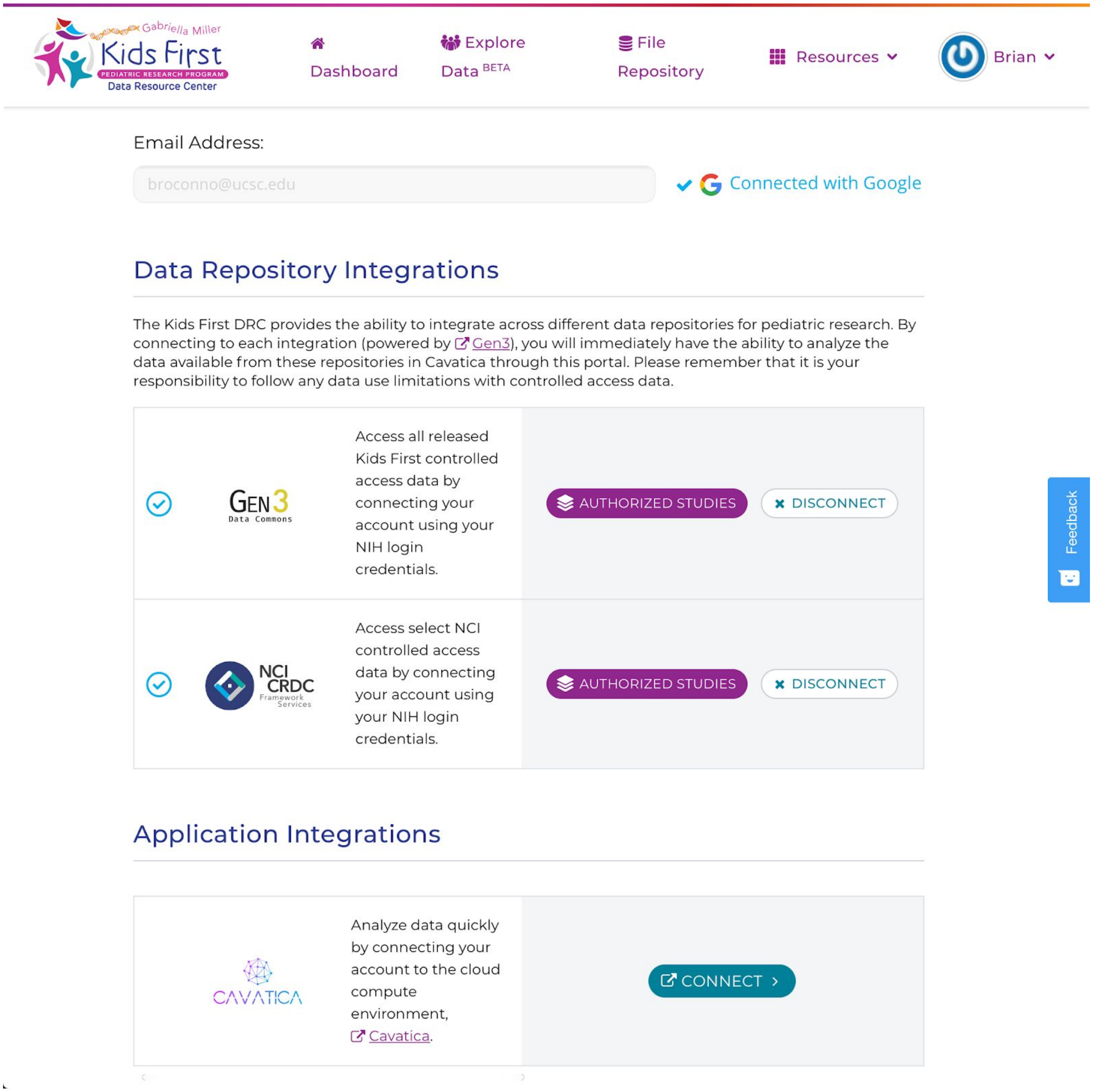
New to Kids First Data Resource Portal? Join now ›

Figure 15: The GMKF Data Resource Portal uses either Google or Facebook login SSO solutions for establishing the identity of users.

Authentication and Authorization

While Google or Facebook OIDC/OAuth-like flows establish the authentication of a given user with a popular identity provider (IdP), the actual identity is not valid for accessing GMKF data. The Data Resource Portal takes the approach of linking these identities to eRA Commons IDs which are associated with data access privileges stored in dbGaP for the vast majority of datasets. For the single project that does not use dbGaP, a whitelist approach is used to associate Google or Facebook identities with access privilege decisions from the Data Access Committee for this project.

The actual linking between Google/Facebook IDs and data repositories is delegated to the Gen3 stack from U. Chicago, specifically the Fence component. The portal presents an interface to end users in the settings section of the site to link their account to two different data repositories, the Gen3 environment on Bionimbus and Gen3 on the NCI GDC. The flow for each is a standard OAuth 2.0 flow where users are redirected to login via eRA Commons for both Gen3 and NCI GDC which, in turn produces a refresh token for the GMKF portal which is then stored and association with the users Google/Facebook identities. This token is then used to create access tokens which are themselves used to access files via Fence in each of the Gen3 environments.



Email Address:

broconno@ucsc.edu ✓ Connected with Google

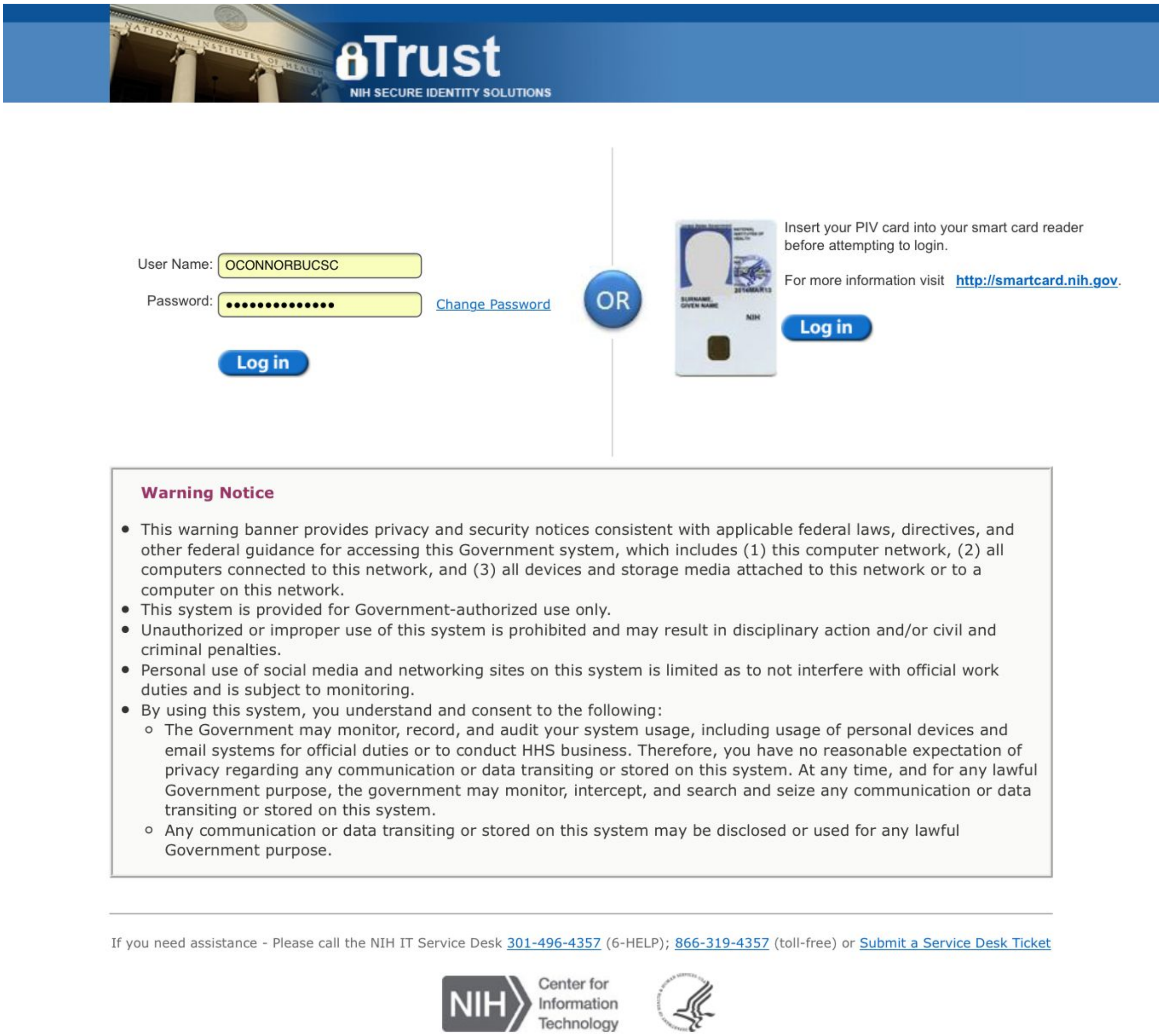
Data Repository Integrations

The Kids First DRC provides the ability to integrate across different data repositories for pediatric research. By connecting to each integration (powered by Gen3), you will immediately have the ability to analyze the data available from these repositories in Cavatica through this portal. Please remember that it is your responsibility to follow any data use limitations with controlled access data.

✓	GEN3 Data Commons	Access all released Kids First controlled access data by connecting your account using your NIH login credentials.	AUTHORIZED STUDIES ✕ DISCONNECT
✓	NCI CRDC Framework Services	Access select NCI controlled access data by connecting your account using your NIH login credentials.	AUTHORIZED STUDIES ✕ DISCONNECT

Application Integrations

CAVATICA	Analyze data quickly by connecting your account to the cloud compute environment, Cavatica.	CONNECT ›



User Name: OCONNORBUCSC

Password: ••••••••••••••

Insert your PIV card into your smart card reader before attempting to login.

For more information visit **http://smartcard.nih.gov**.

Warning Notice

- This warning banner provides privacy and security notices consistent with applicable federal laws, directives, and other federal guidance for accessing this Government system, which includes (1) this computer network, (2) all computers connected to this network, and (3) all devices and storage media attached to this network or to a computer on this network.
- This system is provided for Government-authorized use only.
- Unauthorized or improper use of this system is prohibited and may result in disciplinary action and/or civil and criminal penalties.
- Personal use of social media and networking sites on this system is limited as to not interfere with official work duties and is subject to monitoring.
- By using this system, you understand and consent to the following:
  - The Government may monitor, record, and audit your system usage, including usage of personal devices and email systems for official duties or to conduct HHS business. Therefore, you have no reasonable expectation of privacy regarding any communication or data transiting or stored on this system. At any time, and for any lawful Government purpose, the government may monitor, intercept, and search and seize any communication or data transiting or stored on this system.
  - Any communication or data transiting or stored on this system may be disclosed or used for any lawful Government purpose.

If you need assistance - Please call the NIH IT Service Desk 301-496-4357 (6-HELP); 866-319-4357 (toll-free) or Submit a Service Desk Ticket

Figure 16: The GMKF Data Resource Portal associates the identity provided by Google or Facebook with data repositories for Kids First and NCI datasets and the Cavatica analysis environment.

Cloud Storage

The GMKF Data Resource Center does not, itself, directly host data in the AWS cloud but, instead, partnered with the University of Chicago to provide cloud storage and access to GMKF data via Gen3 in the Bionimbus compliance environment. Likewise, since an almost identical stack is used for the NCI environment, the GMKF portal can integrate with that project as well.

Data on the cloud for both GMKF and NCI datasets are indexed using the Indexd service which maintains a mapping of IDs to file locations on the AWS cloud. For users to access data files, though say Cavatica, the refresh token stored for Gen3 (both GMKF and NCI) can be used to create an access token that can, itself, be used to generate signed URLs or native paths with temporary cloud access credentials using the Fence service. This is only done for users that are authorized to access these files, the refresh/access tokens establish the user identity and Fence makes use of this, plus whitelists from dbGaP, to ensure only authorized users can access controlled data.

What Works Well?

It is remarkable that the GMKF Data Resource Portal was created and operational in approximately one year. The level of sophistication and data access abilities available in the portal suggest a much longer development process. One of the reasons the work was able to move so quickly was the reuse of key infrastructure components from U. Chicago including both the Gen3 software stack as well as the Bionimbus compliant environment. To create an environment with FISMA moderate certification is no small task and typically can take on the order of 2 years. Partnering with U. Chicago and using an existing environment and software for data storage and access allowed the GMKF Data Resource Portal to fastrack some of the most complicated and time consuming aspects of deploying a new DCC portal built to redistribute controlled access data.

Single Sign On

The use of Google and Facebook as IdPs in the GMKF portal was a simple and practical choice. These identity providers are well known, have detailed documentation on using each as an identity provider to authenticate users, and software libraries exist in many languages, making it easy to incorporate into the portal.

While Google and Facebook logins are ubiquitous and easy to setup and provide SSO interactions for the end user, they provide no identity verification and attributes necessary to authorize users. Instead, delegating this to the Gen3 stack meant the existing implementation of SAML-based eRA Commons authentication could be used with the portal.

Authentication and Authorization

As mentioned in the SSO section, the use of Google or Facebook for authentication was sufficient for identifying a user and saving queries and other state in the portal. However, these identities are not suitable for accessing controlled access data.

The use of Gen3 was important for two reasons, first it included a self-contained eRA Commons login ability, so the GMKF portal could redirect to Gen3 to perform the eRA Commons login over SAML. Once the identity was verified, the Gen3 stack could then issue a refresh token for providing access to files in both the GMKF and NCI instances of Gen3. This off-the-shelf functionality hid the complexity of interacting with the NCBI’s eRA Commons system. Behind the scenes Gen3 synchronizes a list nightly of those eRA Commons IDs that should have access to files from projects represented in the system. Again, the details of this are hidden from the portal or the Cavatica system which streamlined and simplified their development.

Cloud Storage

The Indexd component of Gen3 provided the ability to catalog and index the available files for GMKF on AWS. It was used to assign identifiers to each file that could later be used to retrieve access to the bytes. The Fence component of Gen3 provided the ability to create signed URLs or native AWS bucket access in read-only mode. The existence of these two service components, much like the authentication and authorization capabilities of Gen3, greatly sped up the development of the GMKF Data Resource Portal.

What Does Not Work Well?

The GMKF Data Resource Center and associated portal were greatly helped along in their development through the reuse of key technologies from the U. of Chicago's Gen3 system along with the use of the Bionimbus FISMA moderate environment. This simplified both the technical challenges of bringing up a functional portal and allowing researchers to have cloud accessible data readily available for compute on Cavatica. However there still remain challenges and areas for improvement.

Single Sign On

The current approach to single sign on utilizes a primary identity with Google or Facebook combined with account linking to Gen3 for accessing GMKF and TCGA data. Each of the latter uses a separate eRA Commons login process to identify the user. Likewise, a Cavatica account needs to be established and linked separately. This account linking activity, while not difficult technically, requires coordination with services that aren't publicly accessible or documented. And it presents a model where each DCC portal will need to develop identical account linking functionality. Ideally, this account linking functionality could be abstracted out of each DCC portal and developed into a common login broker process where a single login to, for example, eRA Commons could be linked to accounts in multiple other systems.

Authentication and Authorization

While the SSO approach described above could yield incremental improvements for both developers of DCC portals and users, a far more pressing issue is the authorization of users in a given system. The dbGaP system represents the current system of record for which NIH studies a particular researcher has access to. Currently, the GMKF Data Resource Center and Portal use a series of whitelists produced on a nightly basis and shared with the Bionimbus system operators. This allows the Gen3 stack to only provide access to files a researcher has access to for all but one of the GMKF studies currently (there is one study that uses a whitelist maintained outside of dbGaP). While this approach works, whitelists are not a real time system so it can take up to 24 hours for a user to be added or removed. Likewise, the representation of consent groups and the nuances of data access are difficult to scale for large numbers of projects. A better approach would allow the GMFK portal to query in realtime the access available to a given user.

Cloud Storage

When dbGaP/SRA closed access to projects wanting to upload large-scale whole genome datasets, the GMKF project was left to find alternatives. The adoption of Gen3 and storage on the AWS cloud environment that resulted was a huge boost to the accessibility and computability of the data. However, the cost of storing data on the cloud now became the responsibility of the project. As the project continues to grow year by year, this cost will grow as well and finding a viable mechanism of supporting this growth will be needed.

What Would a Shorter-Term Improvement Look Like?

The GMKF Data Resource Center and Portal is surprisingly advanced and functional given the length of time the project has been active. While other DCCs struggle with making data accessible on the cloud, the GMKF Portal has managed to work with existing infrastructure and provide this service in a facile compute environment that is flexible and powerful for researchers.

There are multiple areas where improvements over the short term (in this case over the next year) would lead to improvements in the usability and sustainability of the portal. For SSO, having a modular login system that still allows for the use of Google, Facebook and potentially other identity providers along with eRA Commons would be a benefit. The project is currently exploring the use of Auth0 (https://auth0.com) or Keycloak (https://www.keycloak.org) as a mechanism of brokering multiple IdPs quickly and easily with minimal code. The NCBI/CIT teams are also beta testing a new OIDC interface to eRA Commons which would make leveraging the authentication through more modern web standards and toolkits easier. Having Gen3 be able to use the SSO token produced by the eRA Commons login rather than having users repeat an eRA Commons login multiple times would also be a benefit. Likewise, having the account linking functionality abstracted out of the portal into component reusable across multiple DCC portals would be a benefit as well.

For authentication and authorization, the major area of short term potential improvement would be centered around the communication of whitelists. This approach could leverage the JWT-based approach the NCBI is currently prototyping to describe both user identities and claims on datasets they should access using this ubiquitous and modern web token scheme (see “NCBI FEDERATED IAM AND CLOUD DEPLOYMENT PROTOTYPE”). This would allow for real time data access information to be provided for projects that researchers have access to. However, with the prototype nature of this service, the feasibility for a short-term replacement to the current whitelist approach is unlikely. Still, this would be an excellent area to prototype over the short term and there is a significant opportunity to start coordinating a common claims language between Gen3 and NCBI’s JWT service.

For cloud storage, a key concern is cost. Each month the project spends approximately tens of thousands of dollars on AWS S3 file storage. With a projected growth approximately 200TB per year that number will continue to consume more and more of the yearly budget. Controlling this cost is of utmost concern and understanding the process to apply STRIDES program (https://datascience.nih.gov/strides) discounts to the storage of data on the AWS cloud is a key priority.

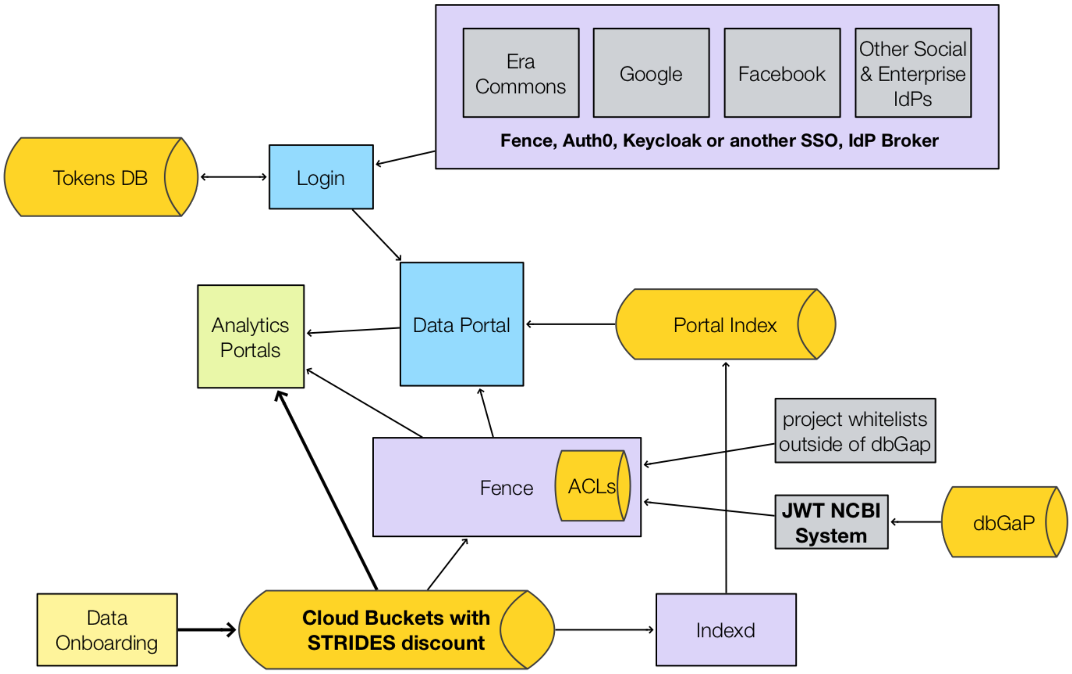

Era Commons
Google
Facebook
Other Social & Enterprise IdPs
Fence, Auth0, Keycloak or another SSO, IdP Broker
Tokens DB
Login
Analytics Portals
Data Portal
Portal Index
project whitelists outside of dbGap
Fence
ACLs
JWT NCBI System
dbGaP
Data Onboarding
Cloud Buckets with STRIDES discount
Indexd

Figure 17: An overview of how a prototypical DCC might work in the next year based on our interview and suggestions by the GMKF Data Resource Center. Text in bold are items that would be fruitful components to improve, prototypes, and test in the short term.

What Would a Longer-Term Solution Look Like?

Longer term, the GMKF Data Resource Center and Portal could benefit from interoperability between the DUOS system (https://duos.broadinstitute.org), the JWT-based system proposed by the NCBI, and Gen3. Specifically, the area of convergence would be the claims language embedded in the JWT tokens produced/consumed by these systems. This would allow the GMKF to, for example, use JWT identity tokens from a successful eRA Commons login to enable a user to apply for access to datasets accessible via the DUOS system. This system provides an automated process of evaluating a user's data access request, claims from their identity (such as being an academic researches at a recognized institution), and data use restrictions (such as dataset X can only be used by a bona fide researcher at an academic organization). Currently these requests are evaluated through a Data Access Committee established per dataset or project in the dbGaP system. An automated system for verifying data access could greatly streamline research. Gen3 could then understand both the claims from the

JWT system from NCBI that encode current approved researchers through the traditional dbGaP flow as well as the claims from the DUOS system allowing for automated data access.

Another fruitful area of longer term improvements, beyond streamlined authorization flows, is data access across systems. The GA4GH is currently defining the Data Repository Service (DRS) API standard that aims to make it possible for multiple systems to enable data access to objects on multiple clouds with a common API. Combined with compatibility between systems in terms of access token claims, DRS will allow researchers to refer to data across clouds and projects. This will enable GMKF researchers to point to data files from other programs and clouds for use in the Cavatica system. To compliment this data access standard, a common format for portals to represent search results, including both metadata fields as well as DRS URIs, would allow search results in one data portal to be “handed off” to multiple environments for computation. A format standard enabling this functionality is currently being proposed to the GA4GH. From a researcher perspective, supporting these two emerging standards will allow her or him to search for data across a wide variety of portals, take search results from each portal, and reference them in their preferred computational environment. This will effectively allow a researcher to create composite synthetic cohorts across multiple projects and access and compute on the data regardless of source cloud of system.

DCC Requirements and Preferences Summary

After interviewing the GMKF Data Resource Center and the GTEx Portal teams, several core requirements and project preferences were identified. These are summarized via questions and responses in the table below.

Projects	**GTEx**	**GMKF**
Single Sign-on (SSO) providers		
Do you currently use SSO in your apps, websites, etc? What is the provider or the mechanism?	Google	Google and Facebook
What specific applications (websites, other?) does your SSO currently cover?	GTEx portal	GMKF Data Resource Portal
Do you share an SSO mechanism with other organizations or projects?	Controlled data access uses dbGaP via eRA Commons currently but now using AnVIL via Terra which does use Google linked to eRA Commons	Account linking with Fence provides access to dbGaP managed data via eRA Commons login

Is it important for your users to be able to login to your app/website/etc. using the same SSO mechanism as another app/site/etc?	Yes	Yes
Do your apps/websites allow login via eRA Commons?	Yes, now using AnVIL via Terra which does use Google linked to eRA Commons	Account linking with Fence uses an eRA Commons login
Authentication and Authorization		
Beyond eRA Commons, what other identity providers can they authenticate with? (E.g., Google, ORCID, campuses)?	Google	Google and Facebook
Is there interest in adding more identity provides?	TBD	Yes, interested in supporting more logins with an identity broaker platform like Auth0
Do your applications use an authorization whitelist via dbGaP for data access control?	Yes	Yes via Fence from Gen3
Does your application offer an alternative source of authorization information beyond dbGaP?	TBD	Yes, at least one project manages authorization outside of dbGaP and GMKF relies on whitelists in Fence to manage
Data storage and authorization mechanisms		
Do your applications use an authorization whitelist via dbGaP for data access control?	Yes	Yes via Fence from Gen3
Do your applications use native cloud storage URLs plus temporary credentials for data access control? What specific mechanisms/providers?	No, the current MITRE system uses signed URLs exclusively. Yes for AnVIL which uses native URL access on Google.	No, the Cavatica system uses signed URLs via AWS.
Do your applications use signed URLs for data access	MITRE system provides but not fully functional. Nice to	The Cavatica platform generates signed URLs to

control? What specific mechanisms/providers?	have	access data in Gen3
Do your applications use a group/attribute service for access control?	Yes, they have an exchange site for research groups to share	Groups are supported in Gen3 from whitelists and used for access control.
What other authorization approach(es) do your applications use for access control?	TBD	In addition to whitelists from dbGaP some projects maintain their own Data Access Committee outside of dbGaP and they support this with whitelists maintained in Gen3's Fence.
Data on the cloud Do you currently store data in the cloud? Which cloud(s)?	Via 1) MITRE solution at SRA for v7 (currently broken) 2) v8 has been onboarded into Terra workspaces as part of NHGRI Anvil.	Yes, Gen3 stores the GMKF data (~3PB) on the AWS cloud.
Do you allow data access via native cloud access mechanisms?	yes	no, only signed URLs are needed
Are your data access controls based on your SSO mechanism or do they use a different authentication method?	In the future they want controlled access data in their portal and this would be behind an eRA Commons login	No, SSO is Google and Facebook whereas data access is done through linking to eRA Commons through Fence

Table 1: Requirements summary for DCCs we interviewed.

Available Solutions

In the process of interviewing GTEx and GMKF several technology solutions were identified that provided SSO, AuthN/Z, or cloud storage solutions for these projects. In addition, there are several software solutions used by the wider community that provide similar (or identical) features. Each solution provides distinct abilities, benefits, and drawbacks. Here we compare solutions identified in our interviews with DCCs and compare their features. This is not a comprehensive list but we prioritized comparison of software solutions that were flagged in either DCC interviews or are known to be used in similar DCC systems. *Over time we expect to add additional systems to this listing and evaluate those solutions as we interview additional Common Fund DCCs.*

SSO

Single sign-on (SSO) is a property of authentication systems whereby a user logs in with a single username/password yet gets access to multiple systems. It could be explored as part of authentication but we break it into its own section since it is an important-enough topic and there are a few approaches we see the DCCs currently taking and/or mentioning.

Google Sign-in and Facebook Login

Both the GMKF Data Resource Portal and the GTEx Portal support Google Sign-in (https://developers.google.com/identity/). This allows the portal to access the Google identities of users such as email address and unique Google user ID. This is not a traditional SSO since multiple IdPs cannot be used and there is some repeated login process as a user moves between systems. For example, when a user logs into the GTEx portal having previously logged into a GSuite service, the user is prompted to select which Google account should be used (if there are multiple) but the user is not required to enter their password again. Overall, this system is flexible and easy to use, Google, GSuite, or university email addresses powered by GSuite can be used but other OIDC-based IdPs cannot. Google Sign-in is based on OpenID Connect (OIDC).

The GMKF Data Resource Portal also supports the Facebook Login authentication API from Facebook (https://developers.facebook.com/docs/facebook-login/). Much like Google Sign-in, Facebook Login allows the GMKF portal to identify the user via their email and unique identity on that platform. Similar to Google, this does not represent a true SSO since multiple IdPs cannot be used and there is some repeated login process as a user moves between systems. However, the system was easily incorporated into the GMKF portal and provided this identity function. Facebook Sign-in is similar to but not identical to the OpenID Connect (OIDC) standard.

eRA Commons/NIH Login

The eRA Commons/NIH Login system provides an SSO solution widely used throughout the NIH and partner sites. While Google uses OIDC and Facebook uses a proprietary, but similar, approach eRA Commons login uses the Security Assertion Markup Language (SAML) for establishing a user identity. The eRA Commons identities are used in dbGaP to authorize users to access data, see the next section.

Auth0, Keycloak, and other SSO Implementations

In addition to calling a given IdP directly, such as eRA Commons via SAML, Google via OIDC, or Facebook via their Login API, various third party authentication and authorization platforms exist to manage users, identities, and privileges. These typically support multiple IdPs and authentication/authorization flows simultaneously and provide various management tools to streamline use. Of these platforms, several support configurations for SSO including Auth0 and Keycloak. When we interviewed the GMKF Resource Center there was a desire to experiment and leverage management platforms such as these to simplify the ability to support multiple IdPs in the future. Specifically, both Auth0 and Keycloak were mentioned.

Auth0 (https://auth0.com) provides the ability to support Single Sign-on (SSO) via their Universal Login feature (https://auth0.com/docs/universal-login). Auth0 supports a wide range of identity providers ranging from social providers like Facebook, Google, and Twitter to enterprise solutions like Active Directory, LDAP, and any OpenID Connect providers. As an identity hub, these multiple Identity Providers supported by Auth0 use various protocols including OpenID Connect as well as SAML, WS-Federation, and others (https://auth0.com/docs/identityproviders#social). Auth0 is a hosted service and charges various monthly rates depending on the features, number of active users, and account integrations requested.

Keycloak (https://www.keycloak.org) is another identity and access management solution that was also mentioned during our interview with the GMKF Data Resource Center. Like Auth0, Keycloak provides SSO abilities and supports a range of social and enterprise identity providers. It supports authenticating users with OpenID Connect or SAML 2.0 identity providers. Unlike Auth0 which is a hosted service, Keycloak is a server that is installed and run for a given project.

Solution	**Google/Facebook**	**eRA Commons**	**Auth0, Keycloak, etc**
Multiple IdPs supported	no	no	yes
Can this mechanism be used by unaffiliated applications? (Could a random researcher use it in an app?)	yes	no	yes
What authentication protocol(s) does this service provide?	OIDC and proprietary respectively	SAML	OIDC, SAML, and others
What identity providers are supported? (I.e, Which organizations can users	Google and Facebook respectively	eRA Commons	Many both social and enterprise

authenticate with?)			
What identity data does the service provide to applications?	Token for identity and possibly other scopes	Token for identity	Token for identity and possibly other scopes
Where does the identity data originate? (Registration mechanisms, organizational support, validation, etc.)	Google and Facebook respectively	eRA Commons	Many both social and enterprise
What are some notable applications that use this service?	These services are widely used	This service is widely used as an identity provider for NIH systems	These services are widely used
How is this service supported on an ongoing basis? (Sustainability mechanism, sources)	Commercial	NCBI/CIT infrastructure	Commercial
What is the user support (help desk) organization for this service?	Self service forums and online resources	NIH Login helpdesk	Self service forums and online resources

Table 2: A comparison of different SSO providers.

Authentication and Authorization

While authentication and authorization are often times intertwined, they represent distinct concepts. Authentication is about identifying the user while authorization provides information about what a user is allowed to do. In the previous section we described authenticating a user through services that provide a Single Sign On (SSO) experience. In this section we examine two systems that provide multiple capabilities but we will focus on the authorization aspects.

Gen3 - Fence

Fence (https://github.com/uc-cdis/fence#token-management) is part of the Gen3 stack (https://gen3.org) and has multiple capabilities including:

- acting as an auth broker to integrate with one or more IdPs and provide authentication and authorization to other Gen3 services
- managing tokens
- acting as an OIDC provider to support external applications using Gen3 services
- issuing short lived, cloud native credentials and/or signed URLs

As used in the GMKF portal, Fence acts as an auth broker, allowing users to log in via eRA Commons and the calling portal is able to retrieve an identity, refresh, and/or access token used to respectively establish identity, obtain access tokens, and access data resources in Fence via native credentials or signed URLs.

In this way, Fence acts as a bridge for the GMKF portal, allowing users to link their Google or Facebook IDs with access to Gen3 hosted datasets for GDC or GMKF via their eRA Commons identity. Fence ultimately manages access to data objects cataloged in Indexd by utilizing a whitelist approach. These whitelists are provided through a secure transfer mechanism with dbGaP and define which eRA Commons users can access data from which projects and consent groups.

NCBI JWT based on Virtual Directory Service from CIT

When Fence acts as an auth broker with eRA Commons the authentication flow uses SAML. This is an extremely common authentication solution but older than the more modern OIDC approach which uses JSON Web Token (JWT) responses instead of XML. Likewise, the identities that can access GDC and GMKF data is represented in whitelists which are difficult to maintain and synchronized on a schedule rather than in real time, meaning there can be a delay for users gaining access to data (or being removed when their access has expired).

NCBI has developed a proposal for a better solution that addresses these two concerns (see “NCBI FEDERATED IAM AND CLOUD DEPLOYMENT PROTOTYPE”). First, the proposal will produce JWT identity tokens for users logged into eRA Commons. This token establishes the identity of the authenticated user. Furthermore, that access token can then be converted into a jwtPassport token that uses JWT claims to represent the projects and consent groups that user has access to. This approach is appealing because it represents a generic mechanism of retrieving information for a user on which projects he or she should have access to. These tokens can be interpreted by systems like Fence and used to provide (or reject) access to data on the cloud.

This proposal will be prototyped over the next year and is an opportunity to replace dbGaP whitelists with a much more scalable and flexible system.

Solution	Gen3 Fence	NCBI JWT Proposal

Protocols supported	Auth broker for OIDC and SAML systems	TBD, leverages JWT user and passport tokens
JWTs used	Yes	Yes
Strategy for authorization	References whitelists prepared by dbGaP	Live representation of data access claims in JWTs
Can this mechanism be used by unaffiliated applications? (Could a random researcher use it in an app?)	Some features	TBD
What authentication protocol(s) does this service provide?	Works with OIDC and SAML	TBD
What identity providers are supported? (I.e, Which organizations can users authenticate with?)	Google and Shibboleth (NIH iTrust, InCommon, eduGAIN)	TBD
What identity data does the service provide to applications?	OIDC with JWT formatted tokens	JWT formatted tokens
Where does the identity data originate? (Registration mechanisms, organizational support, validation, etc.)	multiple	dbGaP for authorization, eRA Commons for authentication
What authorization mechanism(s) does this service provide?	OAuth2 with JWT formatted tokens	JWT formatted tokens
Does this mechanism provide a group service?	Yes	TBD
What are some notable applications that use this service?	GDC, the NCI Cloud Resources, NHLBI Data STAGE, NHGRI AnVIL and other projects	Prototyping at NCBI now
How is this service supported on an ongoing basis? (Sustainability mechanism, sources)	Community and supported through GDC and other projects	Prototyping at NCBI now
What is the user support	Community and help desks	Prototyping at NCBI now

(help desk) organization for this service?	per project	

Table 3: A comparison of different authentication and authorization solutions.

Cloud Storage

Cloud storage looks at how data files can be stored in AWS S3 or GCP storage buckets and then accessed by authenticated and authorized users. We examined two solutions for storing and accessing data on the cloud. For the GMKF Resource Center, Gen3 was used for storing and sharing genomic data files on the cloud while GTEx used dbGaP with a solution built by NCBI/MITRE for providing access to the files on the cloud. Later GTEx used plain Google Storage buckets for storing V8 release files and shared these via the Terra analysis platform.

Gen3 - Indexd + Fence

In Gen3 there are two micro services that control access to data on the cloud: Indexd and Fence. Indexd is an indexing service, it catalogs the location of data files in buckets in both Google and Amazon clouds and associates each file with one or more Globally Unique IDs (GUIDs) such a Data GUIDs (https://dataguids.org). This allows systems and users to refer to data files via these IDs and resolve actual cloud locations via Indexd.

As described in the previous section, Fence verifies a user's access to files in Indexd. For authorized users, Fence allows them to retrieve credentials to access the files cataloged in Indexd. This includes returning native cloud credentials that can be used to temporarily access the file paths in S3 or Google Storage or signed URLs which can be used directly by a variety of applications.

Gen3 is used to provide access to GMKF data on AWS and the signed URL approach is used by Cavatica for researchers to access and work with the data on the cloud.

SRA Cloud Storage

To facilitate access to SRA data, MITRE created a solution for accessing dbGaP data copied to AWS and Google via a service that produces signed URLs. This is the current mechanism the GTEx portal uses for data access to V7 of their dataset on the cloud. For files in SRA, the MITRE solution allows users to obtain an API token per dataset they are authorized to access. This token can then be used, along with the sample ID, for retrieving signed URLs on AWS and Google. This is an early service and is currently limited, for the GTEx project, to only data available in SRA (release 7 only). Signed URLs for the Google cloud are not publicly available, only users in a closed beta can access signed URLs in this cloud. On AWS URLs can be signed but only within the AWS environment, the signed URLs are not intended for use outside of the cloud.

Solution	Gen3 Fence + Indexd	SRA Cloud Storage
Clouds supported	AWS, Google, private cloud	AWS and Google (beta)
Signed URL support	Yes	Yes
Native, temporary credentials support	Yes	No
Can this mechanism be used by unaffiliated applications? (Could a random researcher use it in an app?)	Yes	Yes
What authentication protocol(s) does this service provide?	JWT token issued by Fence	Token available via the dbGap website per project
What identity providers are supported? (I.e, Which organizations can users authenticate with?)	The same as Fence described previously	eRA Commons
Where does the identity data originate? (Registration mechanisms, organizational support, validation, etc.)	Google and Shibboleth (NIH iTrust, InCommon, eduGAIN)	eRA Commons
What are some notable applications that use this service?	GDC, the NCI Cloud Resources, NHLBI Data STAGE, NHGRI AnVIL and other projects	SRA for data in dbGaP
How is this service supported on an ongoing basis? (Sustainability mechanism, sources)	Community and supported through GDC and other projects	Unknown
What is the user support (help desk) organization for this service?	Community and help desks per project	SRA help desk

Table 4: A comparison of different cloud storage solutions.

Emerging Themes

After interviewing a small collection Common Fund DCCs we are starting to build up a profile of the general needs of DCCs. From this, common themes are emerging for both current approaches but also desires for how systems can work better in the future. This section looks at common themes emerging for the DCCs, specifically areas they would like to see improvements in.

Shorter Term

These are themes the DCC we interviewed flagged as being areas for improvement over the next year or so (September 2019 - October 2020):

- An improved system for whitelists from dbGaP, the JWT proposal from NCBI being quite appealing
- Maintaining the ability to have whitelists for projects not in dbGaP
- Getting data on the cloud and accessible
- Being able to use data in multiple analysis systems, DRS for common interface to cloud storage
- Native cloud URI support in addition to signed URLs
- OIDC support in addition to SAML for eRA Commons login
- Support for site logins with multiple social/enterprise IdPs using Auth0 or something similar
- Common claims language between dbGaP (JWT Proposal) and Gen3
- Clear onboarding guide for STRIDES

Longer Term

The DCCs we interviewed shared longer term themes that can be worked on and will likely be multi-year efforts. Here are some of the longer term themes that emerged from our conversations:

- Ability to link eRA Commons with Cloud accounts and other IDs (passport)
- Streamlined trusted partner program so DCCs can redistribute data via their portals
- DUOS for automated data access management

Shorter Term Proposal

Overview

In the previous section we examined the themes from the DCCs we interviewed that are opportunities to improve in the short term (approximately the next year, September 2019 - October 2020). Taking stock in what the GMKF Data Resource Center and GTEx DCC have built over the last several years, **this section looks to identity what we might do in the next year that will identify areas that could be improved, develop those improvements, and ultimately document a prototypical cloud-based DCC template for either creating new Common Fund DCCs or advancing an existing one**. The goal is not to be proscriptive for portal development, metadata/data curation and harmonization, and all the other elements needed by a DCC. The intent, instead, is to develop clear guidelines for a DCC to stand up SSO, authentication, authorization and cloud storage solutions that works in the current, or emerging, ecosystem of NIH and cloud services and to do this expeditiously. By streamlining this process, we would expect more reuse of components across DCCs and also an increased ability to focus on other aspects of creating a successful DCC, such as portal development and data/metadata curation.

Goals

Overall, we have identified the following as possible goals for shorter term work over the next year:

- Identify one or more Authentication solution(s)
  - OIDC for eRA Commons prototype
  - Explore use of Authentication broker (Fence, Auth0, Globus Auth, etc)
    - SSO
    - account linking

- Identify one or more Authorization solution(s)
  - Identity tokens from the proposed NCBI JWT system with claims representing data access in dbGaP
  - Gen3 using the same JWT claims language as the NCBI proposal

- Identify one or more Cloud Storage Systems
  - Gen3 Indexd + Fence
  - Overture

- Look for cloud agnostic solutions for the above when possible

- Look for free and open source solutions for the above when possible

- Produce a guide, one step above an install guide, that documents the components of a prototypic system, how the system would work, and how users would interact with it either directly or through third party systems. This would constitute a best practice guide and checklist for building a DCC cloud environment with specific component options presented.

Anti-Goals

- This is not an installation guide for specific SSO, authentication, authorization, or cloud storage solutions but that may be a future output of related work

- We do not want to build any solutions from scratch, this is a proposal for documentation, integration, and augmentation work based on existing solutions that are demonstrated to work for existing projects.

Requirements

A Common Fund DCCs needs to support the following. So the work proposed in the short term should support these requirements:

- Cloud Support: both AWS and Google

- Standards: use modern, industry standards whenever possible (OIDC, OAuth2, etc)

- Community Standards: Contribute to in progress standards (GA4GH DRS)

- Existing Components: Use and/or extend existing implementations of proven software components

- Data Storage:
    - Support access to data on the cloud using both 1) signed-URLs and 2) native URLs
    - Buckets should be supported as either centrally managed or managed by a given DCC
    - Cloud storage should be compatible with the STRIDES program
    - Must support ability to enable requestor pays

- Authentication: need to support eRA Commons and other identities, linking together

- Authorization: need to support access control information from dbGaP as well as a self-administered whitelists for projects not in dbGaP

Proposed Work

To accomplish the goals of this short term proposal, we are looking at the GMKF Data Resource Center and Portal as an excellent candidate to work with in the next year. This group has already identified key areas to improve, how they might make these improvements, and are willing and interested to distill their experience into guidelines for other DCCs. Figure 18 illustrates the areas in which the GMKF Data Resource Center would like to improve over the next year. These are aligned with the short term work proposed here and this DCC would be a wonderful partner for prototyping the improvements described below.

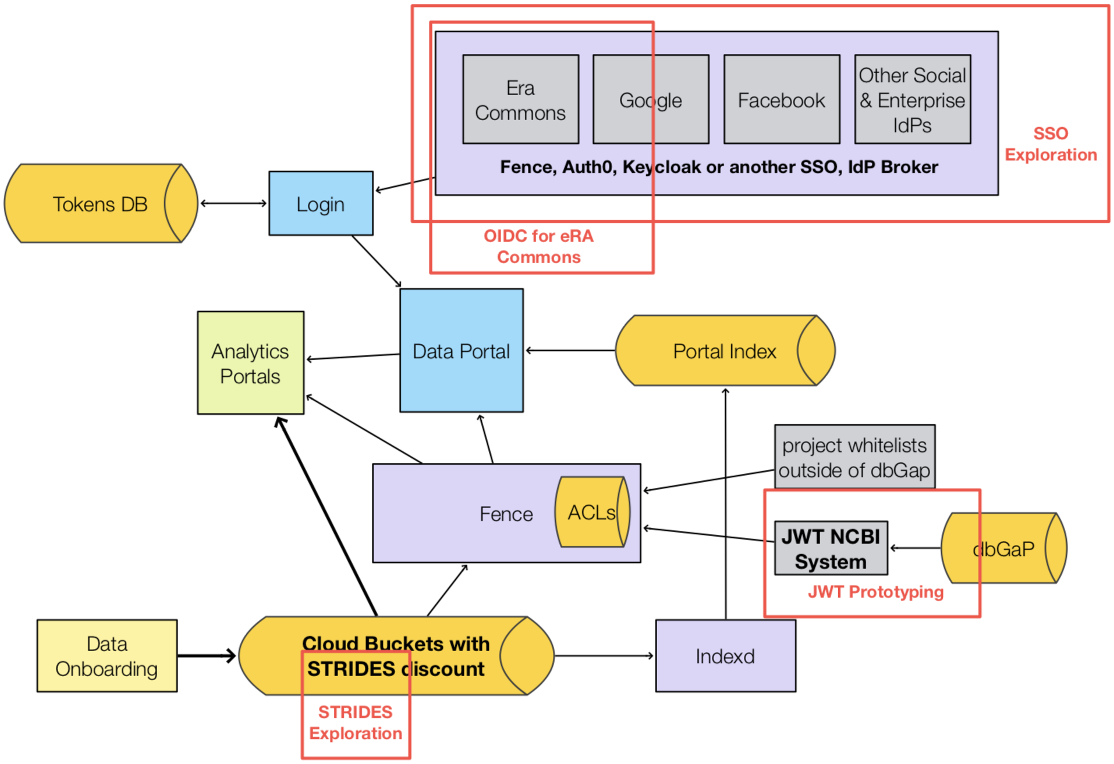


Figure 18: Areas (highlighted with red boxes/text) where a short term effort could both improve the functionality of the GMKF DCC but also provide a template and working example for other DCCs to leverage.

SSO

The GMKF Data Resource Center has expressed interest in evaluating commercial and open source solutions for SSO. Options include Fence, Keycloak, Globus Auth, and Auth0. The proposed work would be to evaluate these different solutions that could be used to increase the number of login options for users.

Authentication and Authorization

Authentication: OIDC for eRA Commons login

Another opportunity is to evaluate the use of SSO providers with the prototype OIDC eRA Commons authentication flow. This is currently being prototyped in the Globus Auth system. The OIDC login support for eRA Commons would make it considerably easier to use SSO/identity brokers like Auth0 and Keycloak.

Authorization: JWT system from NCBI

The JWT system being developed by NCBI could prove to be more effective way of conferring information about which projects a user can access than whitelists. The short term work could look to incorporate support for this system in Fence.

Data Access on Cloud

One key concern about moving data to the cloud is the long term costs associated with storage and use. The STRIDES program provides a way to receive discounts. THe GMKF Data Resource Center has, to date, not set up their buckets to use the STRIDES discounts. A key goal for the short term is to set up their budgets with the STRIDES discount and to document the process to help others with the same task.

Sharing Knowledge

At the end of the proposed short term work, we will produce an overview of deploying a Common Fund DCC (SSO, AuthN/Z, and cloud storage specifically) based on the experiences of the GMKF Data Resource Center and the improvements, testing, and development proposed here. The goal is to document a prototypical cloud-based DCC so others can use this as a template for either creating new Common Fund DCCs or advancing an existing DCC.

Longer Term Proposal

Beyond the work proposed in the short term of the next year, there are two areas of interest for longer term enhancements for DCCs like GMKF: Passports and systems like DUOS. Standards around these topics are actively being developed by the GA4GH Data Use and Researcher Identity (DURI) work stream (https://ga4gh-duri.github.io/categories/welcome.html).

Passport

Researchers have multiple identities such as an eRA Commons account, institutional email addresses, and cloud accounts. The concept of the passport is that these identities can be unified in a single digital identity. This passport would not only combine these multiple digital identities, but it would also include claims about the user, such as being a bona fide researcher at a known research institution or university. Researchers can use this passport for accessing services but also for gaining access to data, which is explored in the next section.

DUOS

DUOS (https://duos.broadinstitute.org/#/home) is a semi-automated management service for determining user authorization to access data based on the claims from a researcher's passport along with the data use consents of the data being accessed. In place of simple whitelists, it can match researchers to datasets they should have access to and can automate the granting of access rights to the user. As with passports, DUOS, and systems like it, are of great strategic value to DCCs like GMKF since they could potentially greatly reduce the time and effort needed to gain access to controlled access datasets.

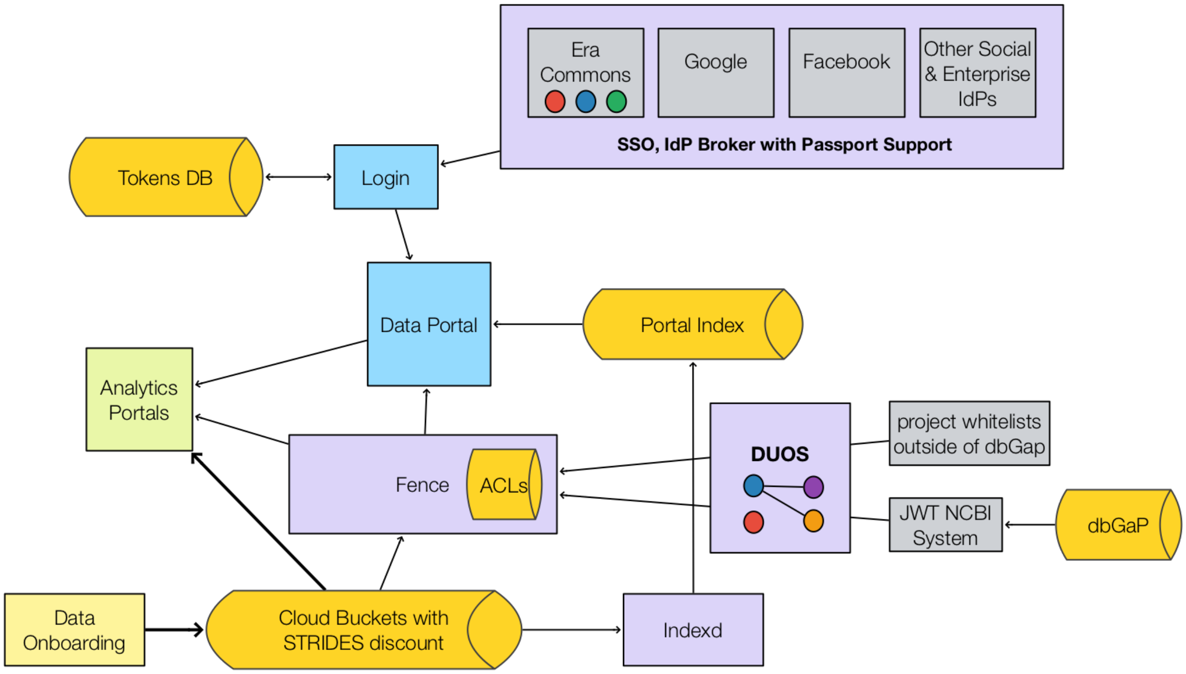

Era Commons
Google
Facebook
Other Social & Enterprise IdPs
SSO, IdP Broker with Passport Support
Tokens DB
Login
Data Portal
Portal Index
Analytics Portals
Fence
ACLs
DUOS
project whitelists outside of dbGap
JWT NCBI System
dbGaP
Data Onboarding
Cloud Buckets with STRIDES discount
Indexd

Figure 19: An example of how DUOS and researcher identity Passports can be overlayed into a DCC similar to the GMKF architecture. User identity tokens include claims about the researcher while DUOS performs matching between the user and dataset access policies, proving the result to a system like Fence which manages access to files on the cloud.

Conclusions

In the course of this report we have interviewed multiple Common Fund DCCs, created a summary of their current abilities and desired future functionality, and profiled various technical solutions for SSO, authentication, authorization, and cloud data storage. We have observed a strong desire to move data and compute to cloud environments where they can be leveraged by researchers more effectively and scalably than downloads to local compute environments. The DCCs themselves have expressed interest to use standardized technologies when possible, to use modern protocols like JWTs, OAuth 2 and OIDC, and to leverage existing components that have been demonstrated to work in other projects. The end result is a short term proposal for work that, if successfully completed in the next year with a DCC such as GMKF, would provide a useful prototypical example of a functional, cloud-based DCC that others can leverage.

Next Steps

We have identified next steps towards supporting both the shorter term proposal as well as a longer term vision. Over the next several months we will:

- Continue to refine this report through September 2019. We will solicit feedback from community stakeholders, fix any factual errors, and continue to refine the document

- Add information from additional Commons Fund DCC interviews

- Add information from the evaluation of additional SSO, AuthN/Z, and cloud data storage systems such as:
  - Overture (https://www.overture.bio)
  - Globus (https://www.globus.org)
  - Others TBD

- Continue to refine the short term proposal for creating a prototypical DCC recommendation, working together with DCCs such as the GMFK Data Resource Center

- Continue to look at ways to enhance and solidify the longer term proposal

Appendix - Technology Primer

Core Concepts

Authentication: The goal of authentication is to verify someone's identity.

Authorization: Authorization is granting someone the power to do something.

Clouds: Services that provide virtual machines and other resources for rent. Examples include Amazon Web Services (AWS) and Google Cloud Platform (GCP). Clouds provide programmatic access to their resources allowing developers to use hardware in much the same way as software components. It can also provide a service where a user delegates a task to a hosted solution that another entity runs and operates and rents out its resources for a fee.

Signed URLs: A stand-alone URL prepared with an authenticated and authorized request that allows for temporary access to secure resources for the holder of the URL.

Single Sign On (SSO): A property of access control of multiple systems, SSO allows users to log in once and access several related services without having to login multiple times.

Identity Provider (IdP): A service that allows users to log in and establish their identity, can then be provided to downstream services.

Auth Broker: A service that can authenticate users to a variety of IdPs.

OAuth 2: A common protocol for providing authorization to system resources on behalf of a user.

OAuth 2 Client: An application which wants to access resources in another system on behalf of a user.

OAuth 2 Authorization Server: A service that gives access tokens to an OAuth 2 Client once the user is successfully authenticated and authorizes the access.

Access Token: A string issued by the Authorization Server to a Client for the client to access particular controlled resources on behalf of a user.

OpenID Connect (OIDC): Based on OAuth 2 but modified to support authentication, this produces an identity token that's used to identify the user.

OpenID Provider: An OAuth 2 server which implements OIDC.

Relying Party: An OAuth 2 client which uses or requests OIDC access tokens.

Authentication

For any of the SSO mechanism, we essentially agree that the "Identity Provider" that will authenticate the user is a trusted entity. For example, if an SSO mechanism is set up for a user to log into a website using Google, it can be trusted that Google is capable of managing the security of the user and that authenticated Google user can be trusted to login into other portals (usually by matching metadata such as an email address and matching it against the portal being accessed). It can be trusted that the user matching joe@gmail.com in Google and joe@gmail.com in another portal is the same individual and that individual is being granted the credentials associated with that email. The drawbacks of any SSO system is that it's as secure as the third party service provider.

SSO can also serve as the Identity Provider authenticating a series of applications. . A login service can be set up internally for a university, for example, and it can be presumed that the login service is secure. If a user is authenticated by the login service, any other apps that use that same service as a login entry can grant a person access to their individual apps since the login service verified the user's identity..

Authorization, is the step that happens after the Authentication process, this step essentially determines what data a user has access to and/or what data they have permission to do.

SAML

[2]SAML is an XML based authentication mechanism between a "Service Provider" and an "Identity Provider".

The Service Provider agrees to trust the Identity Provider to authenticate users. The Identity Provider generates a response that validates the user's authenticity — essentially confirming that the user is allowed into a website or service.

It's an Identity Provider that enables seamless authentication, mostly used between businesses, enterprises and academic[3] and research facilities[4]. It's also the older implementation that has primarily been replaced by Open Connect / OpenID[5].

Benefits:

- Standardized
- Platform neutrality
- Does not require data synchronization of users
- SSO, Improved User Experience (sign in once and access all resources under the same purview)
- Increased security, relying on a central authority

[2] https://auth0.com/blog/how-saml-authentication-works/
[3] https://incommon.org/federation/federation-join/
[4] https://www.geant.org/Services/Trust_identity_and_security/Pages/eduGAIN.aspx
[5] https://en.wikipedia.org/wiki/Security_Assertion_Markup_Language

Drawbacks:

- If Central authority is compromised, auth fails globally. It is only as secure as its weakest link. You are trusting the Identity Provider to be secure.
- Complexity is built into the app (same with any centralized auth).
- If SAML provider isn't internal, you trust a third party with user data.

Very Simplified Auth Process:

1. Request sent to SSO provider
2. If authenticated, then redirects to website as an authenticated user, otherwise user is redirected to SSO URL provider to login (via user/password, 2FA, etc)
3. SSO Provider sends SAMLResponse to client which is validated to ensure authentication succeeded.

OIDC (OpenID Connect)

OpenID Connect (OIDC) is an authentication protocol, based on the OAuth 2.0. It uses JSON Web Tokens (JWT), which you can obtain from the Identity Provider.

While OAuth is primarily used for authorization, OIDC is a way to authenticate users.

Key Concepts:

[6]**Access Tokens**: are credentials that authenticate a user and can be used to verify a user's identity. It can be used to obtain further information about the user.

ID Token: Unlike Access token these tokens contain specific information about a given user, such as username, email, or other PII.

Claims: JWT Tokens contain information or a claim about a user. A set of predefined claims already exists defined by the OIDC standard. Implementation also supports custom claims that extend the support metadata provided by the protocol.

1. When you choose to sign in to a given website (client) using an OID provider, the client sends an Authorization Request to the Identity Provider.
2. The Identity Provider authenticates your credentials or asks you to login if you are not already signed in, and asks for your authorization (lists all the permissions to the resources that the client is requesting).
3. Once you authenticate and authorize the sign in, the Identity provider will send an Access Token and optionally an ID Token back to the client.

[6] https://auth0.com/docs/protocols/oidc

4. The client can then use the Access Token to invoke an Identity Provider API as long as each request is signed with the access token.

Comparison:

- OpenID Connect is a rewrite of SAML using OAuth 2.0.
- OpenID Connect is newer and built on the OAuth 2.0 process flow. It is tried and tested and typically used in consumer websites, web apps and mobile apps.
- SAML is its older cousin, and typically used in enterprise settings as well as many academic institutions eg. allowing single sign on to multiple applications within an enterprise using our Active Directory login.

Author Notes:

XML in general is a bit antiquated these days, JSON tokens tend to be easier to work with, parse and read. The new implementation of OID is essentially an evolution of SAML and improved upon the existing mechanism. The RFC defining the OAUTH 2.0 Standard, which OIDC and OID is based on came out in 2012. SAML 2.0 is from 2005.

Other differences to keep in mind.

- OpenID: is an authentication mechanism.
- OAuth 2.0: is the authorization protocol that OpenID is based on. (It's also the defacto standard for authorization to date)
- OpenID Connect: is a combination of OpenID and OAuth 2.0 mixed together serving as both authentication and authorization solution.

Authentication and AWS/Google

Authentication on AWS:

- Amazon Cognito provides a service that allows your app to authenticate via their service.
- Supports a variety of open standards such as: Oauth 2.0, SAML 2.0, and OpenID Connect.
- Allows for user registration as well as auth.
- It adds Active Directory Auth proxied so, in theory, it can seamlessly integrate into a corporate infrastructure.
- User pools are user directories that provide sign-up and sign-in options for your app users.
- Identity pools provide AWS credentials to grant your users access to other AWS services.

Amazon LWS:

- Is a hosted service that allows user to login, but seems very much tailored to the shopping experience and tied to your amazon account.

User Pools has some value, the Identity Pools seem to be the authorization solution though it is inherently tied to AWS infrastructure.

Authentication on Google:

- Google Identity: https://developers.google.com/identity/
- This is tied to a Google account. In comparison, AWS Cognito allows you to simply implement an authentication platform on the cloud.
- Very tied and integrated into Google services for better or worse.
- Options for Google seems more limited and less robust

Author Note:

Hosted solutions such as Cognito and Google Identity have various benefits and disadvantages in comparison to an in-house solution. While these systems make implementations dependant upon the resources provided, their SLAs are the best you can expect and in-house solutions rarely are able to improve upon them. These solutions can streamline development since their primary business model is to make it easy to build on top of them. This can mean increased productivity since the complexity of deploying authentication and SSO solutions can be enormous. Still, solutions need to be carefully evaluated and options weighed before committing to such a core infrastructure component.

Authorization

Authorization, is the step that happens after an Authentication process, this step essentially determines what data a user has access to and/or what they are able to do.

SAML

SAML does not do authorization explicitly. It simply provides the attributes in the SAML token and it's up to the application as to how these are handled and interpreted.

OAuth 2

Is essentially a protocol that allows a user to grant access to a resource. User A can read the employee salary history. User B can issue a payroll check and so on.

[7]OAuth has the following roles:

- Resource Owner: The entity that can grant access to a resource.
- Resource Server: The server hosting the resource
- Client: the user or app making the request.

[7] https://auth0.com/docs/protocols/oauth2

- Authorization Server: the server ensures that the user is issued access after authentication.

OAuth utilises two endpoints which need to be accessible. (Keep in mind unless this is a hosted solution these can be internal.)

Authorization EndPoint:
Used to gain access to the protected resource. I.e., used to gain access.

Token Endpoint:
Is used to get an access token. Also used to refresh the token once expired. The Authorization code return by the auth endpoint is passed to the token endpoint. The access token received is then used to sign all requests to access the given resource.

OAuth uses a JSON Web Token same as OIDC.
header: contains type of token, and crypto algo information
payload: contains a set of claims (ie permissions) that are allowed.
signature: used to validate token.

Authorization and AWS/Google

See the previous Authentication section.

Storage on the Cloud

S3

S3 is Amazon's cloud data storage solution. It is a key based data store that mimics a directory structure on a computer. The data is stored on the cloud and can have private and or public access.

The S3 storage format is usually as follows:

s3://<bucket-name>/<path>

Permission and Authorization can be controlled at the bucket level as well as on a per path level.

Example:

Bob has access to s3://private/ and can read/write everywhere except s3://private/confidential

Joe has access to s3://private and can only read/write to s3://private/confidential

Control can be gradual and can be customized based on the need.

S3 is primarily used to store data. It has a limited functionality when it comes to searching and its performance is very slow.

Case in point:

If I would like to retrieve a certain file, such as demonstrated in the example below, I can easily do so via:

```
aws s3 cp s3://my-private-cloud/genomics/2019/s/smith/john/20190613.txt  .
```

On the other hand if all I know is that a file called 20190613.txt exists somewhere in my bucket, and I need to find it, it’s a very slow process and will be equivalent to searching your entire hard drive file by file.

There are different tiers for S3 including a ‘cold storage’ concept where data is stored in a less accessible service but at a lower pricepoint.

If the ’servers’ are AWS instances, then the instances can be provided with Roles, which can be automatically granted rights based on the Role assigned to the server.

E.g. genomics-host1 is granted genomics-research role, which it automatically has read/write access to.

s3://private/genomics-research.

Importantly, S3 buckets and the data they contain can be setup with a requestor pays model, ensuring that the user requesting data actually pays for egress charges rather than the owner of the bucket/data. This is extremely useful to prevent enormous egress charges for downloads from outside the AWS cloud environment.

Amazon S3 and the other cloud services offered by AWS are rich, complex, and highly functional. This description only scratches the surface but highlights some important features of storing and sharing data on the AWS cloud.

Google Storage (GCP)

All the relevant object storage services outlined previously on AWS are also available on Google under the Google Storage service. The Google Storage format is usually as follows:

gs://<bucket-name>/<path>

Permission and Authorization can be controlled at the bucket level as well as on a per path level. Authorization is done via: https://cloud.google.com/iam/

There are different tiers for Google Storage including a ‘cold storage’ concept.

Currently the cost of storage is slightly more for Google over Amazon/Azure at the consumer level. However, discounts can be setup for use in corporations or institutions, such as the NIH STRIDES program.

Like AWS S3, Google Storage buckets and the data they contain can be setup with a requestor pays model, ensuring that the user requesting data actually pays for egress charges rather than the owner of the bucket/data.

Like AWS, Google Cloud Platform offers a wide variety of feature rich and highly functional cloud solutions. This short introduction of storage on the Google Cloud Platform just scratches the surface of a much larger collection of services and features of this platform.